DC_TEL: Framework for assisting the data production and analysis in TEL

Nadine Mandran, Isabelle Girault, Cédric Ham

DC_TEL: Framework for assisting the data production and analysis in TEL.

Mandran Nadine
CNRS, LIG
F38000 Grenoble, France
Nadine.Mandran@imag.fr

Girault Isabelle
Univ. Grenoble Alpes, LIG
F38000 Grenoble, France
Isabelle.Girault@imag.fr

D'Ham Cédric
Univ. Grenoble Alpes, LIG
F38000 Grenoble, France
Cedric.dham@imag.fr

This paper proposes a framework, DC_TEL, to support the data processing in the field of Technology-Enhanced Learning (TEL) environments. The DC_TEL provides a guide to enhance the collaborative work between a TEL researcher and a specialist in data production and analysis. The framework formalizes and details the data production and analysis in six steps. Each step is split in objectives and tasks. For each task, DC_TEL proposes 7 attributes, which are important to assist the data production and analysis and also to use the new educational datamining tools. In this respect, it takes advantages of the works on the data processing, on the data quality and on the educational data mining tools. The DC_TEL framework is implemented in a new computing platform, called Undertracks. An instantiation of this framework is carried out by a data analyst specialist to assist two researchers in chemistry education to enrich data produced within the TEL environment "copex-chimie". This instantiation shows the advantages, weaknesses and ways to improve globally this framework for the multidisciplinary TEL researches.

1. INTRODUCTION

The content of this paper is in line with in the enhancement of methods to guide the data processing in Technology-Enhanced Learning (TEL) Research. The proposition allows the TEL researcher to carry out the datamining in relation to the pedagogical theories and the practices of teachers and students. The goal of our proposition is to enhance the collaboration between the specialist in data production and analysis [1] and the TEL researcher, and more particularly when he needs to produce data and analyze them with the educational data mining tools. Indeed, today the number of data and educational data mining tools increases rapidly. The performance of the tools increases, sometimes the complexity can increase too. Thus, the TEL researcher no accustomed with these tools, is face-to-face with the complexity to

[1] Specialist in data production and analysis is called "data analyst" in the rest of the paper. Also, the role of the data analyst can be the creation and the development of the data mining tools.

implement these tools on its data. This difficulty in communication and collaboration between the different actors around the TEL research on the means of analysis and data production is mentioned by [Siemens and Baker 2012]. Our proposition wants to reduce this difficulty and can help the dissemination and combination of these tools.

Roughly speaking, in data mining, the data processing consists of two major steps. The process starts with data production, based on a study protocol, and continues with data analysis. While these two steps are often described, the transition between them is not clearly defined. This transition is referred to as the pre-processing step by [Romero et al. 2014] who indicate that "this important step is rarely described by the authors". The framework proposed includes this transition by two iterative and additional steps: data validation and data enrichment. These steps require data analysis tools, such as descriptive statistics, visual data analysis, and action by human specialists. The framework we propose also adds a data quality dimension. We can use four indicators to enhance data quality: “relevance”, “accuracy”, “temporal precision”, and “ease of interpretation” [Di Ruocco et al. 2012]. Our proposal takes into account the data process scheme, the data quality indicators and the works on the educational data mining tools to create a framework supporting the data processing in TEL researches.

The paper is organized in six sections. The next section describes the related data governance works, according to data process and quality. Section three describes the DC_TEL framework that is decomposed in steps, objectives and tasks. Section four briefly presents the Undertracks platform that we designed to implement this framework. Section five depicts a case study using data produced during a chemistry training session supported by a TEL system [Girault and d’Ham 2014]. This instantiation illustrates the collaboration around the framework between two researchers in chemistry education and a data analyst specialist. Section six presents the added value of this framework, together with its limits and future works.

2. RELATED WORKS

The framework described in this paper needs to be structured with the literature. The related works concern data processing and quality.

2.1. DATA LIFE CYCLES

Many data organization processes are available, and their terminology and organization are specific to the activity sector (e.g. industry, management, engineering) and the business process. The terminologies used to qualify the data process steps are numerous, and may be confusing. It is therefore essential to clarify the terminology and determine the aim of each step.

In EDM, several processes are presented. The transition between data production and data analysis is referred to as the “pre-processing” step by [Romero et al. 2014] (e.g. the step where data must be validated). Their total process includes seven main steps. The first two steps (Data gathering - Data aggregation/integration) fall within the data production step; while the third step (Data cleaning) can be associated with the “Pre-processing” step, and the last four steps (User and session identification - Attribute selection - Data filtering - Data transformation) are related to data analysis and are partly on a technical level. Also, the process is depicted as linear.

Another process, the Knowledge Discovery in Database (KDD) process, was proposed by [Fayyad et al. 1996]. This process is also linear, and includes a pre-processing and transformation step, which is not clearly defined.

Regarding social sciences and humanities, the UK.DATA ARCHIVE[2] [Bishop 2012] proposes a data processing in the form of a "data life cycle". This process includes 6 steps: - Creating - Processing - Analyzing - Preserving - Giving access - Reusing. We find the idea of a cycle interesting since a data life cycle provides some results, which can lead to other issues. With these new issues, a study design can be once again defined. However, there is a combination of steps: three steps correspond to the data process itself (creating, processing and analyzing), while the three other steps correspond to dissemination issues (preserving, accessing and reusing). The "creating data" step is the "data production" step, while the "processing data" step could be what [Romero et al. 2014] called the "pre-processing" step to prepare data analysis. If we take a closer look at the "processing data" step, it also includes different issues: data process and dissemination. The data process, described by "check, validate, clean data", is more complex than the "pre-processing" step of [Romero et al. 2014] that only includes "cleaning". Last but not least, the "analyzing data" step includes "interpret data, derive data, produce research outputs, author publications, prepare data for preservation", which are tasks that we believe do not correspond to the analysis step itself but rather correspond to tasks than can be carried out after the analysis.

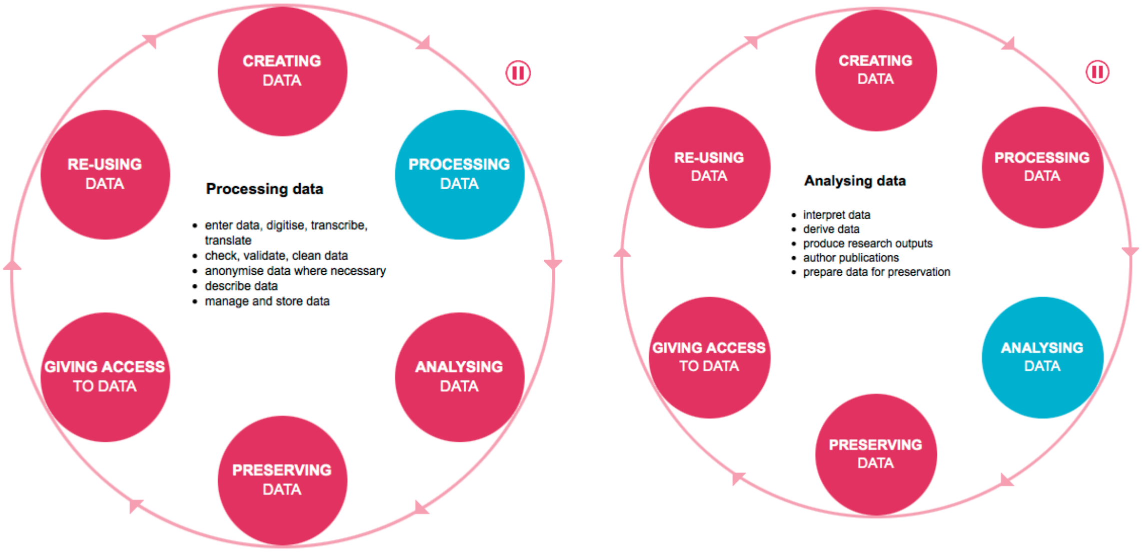


Figure 1: Data life cycle described by UK.DATA-ARCHIVE,

Comparison of the different data process schemes shows: 1- the two major steps are split into several minor steps 2- there are some steps between production and analysis, but they are not well defined, and the content is different for each scheme, 3- the terminology used in each scheme is often different, 4- the content of the tasks used in each step is different, 5- the position of the tasks is not the same according to the scheme, 6- the human skills necessary to

[2] http://www.data-archive.ac.uk/

lead the steps are not defined 7- there are two forms of process: linear and circular. This literature study shows a need to specify the terminology, define the intermediate steps and the main tasks needed to perform these steps, and situate these steps and tasks in the process.

