DOP8_Qpp: Model to pre-process educational data

Nadine Mandran, Laura Dupuis, Vanda Luengo

DOP8_Qpp: Model to pre-process educational data

Nadine Mandran
Lab. d'Informatique de Grenoble
BP 46 - 38402 St Martin d'Hères Cedex
+33 476 574 896
nadine.mandran@imag.fr

Laura Dupuis
Lab. d'Informatique de Grenoble
- 38402 St Martin d'Hères Cedex
+33 476 574 896
laure.dupuis@imag.fr

Vanda Luengo
LIP6
4 Place de Jussieu
75005 Paris
+33 476 574 896
vanda.luengo@lip6.fr

ABSTRACT
This paper addresses problem of reuse of data to help researchers in Technology Enhanced Learning (TEL). The article describes a process to take over data that come from educational context. This process to reuse TEL data contains: major tasks, identified data properties and a set of quality criteria to reach these data properties. The objective of the process is to evaluate if data are reusable to serve learning analytics. This process is integrated in an existing data life cycle DOP8. The model was elaborated from several works with set of data since 2012 and from interviews with data-scientists. Also, we have administrated an on-line survey with data-scientists to evaluate feasibility of our proposal.

Keywords
data quality, data properties, pre-processing, reusing data, dashboard, data validation

1. INTRODUCTION
This paper proposes one process to evaluate if data that come from educational context are reusable by research purposes in TEL domain.

Nowadays, masses of data are produced when students use learning platforms, when teachers produce dashboards of scores to follow up their students and when knowledge of these students are assessed through several tools such as a Multiple Choice Quiz (MCQ). These ecological data built by and for teachers in the field are valuable. They are crucial (1) to well understand the teachers' practices (2) to understand teaching and learning phenomena and (3) to provide valuable indicators to the TEL designers. For example, Márquez-Vera, C. & al [8] use ecological data to predict student failures. In their case they use information about high school students enrolled on Program II of the Academic Unit Preparation at the Autonomous University of Zacatecas (UAPAZ) for the 2009/10 academic year. All the information used in this study has been gathered from three different sources during the period from August to December 2010. Their dataset has 77 attributes, variables, or features for each student. The comprehensive dataset describes socioeconomics factors, personal factors and final scores from several knowledge domains (mathematic, physics, writing and reading Spanish, etc.).

Whether the data should be used for research purposes, it is necessary to ensure a good level of data quality. To achieve this, it is important to have criteria to know if data are reusable. Reusing data is difficult because data that come from educational field are not structured by a predetermined guideline, metadata to describe data production are not always available. Although Haugh & al [6] indicate that poor data quality generates loss of data and also bad analysis, when data are produced without validation. So, before to analyse data it is useful to pre-process them. However, when data come from field, it is more difficult to pre-process them because researcher needs to interact with data producer, to understand how and why data have been produced. To pre-process this kind of data is important to identify some criteria to know if data are acceptable to lead data analysis. Several criteria are proposed by Berti Equille & Di Ruocco [1, 4] to ensure and assess data quality, such as the relevance, the ease of interpretation, the temporality coherence. These criteria are generic and there are not processes or tasks to reach them in the TEL context.

Our proposition is DOP8_Qpp: a model of process to shape decision making about data quality. It includes process with major tasks, data properties and set of quality criteria.

We evaluate our proposition by two ways. First, we instantiate our contribution using data come from MCQs realized in the first year in medical education in Grenoble, France (1800 students). Secondly, we have asked opinion of specialists with an online questionnaire to evaluate quality criteria and process.

First, paper presents related works about pre-processing of data and data quality criteria. Then, we described our proposal. Last part presents evaluation of our proposal.

2. RELATED WORK
2.1 Pre-processing steps in data mining
Lot of works focus on data processing and they propose several data life cycles [2][14][5][12]. Several of them contains pre-processing steps precisely described. This first step is data transformation into an appropriate form for data analysis process. This step requires an important manual-work; it consumes 60-90% of the time, efforts and resources employed in the whole knowledge discovery process [9]. Romero & al [12] split the pre-process in 6 steps after data gathering: Data aggregation, data cleaning, user and session identification, attribute selection, data filtering and data transformation.