2.2. DATA QUALITY

According to the standard (ISO 8402:1986), the quality is defined as the totality of features of a product or service that fulfill stated or implied needs, for instance the correspondence to specifications, expectations or usage requirements, the absence of errors. [Brasseur 2005] gives some properties about the data quality: “Data quality can address the needs of its users”, “Data quality is dependent of their use”, “The understanding of user needs is a prerequisite for defining and obtaining data quality requirement”, “A big difficulty is that the bad data quality is not easily detected. There is often some incidents or anomalies during the operational work, which reveal, here and there, of inconsistencies on data.” Therefore, throughout the data life cycle, it is important to have methods and tools to control data. Latter two are important to take into account because the non-quality of the data has a cost.

The impact of the poor data quality has a strong cost on the technological development [Haug, et al. 2011]: “75% of organizations have identified cost stemming from dirty data, 33% of organizations have delayed or cancelled new IT system because of poor data, less than 50% of companies claim to be very confident in the quality of their data. Business intelligent (BI) projects often fail due to the dirty data, so it is imperative that BI-based business decisions are based on clean data, the organizations typically overestimate the quality of their data and underestimate the cost of the errors". They define “the optimal data maintenance effort”, with two assumptions “1- During data maintenance the focus is on the most critical data before moving on to less critical ones.” 2- the costs of the efforts to ensure high data quality are not causally related to their importance (e.g. focusing on a set of poor quality data with great impact on costs is not necessarily cheaper than focusing on data with little impact on costs). Thus, the costs of assuring data quality is a linear relationship between data quality and assurance costs.” The quality has a cost and it is necessary to limit this cost to ensure the data quality. In our case, the choice to use a quality approach must be controlled to limit the costs of the data quality. Therefore, a selection of data quality indicators is necessary.

In computer information systems, data quality proposes four approaches dedicated to improve data quality before the analysis step: (1) preventive, (2) adaptive (3) corrective and (4) diagnosed approaches [Berti-Equille 2007]. The preventive approach allows upstream control before production. It is based on the quality of the model and the quality development of the software. The adaptive approach allows data verification in real-time. The corrective and diagnosed approaches are conducted after data production. The corrective approach mainly includes: comparison with field reality, missing data imputation and elimination of duplication. The diagnosed approach mainly includes: exploratory data mining, descriptive statistics and metadata management.

Also, [Di Ruocco et al. 2012] quote several indicators of data quality:

- Relevance: responding to the needs of the study now and for the future
- Accuracy: data compliance compared to reality
- Completeness: verification that the necessary objects are present in the data model
- Consistency: of data when the databases are copied or duplicated
- Temporal precision (Timeliness): accuracy versus time where the data are represented.

- Accessibility: ease of locating and accessing data
- Ease of interpretation: ease of understanding data, their analysis and their use. Data must be understood without ambiguity.
- Uniqueness: a single object, a single record in the system, represents a real-world entity.
- Coherence: the absence of conflicting information.
- Conformity to a standard.

In relation to the data produced with a TEL environment and the four approaches to improve data quality, we can organize these indicators in three categories:

1- Before data production: "Completeness", "Conformity to a standard", “Uniqueness”. This kind of indicators cannot be easily respected with TEL environments that provide educational data. To our knowledge, there is no standard for the production and storage of educational data. To ensure completeness, there is not always a data model. Tracks are rarely produced with a data model, and if so the model is rarely provided to describe data production.

2- To store data: "Consistency", "Accessibility". This second kind of indicator is not relevant in the context of our framework. This point ensures data reuse, but this paper does not approach the issue of data storage.

3- After data production and during the intermediate steps: the “relevance”, “accuracy”, “temporal precision”, “ease of interpretation”, and “coherence” can be taken into account.

In our framework, we propose to introduce some quality indicators because they ensure the data quality and consequently the results quality. However, take into account all of these indicators are expensive. Nowadays, in our framework we not use all of these indicators. As we have not the model of the software and we are not in real-time, we can use only the diagnosed and the corrective approaches.

3. Steps and tasks to guide the data processing in TEL research

The literature study reveals four gaps. 1- The terms used to qualify the steps of the data life cycle are numerous, not well defined and may prove confusing, 2- The transition between production and analysis, as well as the tools and practices for validating data, are not well described, 3- To our knowledge, no quality indicators are used to ensure data quality produced with a TEL environment, 4- The human skills useful for leading the different steps are not defined.

In an attempt to address these four gaps, we propose here a new framework, called DC_TEL. About the first three gaps, our purpose is to enhance and clarify the data life cycle. To reach this goal, DC TEL proposes a circular data life cycle that includes six steps to completely describe the data processing. Each step is well defined with objectives and tasks. Each task is characterized by seven attributes. One of these attributes is the implementation of the data quality indicators. Regarding the fourth gap, two kinds of role are necessary and complementary: TEL researcher and data analyst. DC_TEL provides a guide to enhance the collaborative work between the TEL researchers and data analysts.

3.1. FOREWORD ABOUT DATA

Before detailing our framework, we would like to clarify several points of terminology about the data and some properties of the data in TEL.

1. We choose the term "study" rather than "experimentation”. The term "study" is more general; it can include experimentation such as those designed in experimental psychology, or studies such as those conducted in sociology with individual interviews or focus groups.
2. We choose in our framework not to use the terms "process", "processing" and "pre-processing" because this terminology is confusing. We will use the verb "to process" only to describe data analysis with operators such as statistics or data mining, and only in the "analysis” step.
3. The term “data” is very broad. We define data like a corpus, which are produced during a study. They can be qualitative (e.g. interview, video) and/or quantitative (e.g. questionnaire, log), produced before the study (e.g. pre-test evaluation), during the study (e.g. logs) or after (e.g. post-test evaluation). They also include the metadata to describe the study (e.g. study protocol, user features). We can split data into several measurements. We use the term “variable” to indicate the basic elements of the measurement. For example, the time stamp and the user code are variables.
4. A “track” is a result of an observation recorded at a specific time-stamp; it contains a set of descriptors that reflect the activity of the observed individuals. This definition takes into account both the interaction of the agents with the TEL environment and the production of the agents [Iksal 2012]. To describe the tracks accurately, we have five categories of data: “event data”, “study description”, “context data”, “agent data” and “action data”. We detail these five kinds of data in section 4.2.

These clarifications allow us to situate the context of use of DC_TEL: we use the tracks (Event data) described with metadata (“description of the study”, “agent data”, “action data”). Our framework is design for the data produced in the context of the didactic research. These data are built with a study protocol, the students use a TEL environment and when the tracks are not as numerous as in the case of big data (e.g. MOOCs).

3.2. FOREWORD ABOUT CATEGORY OF OPERATORS

First, we choose not to use the terms “tools” because this terminology is confusing. We will use the noun “operators”. An "operator" is a person or an automatic system, like computer, whose are employed to operate or control a machine.

Nowadays lots of operators are available, to guide researchers in this labyrinth, we need a first level of classification. To lead the data analysis, we choose the categorization proposed by educational data mining. In educational data mining (EDM), operators are classified into six categories [Peterson et al. 2010] : 1- Prediction, 2- Clustering, 3- Relationship mining, 4- Distillation for human judgment (DHJ), 5- Visual data analysis (VDA) and 6- Discovery with model. Descriptive statistics (DS) are not explicitly mentioned by [Peterson et al. 2010] However, in this paper we choose to mention them explicitly, as DS are the first category of operators controlling data value relatively on the field and ensuring control of the quality indicators. In our different studies the data are not numerous. Thus, the four others categories of operators are not appropriate in our framework since their results are too synthetized. Also, raw data can be crushed by these operators’ categories. (reference de baker 2014 EDM)

Another category of operators can be used to manage, control and store the data. These specific operators belong to the "data management" (DM) category of operators. During the data life cycle, the data management is essential. It is a set of different operators, which allows the evolution of data or/and the evolution of the data files. In the case of the data transformation, the operators can have an impact on the variables (e.g. computing a ratio, creation of new variable by the combination of the others) or the records (e.g. selecting a set of students). The operators that allow the merging or the splitting of files have an impact on the data files changes.

3.3. COMBINATION OF OPERATORS

Ensuring an efficient combination of operators for understanding the behavior of TEL users always presents a major challenge [Baker and Yacef 2009]. [Johnson,L. et al. 2010], indicate that the blend of different technics is a mean to assist the understanding of the complexity of the behaviour, like in a use of the TEL environments. “Visual data analysis field is an emerging fields, a blend of statistics, data mining and visualization that promises to make it possible for anyone to sift through display, and understand complex concepts and relationships”. In our framework, we implement this concept of combination. Therefore, we detail some of these operators and a combination that we use in our instantiation.

DHJ is a “technique that involves depicting data in a way that enables a human to quickly identify or classify features of the data.” [Bienkowski et al. 2012]. DHJ allows more precise mining than automatic analysis methods, and gives more sense to the research question. The human skills provided by DHJ are essential for validating and enriching data. They refine semantics around data, and can ensure data accuracy. However, human specialists need help to explore and mine data. For this, data and their representation must be presented to the researcher in an intelligible form. “Human beings can make inferences about data, when it is presented appropriately, that are beyond the immediate scope of fully automated data mining methods" [Baker 2010].

In the first approach to explore the data, the descriptive statistics and the visualization data analysis (VDA) can guide the human experts. For the descriptive statistics (DS), the simple statistical indicators are: frequencies, percentages for qualitative data, mean, standard deviation, median, quartile and coefficient of variation for quantitative data. [Howell et al. 2007]. The visualization data analysis (VDA) allows the representation of data, with the goal of guiding human expertise, “Visual data is a way of discovering and understanding patterns in large datasets via visual interpretation” [Johnson,L. et al. 2010]. In the next section we describe a specific visualization operator that we created to address researchers’ expectations in TEL. The intelligible form can be provided by an appropriate visualization and by descriptive statistics. In data mining, a large number of operators and software are created to visualize data in various ways [Rosling 2009] [McCandless and Cunéo 2011], Many eyes (IBM website)[3].

To the best of our knowledge, DHJ works are less frequent. Regarding VDA, interesting operators are available but data visualization is usually implemented at the end of the process

/

to summarize data and communicate the results. Some EDM works present the combination of DHJ and VDA for analyzing data [Sao Pedro et al. 2010][Baker and de Carvalho 2008][Desmarais and Lemieux 2013] [Gobert et al. 2013]. In our framework, we explore a similar combination of DHJ and VDA operators to validate and enrich data.

3.4. INTRODUCTION OF FOUR INDICATORS OF DATA QUALITY

The implementation of the data quality could be expensive and the terminology presented in the section 2.2 is very generic. To deploy the data quality in our context, we choose four indicators that we refine in relation to our data. Table 1 clarifies the use of these four data quality indicators for our framework.

Indicators and definition	Clarification to use data quality indicators in our framework
"Relevance": responding to the needs of the study	The data must be necessary and sufficient to adress the resarch question and the needs of the data analyst. For instance, data produced by the TEL environment must be used by the TEL researcher to study the students' behaviors, and also by the data analyst to manage and analyze the data.
"Accuracy": data compliance compared to reality	The data produced by the TEL environment must be in line with the expected activity of the TEL users. For example, before the study, the researcher chooses some interesting actions to address the research question. When the data are produced a control must be done to verify that the same action is always coded in the same way.
"Temporal precision": accuracy versus time where the data are represented	The time unit produced by the TEL environment must ensure that the sequentiality of the data is maintained. If it is not the case, we must produce other element to maintain it. Creating a variable sequence of the data production is a mean to control the time stamp. For example, if the time stamp does not include the second and if two data are produced during in the same second, we can lose the sequentiality of the data.
"Ease of interpretation": ease of understanding data, their analysis and their use. Data must be understood without ambiguity	The data produced must be described and also the new data created during the data processing. For example, the computation between initial variables (from raw data) can be useful to enrich the analysis (e.g. the difference between the post-test and the pre-test create the variable learning). These new variables can be documented to ensure the ease of the interpretation.

Table 1: Quality indicators, generic definition and clarification of the chosen indicators in the DC_TEL context

3.5. DATA LIFE CYCLE: 6 STEPS

The step is the first level in breaking down the data life cycle. We propose that a new step be defined whenever data status evolves. To clarify the data life cycle leading data evolution, we identified six data statuses and steps. These are: 1- design the study, 2- collect data, 3- validate data, 4- enrich data, 5- analyze data, 6- summarize the results.

Figure 2 describes the organization between the six steps and the ability or not to iterate between them. Between the steps “Validate”, “Enrich”, “Analyze” and “Synthesize”, iterations and comebacks are often useful. However, the progression between the steps throughout the framework complicates the comebacks and increases the duration of the data life cycle. In contrast, the comeback between the two steps “Validate” and ”Collect” creates a major problem, as non-validity of data can lead to data destruction, collection of other data and redesign of the study protocol. If we diagnose at the end of the validation step that data are invalidated, we need to recollect them. In some cases, we need to redesign the study protocol. Obviously, in this case the duration and cost of the data life cycle increase rapidly. Control of the two first steps of the DC_TEL seems essential to produce suitable data for addressing the research issues, prior to validating and enriching data.