DOP8 proposed by Mandran & al [7] combined the data life cycle and operator life cycle to lead data analysis. In this scheme, pre-processing is composed with two steps: "Validate" to control data in relation to the reality and "Enrich" to enhance the meaning of data with the creation from new variables. If Romero et al. [12] define more precisely data pre-processing, arrangement of these steps could be a problem. For instance, data aggregation before data cleaning, without non-validated data, could introduce mistakes during data aggregation step. Moreover, there is no

indicator to control validity of data and coherence of data files during data pre-processing. Regarding Mandran et al. [7], DOP8 was setup to follow data produced in a research context. That is, once research issue is defined, an experimental protocol is designed to lead data production. Thus, researcher controls data production. In DOP8, such as Romero et al. [12], there is no means to take into account data gathered outside research context and to qualify them in order to reuse these data.

2.2 Quality approach and data quality

Quality approach is process and tools to follow tasks which are needful to manage a project. During process, «actual results of an action are compared with a target or a set point »[13]. Brasseur [3] gives some features about data quality: "Data quality can address the needs of its users", "Data quality is dependent on their use", "The understanding of user needs is a prerequisite for defining and obtaining data quality requirement", "A big difficulty is that bad data quality is not easily detected. There are often some incidents or abnormalities during operational work, which reveal, here and there, inconsistencies on data." Therefore it is important to acquire methods and tools to control data throughout data life cycle.

In information systems, data quality proposes four approaches dedicated to improving data quality before analysis step: (1) preventive, (2) adaptive, (3) corrective and (4) diagnosed approaches [1]. Preventive approach allows upstream control before production. It is based on quality of the model and on 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.

Beside, [4] quote 10 data quality indicators:

- 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. Completeness is split in 4: entities, attributes, relations and occurrences.
- 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.

Polańska and Zyznarski [11] elaborate a framework to guide the quality of the data process. They propose measures to follow the data process. They describe them with "name, purpose, measurement method, type of the method, scale, type of scale, unit of measurements".

These works about data quality are useful to lead a data re-engineering. In the case of the data produced by teachers on the field, only the corrective and diagnosed approaches are possible, since data are produced before. Definitions proposed by Di Ruocco & al [7] are generic and all are not adapted in our case. Polańska and Zyznarski [15] provide a framework to describe and to adapt the data quality criteria. This framework is interesting but we have to adapt it. Indeed it only takes into account quantitative measures to produce quality criteria. But to understand why and how the data are produced and to lead a data re-engineering, it is crucial to interact with data producer; it is a qualitative measure.

To conclude on this point, data produced in real conditions by and for teachers are valuable and especially to lead a research in the context of the design based research methodology[15]. However, data life cycles do not clearly mention difference to take into account between data produced in a research context and out of it. Moreover, in data life cycles there are no data quality criteria to guide steps of these cycles and to validate data. Thus, our proposal has two targets: 1-Defining a process and tasks to help researcher to take up data and to co-operate with data producer and 2-Assigning a properties to data to qualify them and to ensure their reusability. Next section describes our proposal.

3. DOP8_Qpp

The name DOP8_Qpp means: data Quality indicators to lead Pre-Processing steps into DOP8 cycle. It help researcher whose needs ecological data produced on the field. In this article, we don't take into account data produced by researcher to address a research problem. Indeed, in this context, data production is controlled with a protocol. To lead our model with ecological data, it is necessary to distinguish two actors: 1- Data producer (DP) that produces data into an educational context and the Researcher-Analyst (RA), which needs to use data in another context (e.g. students marks produced by a teacher and data analyses lead by a researcher to find a model to describe students' learning evolution).

Also two objects are useful: 1- Structure of Data (SD) which is the physical organization into an information system (e.g. data files, directory, folders, database, etc.) and 2- Data (D) which are a set of values that can be stored and used by a software.

DOP8_Qpp is described at figure 10. Before to explain it, we explain definition and concepts used, the process and dashboard concept.

3.1 Definition and concepts

Three elements are presented below: 1- Collect and pre-processing steps in DOP8 to reuse data, 2- UCUA Data properties, and 3- Quality criteria in relation with data properties.