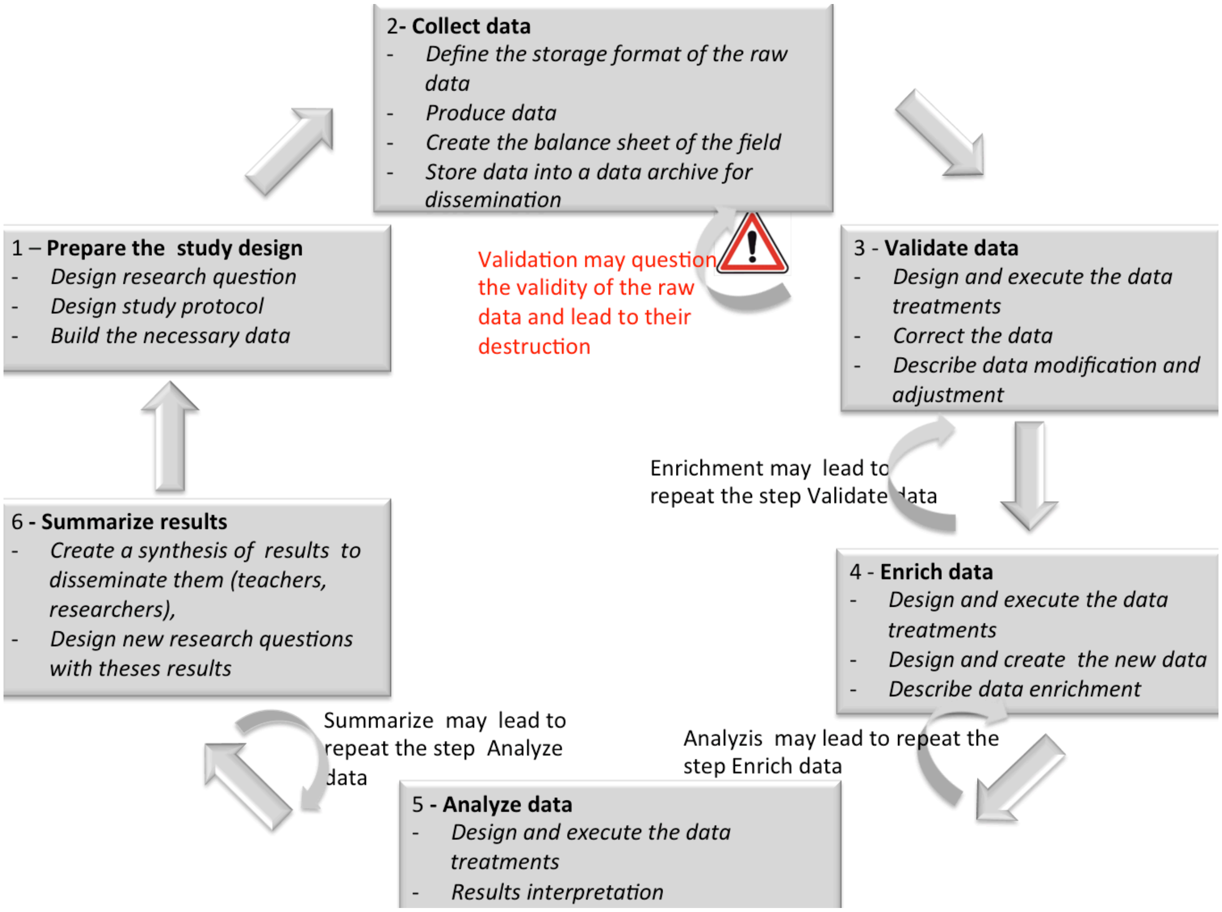

2- Collect data
- Define the storage format of the raw data
- Produce data
- Create the balance sheet of the field
- Store data into a data archive for dissemination
1 – Prepare the study design
- Design research question
- Design study protocol
- Build the necessary data
Validation may question the validity of the raw data and lead to their destruction
3 - Validate data
- Design and execute the data treatments
- Correct the data
- Describe data modification and adjustment
Enrichment may lead to repeat the step Validate data
6 - Summarize results
- Create a synthesis of results to disseminate them (teachers, researchers),
- Design new research questions with theses results
4 - Enrich data
- Design and execute the data treatments
- Design and create the new data
- Describe data enrichment
Summarize may lead to repeat the step Analyze data
Analyzis may lead to repeat the step Enrich data
5 - Analyze data
- Design and execute the data treatments
- Results interpretation

Figure 2: Organization and iteration between the 6 steps and their associated objectives in DC_TEL

3.6. FOR EACH STEP A COMPREHENSIVE LIST OF OBJECTIVES AND ASSOCIATED HUMAN ACTORS

At each step a set of objectives must be carried out to transform the data. We propose a list of objectives to guide the data life cycle and a list of objects that must be produced at each step.

The objectives can be chained. However, none of them are mandatory. For instance, in the step “validate data”, the objective "describe data modification and adjustment” is not mandatory to ensure evolution of data statuses but it is important to ensure dissemination and reuse of data. This list is an initial proposal, and may be enhanced in future work. In this list of objects, we propose several balance sheet documents. These allow us to evaluate each step of the data life cycle with step strengths and weaknesses. Knowledge of weaknesses is required to enhance a future study. These requirements are necessary and essential for conducting a study, the data of which are stored for dissemination and reuse.

With respect to the role of human expertise in our framework, two actors can be identified: data analyst and TEL researcher. After the data production, the data analyst is in charge of: 1- identifying the operators required to analyze data, 2- implementing them with the best application conditions, 3- guiding the TEL researcher in analyzing data. The TEL researcher is in charge of: 1- identifying the research issue, 2- identifying useful data, 3- controlling the data validity and accuracy and 4- adding some semantics for data and results.

Table 2 proposes a definition of the six steps and the associated actors. It also lists the objects produced in relation to the change in data status.

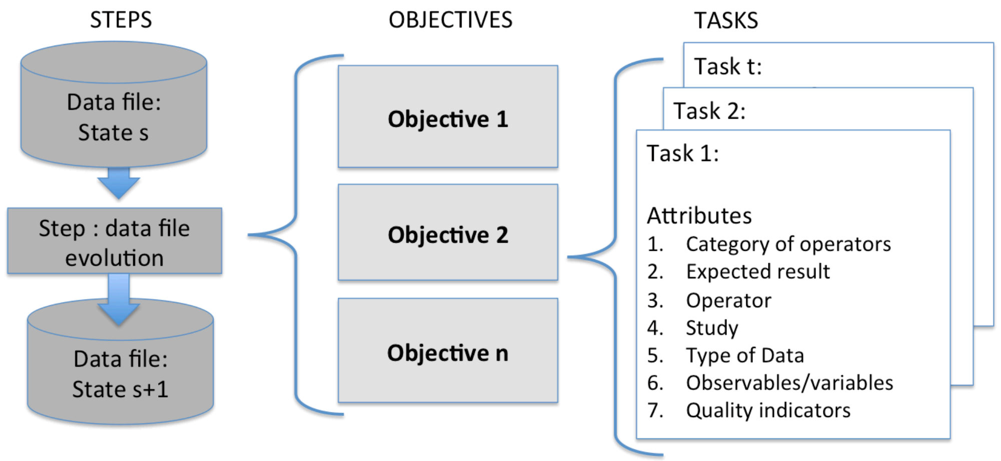
Step	**Definition**	**Objectives and associated human actors:** TEL researcher (TR), data analyst (DA), assisted or not with computer and software (CS) (most important role in bold).	**Objects produced**
Prepare the study design	In relation to the research question, a study protocol is built. This protocol designs the data required by the researcher to address the research question	(TR) defines research question (**TR**/DA) designs the study protocol to ensure efficient data collection. **TR**/DA) builds the necessary data.	**Study protocol** Study or experimental material (questionnaire, grid for interviews, etc.) Description and format of tracks when a TEL is used.
Collect data	Raw data are produced on the study field and are stored with metadata to describe the study context, data and balance sheet of the study field. Raw data are produced on the study field	(TR/**DA**) defines the storage format of the raw data (TR) and/or (CS) produces data (TR) creates the balance sheet of the field (TR)(CS) stores data on a data archive system for dissemination	Meta data file **Raw data file** Description of format storage Balance sheet of the field study

Validate data	From raw data, some data treatments are performed to validate coherence of data in relation to the expected values. The enriched data must be validated too. For each, new data created the data must be validated. Therefore, the two steps are iterative.	(TR/**DA/CS**) designs and execute data treatments (TR) corrects data (TR) describes data modification and adjustment	**Validated data file** Description of data modification and adjustment
Enrich data	Other data are added to the validated data. These data may come from external sources, may be an encoding of validated data or may be a combination of two or more validated data items	(TR/**DA/CS**) designs and executes data treatments (TR) designs, creates new data (TR) describes data enrichment	**Enriched data file** Description of data enrichment
Analyze data	The enriched data are processed, and the results are produced to address the research question	(TR/**DA/CS**) designs and executes data treatments (TR) interprets the results	**Results files** Data processing scheme or algorithm or program Data treatments balance sheet
Summarize results	The analyzed data are summarized to provide a synthesis for communicating the results	(TR) creates a synthesis of results to address the research questions and creates documents to disseminate these results (e.g. paper, report) (TR) defines new research questions with these results (e.g. outlooks in research paper)	Documents to disseminate these results Perspectives of the research questions

Table 2: Definition of the steps for attaining the different data statuses, the list of objectives and the objects produced at each step. The associated actors and the need for computer operators are indicated. The most important actor to attain the objective is indicated in bold.

3.7. TASK ATTRIBUTES

While splitting the data life cycle into steps and objectives is a guide, the definition of these objectives continues to be too coarse and must be refined. An objective is reached with a succession of several tasks, in other word a task is an instantiation of the objective and each task has some attributes that describe the task. We identify seven attributes: 1- the category of operators as described in 3.2, 2- the description of the expected results, written by the TEL researcher 3-the operator to lead the task, 4- the study, 5- the type of data as described in 3.1, 6- the observables/variables, 7- the quality indicators as described in 3.4.



At least one data file is created to store and to follow up the data evolution.

For each step, the list of objectives is comprehensive.

For each objective, several tasks can be made. Each of them is described with seven attributes.

Figure 3: Generic description of the DC_TEL framework, with its steps, objectives, tasks and the attributes for the tasks.

Depending on the needs and skills, each of the actors will inquire the attributes. The refinement of the attributes allows the enhancement of the comprehension between the actors. There is a mean to assist the data analyst to find or to create the most appropriate data mining tools in relation to the TEL researcher's goals.

4. UNDERTRACKS PLATFORM

Before presenting an instantiation of DC_TEL, we describe the Undertracks platform (UT), which can support our framework. (http://undertracks.imag.fr),

Undertracks [Bouhineau, Luengo, et al. 2013]consists of the two main software: UTP and UTA. UTP is dedicated to the production of data and specific operators, while UTA is dedicated to the combination of data and operators, in order to validate, enrich or analyze data. It is the place where the task chain is built.

4.1. UTP AND UTA

From the operators point of view, each data mining, statistical or visualization algorithm that could be applied on data is called an "operator" in UT. An operator is an entity that takes data as an input and may (or may not) provide new data or results as an output. Combining data and operator is equivalent to a task element of DC_TEL.

UTP is the part of the UT platform that produces and stores data and operators. UTP is based on the data base management system (DBMS) model. This database is accessible via a web application. UTP provides five tables with respect to DBMS: study description, event data, agent data, action data and context data. Only two tables are mandatory: study description and event data. UTP provides an interface to store and modify operators. Each new operator has to be described with mandatory fields: (1) the input and output data format, (2) the parameters modifying operator behavior, and (3) the documentation describing the usage, version, and the algorithm used.

UTA is a software designed to combine data and operators stored in UTP [Bouhineau, Lalle, et al. 2013]. The user can graphically link data and operators to create a visual workflow (Figure 4). In this visual interface, a box is a dataset or an operator. A blue box is an operator for loading data. A green box is a computational operator. A yellow box is a visualization operator. This visual interface allows tasks and task chains to be created. Concretely, the user combines data and operators and decides to execute the workflow. Once it has been executed, the user can consult the final results as well as the intermediate data.

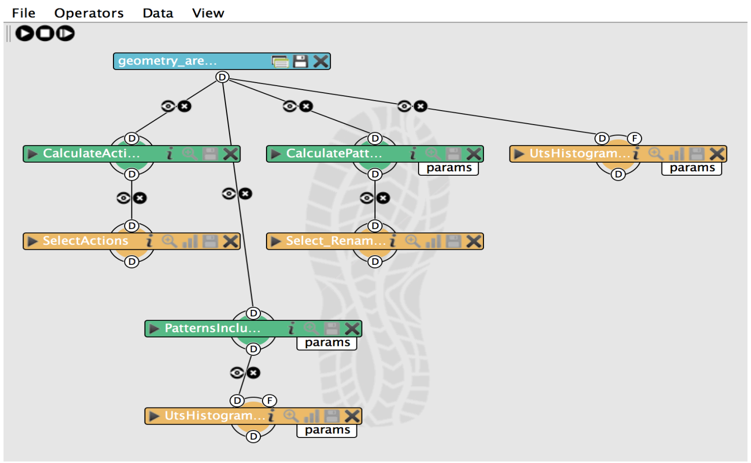


Figure 4: Screenshot of the UTA software. This connects data with operators, with the goal of creating the sub-task chains.

4.2. DATA ON UT

As explained in section 3.1, we have five categories of data: "event data", "study description", "context data", "agent data" and "action data". The latter four are considered to be metadata that allow data capitalization and dissemination.

Description of the study stores research questions, hypotheses and the implementation conditions. These data are used to clarify the aims of the study, to control data production and to ensure reusing and dissemination of data. On UT these tables are called “description study”.

Context data are data providing information about the educational context used in the study with the TEL environment.

Agent data are the characteristics of the agent. They are both interesting and useful to enrich analysis and to understand data with individual features. We use the term "agent" because it is less restrictive than the term "user". Indeed, sometimes we need to gather the events produced by the computer system or any other technical devices. If the agent is a student, data can, for example, describe the demographic features of the student (e.g. sex, age) and the student’s curriculum. The characteristics of an agent are a set of variables.

Event data are collected when an agent uses a technology-enhanced learning (TEL) environment. The event data format is represented by at least a time code, an agent code and an “action”. An agent that interacts with a TEL environment builds the variable “action”. During study design, the researcher forecasts a finite list of “actions” with, for each “action”, a comprehensive list of setting parameters. The data produced are a series of actions that allow patterns to be created. In our context, we consider the tracks like event data (see definition in 3.1).

Action data describe and refine the actions stored in the event data. To give an example, with software used by students to learn mathematics, we can acquire the action "make a division". The setting parameters can be the operands of the division ("numerator” and “denominator”) but also the result of the division (“result”), or the validity of the result (“validity”). For instance, with the Mathematics TEL, the action can be a “division of two integers” and the characteristics can be “denominator”, “numerator”, “result” and the “remainder of the division” On UT these tables are called “action data”.

4.3. SPECIFIC VISUALIZATION OPERATORS ON UT

We created a specific operator on UT, which produces a representation of the sequence of the actions. The sequence is represented on a time lime. Temporality can be represented by the time stamp of the action, the order of the action or the duration of the action. All the time lines are presented on the same screen, enabling them to be dragged and dropped, and thus allowing visually similar time lines to be brought closer.

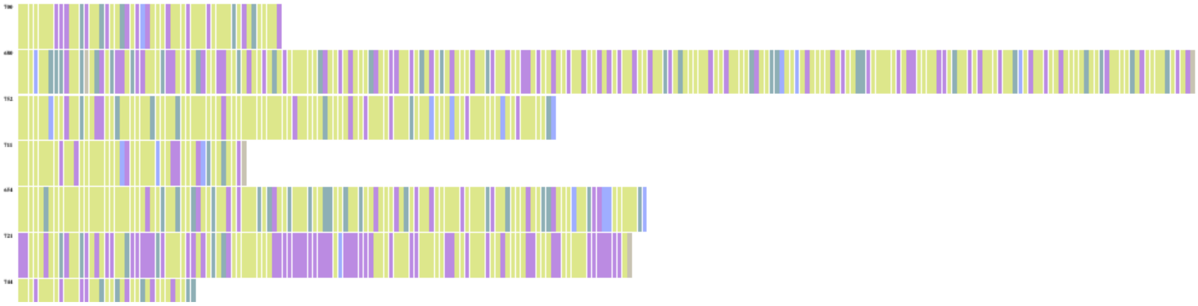


Figure 5: Time lines for sequences of actions for 7 students, where each action is represented by one bar. The color is relative to one action.

As UT is for sharing and reusing data and processes, UTA provides a operator to describe and store new processes in the UT database. Users can also download processes in order to execute or modify them. We use the UT platform to implement the instantiation of our framework presented in the next section.

5. INSTANTIATION OF OUR FRAMEWORK TO ENRICH DATA

To instantiate our framework, we choose a study performed with the TEL environment copex-chimie. This section describes our work related to the step "enrich data" of DC_TEL. We choose this step because it is rarely described while it is important to ensure the data quality.