3.1.1 Collect & Pre-processing in DOP8

DOP8 cycle begins by "Prepare", step where a protocol is designed to produce data. In this case, RA makes research questions and produces data to address these issues. After, it is step "Collect", where data are produced in an experiment field. Then, there are two pre-processing steps. First step "Validate" which goal is a validation of raw data to ensure coherence and relevance of SD and D. During this step, several major tasks are lead to control SD and D in link to the reality. Second step "Enrich" which goal is enrichment of data. During this step, several majors tasks are lead to add new data. The enrichment allows ease of understanding data, their analysis and their use.

3.1.2 Four data properties: UCUA

We propose to use four data properties.

- **Utility:** SD and D satisfy of user's needs now and in the future.
- **Compliance**: conformity between metadata provided by producer with SD and D.
- **Usability:** ability of SD to be used by a data analysis software and ability of D to enhance meaning of results.
- **Acceptability**: is synthesis of the previous three properties. If they are reached, SD and D are acceptable to lead data analysis.

Process (Figure 2) to control these three data properties is organized as follows: first "Utility" of SD and D are validated, then "Compliance " and "Usability" of SD, then "Compliance and Usability" of D. It is organized in this manner because it is necessary to collect metadata about SD and D and to meet RA's needs first, then to control data structure to provide exploitable structures by automated tools and finally to validate data and transform them to enrich data semantics. At the end, data can be qualified with "acceptability" or not.

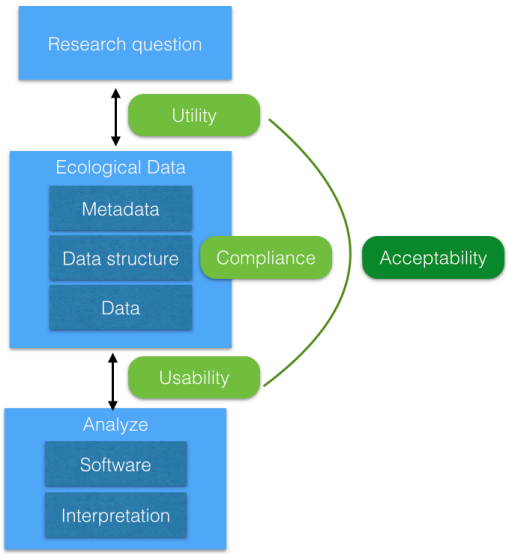

Figure 2: Description of the model with four data properties and place in the process.

3.1.3 Data properties and quality criteria

For each data properties, we use one or several quality criteria coming from data quality works [4].

- Utility of data and structure of data is controlled with criteria 'Relevance'. Compliance of structure of data is controlled with: 'Conformity', 'Completeness of entity', 'Completeness of attribute' 'Completeness of relation'
- Compliance of data is controlled with 'Accuracy', 'Uniqueness', 'Temporal precision', 'Consistency'
- Usability of structure of data is controlled with "structural consistency".
- Usability of data is controlled with: 'Completeness of occurrences', 'Ease of interpretation'.

3.2 Operationalization

Quality criteria are a result of one or several major tasks, which produce values of measures. These values give making decision rules to qualify SD and D based in data properties. Measures are qualitative and quantitative. Qualitative obtained by qualitative methods (e.g. interview with producer to gather metadata). Quantitative, obtained by an automatic process of data (e.g. ratio of missing values). Sequence of major tasks creates a process to qualify object (SD and D). This process helps to qualify data with the four UCUA properties.

Global process has two starting points: designing a research question and existence of a data produced in an educational context. To reach each property, major tasks allow production of quality criteria, then RA can takes a decision and qualify objects (SD and D). One or several dashboards must be designed to follow the process.

Now, we detail process for each properties and each object. To exemplify, our proposal we use a research question which is "There are a model which explain students' evolutions during two semesters in mathematics and biology?" and we have data produced by university during several years.

Utility of SD et D: main goal of tasks to reach first property "Utility" is to identify if all useful information has been provided by producer. That corresponds to data quality "relevance" of data. RA needs information about structure of data and data. RA must achieve two iterative majors tasks: 1- interviewing DP to understand data and to obtain information about data 2- controlling if information is sufficient. If initial information is deficient, additional information must be requested to DP. A new interview can be made.

A dashboard is set up to record this information, such as a checklist where RA checks every needful element. In this part of process, quality criteria are set up with qualitative information. These information provided by data producer set up these data quality criteria and ensure validity of SD and D.

If all quality criteria are satisfied, data are qualified by "Utility"; else, the dimension of "Utility" is not reached. Data can not be used to address the research question.