We first briefly describe the context of the study and the copex-chimie environment. Then we detail the data produced during this study. Finally we present a succession of tasks (and their attributes) performed on these data, to enrich them in order to answer the research question. This instantiation is led with two researchers in didactic (TEL researcher) and a data analyst.

5.1. PURPOSES AND DESCRIPTION OF THE STUDY

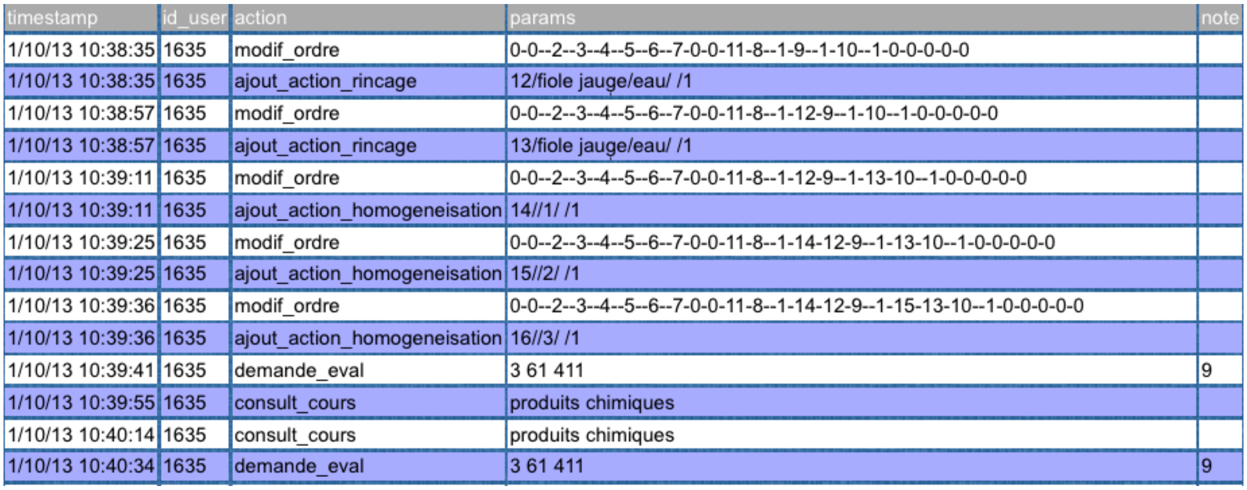
Copex-chimie is a computer environment that scaffolds the activity of experimental design. It has been conceived to help learners design a specific experiment in Chemistry.

The TEL researchers want to understand if the students manage to design an experiment under different conditions of the computer environment. The different conditions of copex-chimie are related to the embedded scaffolds [Girault and d'Ham 2014]. The first scaffold included in copex-chimie is generic and related to the pre-structuring of the task and aims to help the students to achieve the task. The other type of scaffold is individualized and is provided through the feedback messages of an artificial tutor that give information on the learners' errors.

One of the studied research questions deals with the strategies used by the students when they interact with copex-chimie. A first trial is performed with 126 students in 2010 with eight different conditions of copex-chimie. Students work individually for around 90 min with copex-chimie. The teacher briefly presents the work to be done and then lets the students work independently. Based on the results of the first trial, another trial is performed in 2013 with 128 students and is limited to 3 different conditions of copex-chimie. The study protocol was co-designed by the TEL researchers and the specialist in data production and analysis.

5.2. DATA PRODUCED

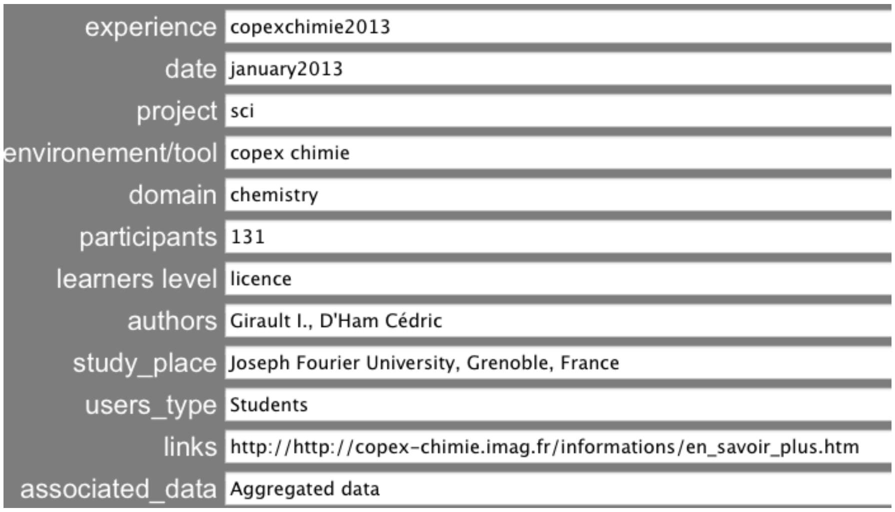
During the trials, event data are produced (Figure 6). They contain the students' activities on the software. Five variables are stored: the time stamp, the user code (id_user) that corresponds to one of the characteristics of the agents, the actions (action) made by the student, the action features (params) that adds information about the action, and the score (note) for students who ask for an evaluation.



timestamp	id_user	action	params	note
1/10/13 10:38:35	1635	modif_ordre	0-0--2--3--4--5--6--7-0-0-11-8--1-9--1-10--1-0-0-0-0-0	
1/10/13 10:38:35	1635	ajout_action_rincage	12/fiole jauge/eau/ /1	
1/10/13 10:38:57	1635	modif_ordre	0-0--2--3--4--5--6--7-0-0-11-8--1-12-9--1-10--1-0-0-0-0-0	
1/10/13 10:38:57	1635	ajout_action_rincage	13/fiole jauge/eau/ /1	
1/10/13 10:39:11	1635	modif_ordre	0-0--2--3--4--5--6--7-0-0-11-8--1-12-9--1-13-10--1-0-0-0-0-0	
1/10/13 10:39:11	1635	ajout_action_homogeneisation	14//1/ /1	
1/10/13 10:39:25	1635	modif_ordre	0-0--2--3--4--5--6--7-0-0-11-8--1-14-12-9--1-13-10--1-0-0-0-0-0	
1/10/13 10:39:25	1635	ajout_action_homogeneisation	15//2/ /1	
1/10/13 10:39:36	1635	modif_ordre	0-0--2--3--4--5--6--7-0-0-11-8--1-14-12-9--1-15-13-10--1-0-0-0-0-0	
1/10/13 10:39:36	1635	ajout_action_homogeneisation	16//3/ /1	
1/10/13 10:39:41	1635	demande_eval	3 61 411	9
1/10/13 10:39:55	1635	consult_cours	produits chimiques	
1/10/13 10:40:14	1635	consult_cours	produits chimiques	
1/10/13 10:40:34	1635	demande_eval	3 61 411	9

Figure 6: Screenshot of the UTP software, showing event data for the copex-chimie study in 2013.

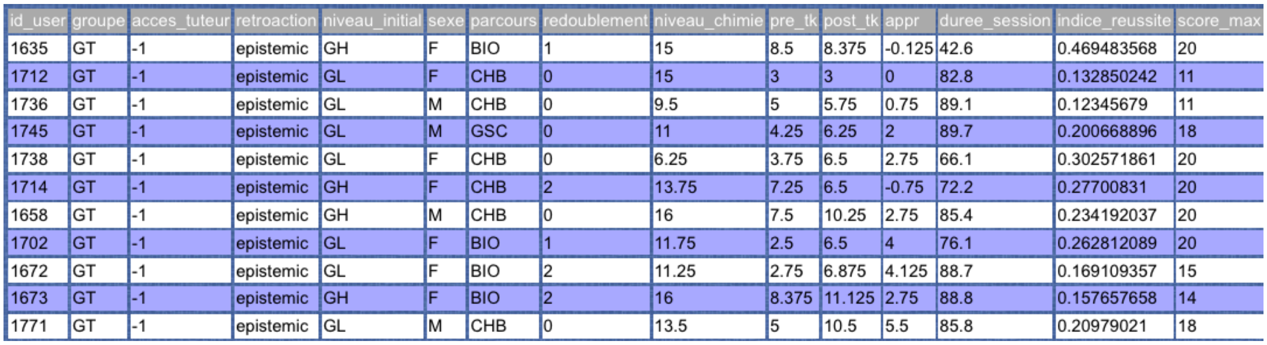
The study description (Figure 7) contains information to identify and describe the study protocol. The experimental protocol is stored as a pdf file.



experience	copexchimie2013
date	january2013
project	sci
environement/tool	copex chimie
domain	chemistry
participants	131
learners level	licence
authors	Girault I., D'Ham Cédric
study_place	Joseph Fourier University, Grenoble, France
users_type	Students
links	http://http://copex-chimie.imag.fr/informations/en_savoir_plus.htm
associated_data	Aggregated data

Figure 7: Screenshot of the UTP software with the description of the study for the copex-chimie study in 2013.

The agent data (figure 8) correspond to personal information about the students who participate to the study. We can classify these data in six groups, (1) the identification of the user (id_user), (2) the personal information (sex), (3) the curriculum information about the students (level in a chemistry course, curriculum, pre-test), (4) the experimental group (group, tutor access, type of feedback), (5) the post-test result and the learning gain (learning: difference between post-test and pre-test) and (6) the results obtained with the software (duration of work, success score and rate of success).



id_user	groupe	acces_tuteur	retroaction	niveau_initial	sexe	parcours	redoublement	niveau_chimie	pre_tk	post_tk	appr	duree_session	indice_reussite	score_max
1635	GT	-1	epistemic	GH	F	BIO	1	15	8.5	8.375	-0.125	42.6	0.469483568	20
1712	GT	-1	epistemic	GL	F	CHB	0	15	3	3	0	82.8	0.132850242	11
1736	GT	-1	epistemic	GL	M	CHB	0	9.5	5	5.75	0.75	89.1	0.12345679	11
1745	GT	-1	epistemic	GL	M	GSC	0	11	4.25	6.25	2	89.7	0.200668896	18
1738	GT	-1	epistemic	GL	F	CHB	0	6.25	3.75	6.5	2.75	66.1	0.302571861	20
1714	GT	-1	epistemic	GH	F	CHB	2	13.75	7.25	6.5	-0.75	72.2	0.27700831	20
1658	GT	-1	epistemic	GH	M	CHB	0	16	7.5	10.25	2.75	85.4	0.234192037	20
1702	GT	-1	epistemic	GL	F	BIO	1	11.75	2.5	6.5	4	76.1	0.262812089	20
1672	GT	-1	epistemic	GL	F	BIO	2	11.25	2.75	6.875	4.125	88.7	0.169109357	15
1673	GT	-1	epistemic	GH	F	BIO	2	16	8.375	11.125	2.75	88.8	0.157657658	14
1771	GT	-1	epistemic	GL	M	CHB	0	13.5	5	10.5	5.5	85.8	0.20979021	18

Figure 8: Screenshot of the UTP software, with the characteristics of the agent for the copex-chimie study in 2013.

5.3. SEQUENCE OF TASKS TO ENRICH DATA

To test our framework, we focus on the data enrichment of the data stored during the copex-chimie study. In this study, the purpose of the enrichment is the production of relevant activity patterns. An activity pattern is described by a sequence of some specific actions. The activity patterns allow the identification of the behaviors, and more particularly the students' strategy.

First, all the patterns of two consecutive actions are explored. We want to identify the patterns that both have an important occurrence and can help us answer our research questions. Then, we cumulate the patterns that correspond to similar strategies regarding our research questions. We select patterns that inform us on what is the action performed by students after they ask for the tutor feedback. Among the actions performed, two different strategies are highlighted: either the students explore the information given by the tutor (we cumulate the patterns "tutor request" followed by a detailed evaluation or by the consultation of a lesson suggested by the tutor) or they modify their procedure (the second action is a modification, an addition or a cancellation of what is written).

We can either give a total number of the selected cumulated patterns (a quantitative analysis) or represent the sequences of the selected cumulated patterns with a visualization tool, which corresponds to a more qualitative analysis. This visualization can show the individual variability among the students of the same experiment and the distribution of the patterns over the time.

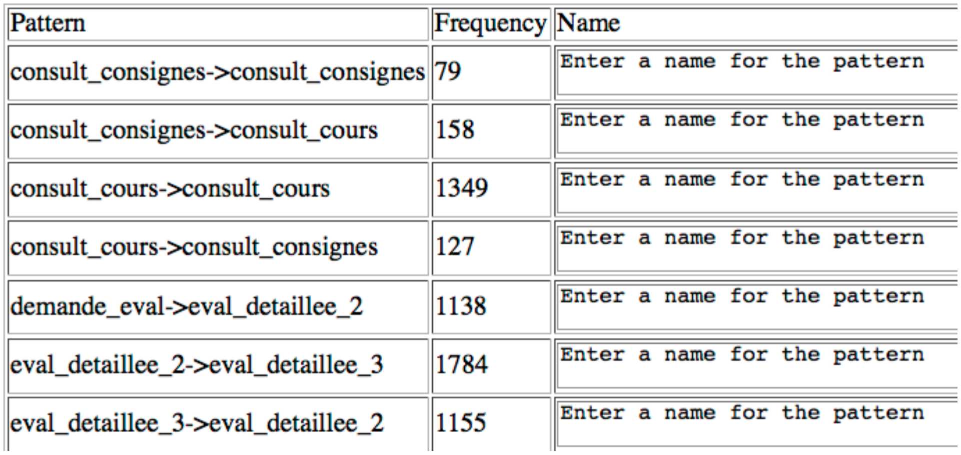
In order to study the influence of the scaffolding conditions on the students' strategies, we study the cumulated patterns per experimental conditions. All this enrichment process helped the TEL researchers to better specify the research questions.

Table 3 details the tasks performed during the enrichment of data described previously, according to the DC_TEL framework.

Step	Enrich data					
Task	Create patterns of (two) consecutive actions	Count patterns	Select interesting patterns for the study	Cumulate patterns	Count cumulated patterns	Visualize cumulated patterns
Objective	Design and create new	Design and execute	Design and create new	Design and create	Design and execute	Design and

	data	data treatments	data	new data	data treatments	execute data treatments
Category of operators	Data management	Descriptive statistics	DHJ	DHJ	Descriptive statistics	VDA
Expected results	Several patterns created	Number of patterns of each type	List of selected patterns	List of cumulated and renamed patterns	Number of renamed patterns of each type	Time line of patterns
Operator (available on UTA)	UT_pattern*: Create patterns automatically with two parameters: list of relevant actions and length of patterns (two)	UT count*: Count the patterns	UT_pattern_rename*: List the patterns, TR selects and names the relevant patterns (figure 9 see results of this operator)	UT_pattern*: Create new patterns with two parameters : list of relevant patterns and length of patterns (two)	UT count: Count the new patterns	UT_visualization*: Draw the time line with the relevant patterns
Study	Copex-chimie 2010					
Type of data	Event data					
Status of data	Raw data	Enriched data	Enriched data	Enriched data	Enriched data	Enriched data
Observables/ variables	Actions	Patterns	Patterns	Patterns	Patterns	Patterns Time stamp */student*
Quality indicator	Relevance	Accuracy	Relevance	Ease of interpretation	Accuracy	Ease of interpretation

Table 3: The sequence of tasks and attributes to enrich the data of the data produced during the copex-chimie study.