About our example, RA needs data to identify: students, subjects taught but only mathematics and biology, date and scores of students. RA needs this minimal set of information to try to address research question. University gives data to RA with description of SD and description of D. Both descriptions exist and minimal set of information exists then the dimension "Utility" of SD and D is reached.

Compliance and Usability of SD: To reach compliance and usability of SD, RA needs to realize 6 majors tasks: 1- controlling conformity between SD and description of SD provided by producer, 2-controlling if all data structures exist, 3- controlling if relation between data structure are usable, 4- controlling if all variables exist in data structures, 5- transforming data structure into a new one if necessary (e.g. aggregation) and 6- controlling that data structure can be used with one of data analysis software. If information is deficient to assume the 6 tasks, additional information must be requested to data producer. Quality criteria used are 'Conformity', 'Completeness of entity', 'Completeness of attribute', 'Completeness of relation' and 'Structural consistency'. The five quality criteria are set up with qualitative information. At the end, SD is qualified with "compliance" and "Usability" dimensions. Process can be continued to qualified D with "compliance" and "Usability"

dimensions. Otherwise, SD can not be used to address research question, process can not continue.

Compliance and Usability of D: Main goal of tasks to reach compliance and usability of D is validated value of data and enrich them. Quality criteria used are 'Accuracy', 'Uniqueness', 'Temporal precision', 'Consistency', 'Completeness of occurrences', 'Ease of interpretation'. RA needs to realize 8 majors tasks: 1-controls if values of each variable match with the description of coding scheme provided by producer, 2- controls that all records into data files are unique, 3- controls that there are sufficient data to ensure data temporality, 4- controls that data are not in contradiction with the others, 5- calculates number of occurrences, 6- calculates number of missing values, 7- controls that data are in relation to constraints of data analyze methods, 8- creates new variables to enrich data, 9- controls that enrichment are sufficient. At the end, D is qualified with "Compliance" and "Usability" dimensions. Now, a first data analysis can be led. Data are qualified with properties "Acceptability".

In our example about students' evolution, one set of tasks can be: controlling that scores of students is between 0 to 20 controlling that several scores exist by students and controlling that ratio of missing values are less than 5%.

Nevertheless, one major difficulty still remained is the monitoring of these processes which combine major tasks, data properties and quality criteria. We propose to use dashboards to address these difficulties.

3.3 Dashboards

Advantage of dashboards is that traces of data pre-processing are kept. Thus, other RAs can reuse these data with a high level of trust. Dashboards are essential into a data quality process. To follow the pre-process, in addition to major tasks, is needful to add 3 elements in dashboard: 1-sub-tasks to refine major tasks, 2-type of measurement method to create quality criteria, and 3-decision criteria.

Major tasks are steps of process, when RA needs to reuse data; (s)he defines several **sub-tasks** that are needful to take over data structure and data. Sub-tasks are defined according to research question and type of datasets chosen. (See example in table 2). **Type of measurement method** indicates type of information produced by sub-tasks. We have adapted a typology proposed by Polańska and Zyznarski [15]: **objective** measures are produced in an automated way (such as ratio of missing values); **subjective** measures are produced by a human (such as a list of available data, or description of data production gathered in an interview with data producer). This distinction is important because an algorithm directly calculates objective measures whereas subjective measures can not be calculated by an automated way.

Table 2: Example of sub-tasks for the major task: "control that all information to pick up data are available"`

Major tasks	Sub-tasks
To control that all information to pick up data are available.	Interviewing produced
	Listing data useful to address research question
	Listing data available in data set
	Obtaining metadata

Decision criteria indicate threshold that is accepted by RA to qualify data and to reuse them. It is fixed before to pick up data, (for instance RA fixed maximum value of missing value at 10%). Decision criteria are logical for subjective type of measurement method and numerical for objective type of measurement method. They validate quality criteria. If all quality criteria are validated, data properties are reached. Thus, data structure or data can be qualify or not with data properties. Dashboards proposed contains: Majors tasks, Sub-tasks, Type of measurement, Decision criteria, data quality criteria and appraisal of data properties**.** Sub-tasks, type of measurement, decision criteria are defined by RA in relation to research question and data that come from field. A first list of sub-tasks, type of measurement, decision criteria must be describe before picking-up data. This list can be completed and modified during process. Next tables present 5 examples of dashboards: 3.1 to reach Utility of SD & D, 3.2 to reach Compliance of SD, 3.3 Compliance of D and 3.4 Usability of data.