Pattern	Frequency	Name
consult_consignes->consult_consignes	79	Enter a name for the pattern
consult_consignes->consult_cours	158	Enter a name for the pattern
consult_cours->consult_cours	1349	Enter a name for the pattern
consult_cours->consult_consignes	127	Enter a name for the pattern
demande_eval->eval_detaillee_2	1138	Enter a name for the pattern
eval_detaillee_2->eval_detaillee_3	1784	Enter a name for the pattern
eval_detaillee_3->eval_detaillee_2	1155	Enter a name for the pattern

Figure 9: Screenshot of the UTA software, with an extract of the patterns list created on the study copex-chimie, the right column allow to researcher to name the relevant patterns. These names are stored in the new variable; it can be used in another analysis like visualizations or statistic tests.

5.4. TEL RESEARCHER AND DATA ANALYST COLLABORATION AROUND DC_TEL

With this test, we can see that the TEL researcher was not able to inquire all of the attributes. The help of the data analyst is always necessary. However, the framework allows the TEL researcher to initiate the data analysis with several elements, which enhance the collaboration with the data analyst.

The TEL researcher:

- writes the scenario of analysis
- cuts the scenario of analysis in several tasks
- establishes a sequence of tasks
- identifies the category of operators to lead the tasks. In our test, the data management was not identified.
- identifies the useful variables to lead the tasks. In our test some variables are not identified as necessary (e.g. time stamp, user).
- identifies three data quality indicators, the temporal precision was not identified.

The data analyst:

- identifies the needs to manage the data.
- identifies the operators to lead the tasks, this activity needs some discussion with the Tel researcher to precise the expected results.
- if the operators do not exist the data analyst must create a new data mining operator.
- adds the variables necessary in relation to the variables necessary to implement the operator.
- adds the quality indicators not identified by the Tel researcher.

6. CONCLUSIONS AND FUTURE WORKS WITH DC_TEL

This paper proposed the DC_TEL framework for assisting the data production and analysis, especially for the collaboration between the Tel researcher and the data analyst to control the data production and to implement the educational data mining tools.

The data life cycle proposed in the DC_TEL allows the identification of the 6 steps, including the steps “Validate” and “Enrich” not really specified in the literature. This proposition may be a first stage to address the question about the importance of the pre-processing indicated by Romero et al. The decomposition of these steps in objectives, tasks, and the list of the attributes allows the TEL researcher to initiate the scenario of analysis before to enhance the collaboration with the data analyst. During the instantiation with the study copex-chimie, we observed that the DC_TEL can support this collaboration. The multiplicity of tasks by objectives allows the combination of the data mining operators, like recommended by [Baker 2010]. Indeed for each objective, several tasks can be implemented. Each of them integrates by only one operator. The consideration of the data quality indicators, rarely mentioned in the EDM, in our knowledge, is a good way to control the validity of the data and the results.

In this paper, we have provided a first version of DC_TEL to enhance the collaboration between the specialist in data production and analysis and the researcher in TEL. However, there are some limitations.

We have tested the decomposition in objectives, tasks and attribute of tasks in the step "validate" (non presented here) and the step "enrich". We must tested this decomposition for the others steps. We must provide some tools more convenient, like a questionnaire, to guide the TEL researcher to inquire the step, objectives, tasks and attributes. A work on the convenient and usable description of the operators must be conducted too.

About the indicators of the data quality, we must explore if the selected indicators are also useful during the steps "design" and "production". The work on the data quality indicators must be continued, to better specify them for the data in the educational context. This work can contribute to create a new category of operators, which allow the creation of operators to compute the data quality indicators, and thus to enrich the category of tools provided by EDM.

This work is a first stage; we must continue our research to take into account these three axes in the TEL research context. Also, the DC_TEL framework must be improved and investigated depending on each actor, which needs to produce and use data, in this multidisciplinary field.

7. SUPPORTING MATERIALS

Data, operators and the chains of tasks used are available on the undertracks.imag.fr. An account must be created on the Undertracks platform. After the registration, the data, the operators and the UTA software can be download. The chain of tasks presented in this paper can be asked to Nadine.Mandran@imag.fr.

8. REFERENCES

BAKER, R. AND YACEF, K., 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), pp.3–17.

BAKER, R.S., 2010. Mining data for student models. In *Advances in intelligent tutoring systems*. Springer, pp. 323–337.

BAKER, R.SJ. AND DE CARVALHO, A., 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining*. pp. 38–47.

BERTI-EQUILLE, L., 2007. *Quality awareness for managing and mining data*. Habilitation à diriger des recherches. University Rennes 1.

BIENKOWSKI, M., FENG, M. AND MEANS, B., 2012. *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics, Department of Education's, Office of Educational Technology* Center for technology in Learning, SRI international., U.S. Departement of education.

BISHOP, L., 2012. Archiving your data: planning and managing the process.

BOUHINEAU, D., LALLE, S., ET AL., 2013. Share data treatment and analysis processes inTechnology enhanced learning. In *Workshop Data Analysis and Interpretation for Learning Environments*.

BOUHINEAU, D., LUENGO, V., MANDRAN, N., BEN-MASSON, T., ORTEGA, M. AND WAJEMAN, C., 2013. Open platform to model and capture experimental data in Technology enhanced learning systems. *Alpine, Rendez-Vous*.

BRASSEUR, C., 2005. *Data Management: Qualité Des Données et Compétivité*, Hermes Science Publications.

DESMARAIS, M.C. AND LEMIEUX, F., 2013. Clustering and visualizing study state sequences. In *Proceedings of 6th International Conference on Educational Data Mining*. Educational Data Mining. pp. 224–227.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G., SMYTH, P. AND UTHURUSAMY, R., 1996. Advances in knowledge discovery and data mining. *AI Magazine*, 17(3), pp.37–53.

GIRAULT, I. AND D' HAM, C., 2014. Scaffolding a Complex Task of Experimental Design in Chemistry with a Computer Environment. *Journal of Science Education and Technology*, 23(4), pp.514–526.

GOBERT, J.D., SAO PEDRO, M., RAZIUDDIN, J. AND BAKER, R.S., 2013. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4), pp.521–563.

HAUG,, A., ZACHARIASSEN, F. AND VAN LIEMPD, D., 2011. The costs of the poor data quality. *Journal of Industrial Engineering And Management*, pp.168–193.

HOWELL, D.C., ROGIER, M., YZERBYT, V. AND BESTGEN, Y., 2007. Statistical Methods in Human Sciences. *De Boeck*.

IKSAL, S., 2012. *Ingénierie de l'observation basée sur la prescription en EIAH*. habilitaton à diriger des recherches. Université du Maine.

JOHNSON,L., LEVINE, A., SMITH, R. AND STONE, S., 2010. *The 2010 Horizon Report*, The New Media Consortium, Austin, Texas.

MCCANDLESS, D. AND CUNÉO, D., 2011. *Datavision: mille et une informations essentielles et dérisoires à comprendre en un clin d'oeil*, [Paris]: R. Laffont.

PETERSON, P., BAKER, E.L. AND MCGAW, B., 2010. International encyclopedia of education.

ROMERO, C., ROMERO, J.R. AND VENTURA, S., 2014. A Survey on Pre-Processing Educational Data. In *Educational Data Mining*. Springer, pp. 29–64.

ROSLING, H., 2009. Gapminder. *GapMinder Foundation http://www. gapminder. org*.

DI RUOCCO, N., SCHEIWILER, JEAN-M. AND SOTNYKOVA, A., 2012. La qualité des données : concepts de base et techniques d'amélioration. In *La qualité et la gouvernance des données*. Série Informatique et SI. Cachan: Lavoisier, pp. 25–55.

SAO PEDRO, M., DE BAKER, R.S.J., MONTALVO, O., NAKAMA, A. AND GOBERT, J.D., 2010. Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. In *EDM*. ERIC, pp. 181–190.

SIEMENS, G. AND BAKER, R.S., 2012. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM, pp. 252–254.

UK.DATA ARCHIVE, http://www.data-archive.ac.uk/, last consultation July 2014.