Differences between these dashboards are major tasks, sub-tasks and quality criteria.

Table 3.1: Dashboard to monitor Utility of SD & D

UTILITY OF SD & D					
Majors tasks	**List of Sub-tasks**	**Type of measures**	**Decision criteria**	***relevance of SD***	***relevance of D***
Control that all information to pick up data files system are available and all variables to address student evolution progress are available too.					
Contrat that all variables to address research question are available too.					
			Appraisal of Utility	**yes/no**	**yes/no**

Table 3.2: Dashboard to monitor compliance of SD

COMPLIANCE OF SD						
Majors tasks	**List of Sub-tasks**	**Type of measures**	**Decision criteria**	***Conformity***	***Completeness of entity***	***Completeness of attribute***
control if SD matchs with information describing SD provided by the producer						
control if all needed data files exist and if conceptual data model exist or if it can be created.						
control if all useful variables are available in data files						
control if all relation between entities and data files are available.						
			Appraisal of compliance of SD	**yes/no**	**yes/no**	**yes/no**

Table 3.3: Dashboard to monitor Usability of D

USABILITY OF SD				
			Decision criteria	***Structural consistency***
control that datasets is usable in an automatic way with data analysis software				
			Appraisal of usability of SD	**yes/no**

Table 3.4: Dashboard to monitor Usability of D

USABILITY OF DATA					
Majors tasks	**List of Sub-tasks**	**Type of measures**	**Decision criteria**	*Completeness of occurrences*	*Ease of interpretation*
control that number of records are sufficient to be use in a datamining analysis					
Control that data are in adequation with the data analysis constraints (e.g. normality of distribution to use ANOVA)					
ensure that some new data are created to enhance intelligibility of data and therefore intelligibility of results.					
			Appraisal of usability of D	**yes/no**	**yes/no**

3.4 DOP8_Qpp enhancement of pre-processing in DOP8

With these different elements describe above, we enhance DOP8 to take into account data that come from educational fields without research protocol. It is enhanced with four data properties, a list of major tasks to qualify data, and a set of quality criteria. In figure 4, we present DOP8_Qpp; it contains specific steps to following data transformation, collecting existing data validating structure of data, transforming structure of data; validating values of data, validating data to sustain analysis and enriching data. Goal of each stage is to evaluate the data properties; a dashboard is produced at each stage. For example, when RA collects data, he uses dashboard described in figure 3.1 to reach criteria relevance. If these criteria are not reached, RA can follow (or not) recommendations made in purple boxes to enhance information to pre-process data. Thus, RA is guided during data pre-processing.

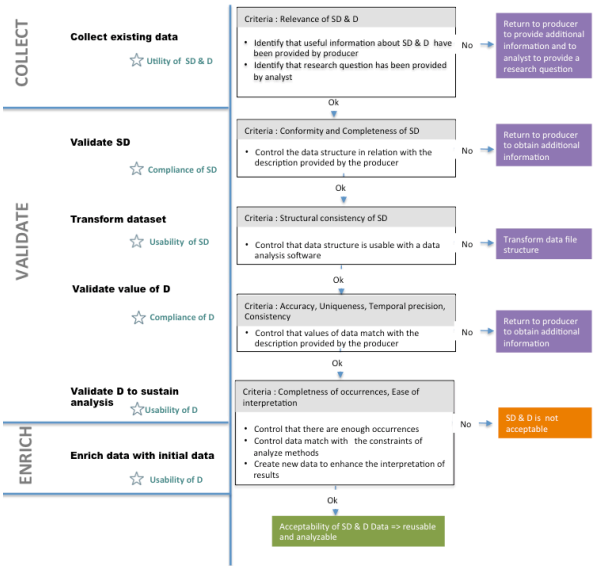

Figure 4: DOP8_qpp: enhancement of steps collect, validate and enrich with data properties and quality criteria.

4. EVALUATION

4.1 Case study : common first year for health studies in France (PACES)

We tested our proposal with data that come from common first year for health studies in France. This case study is based on a set of data produced in a flipped classroom pedagogical approach.

With two research questions about students' evolution and data produced by PACES, our proposal is evaluated by testing dashboards with data properties, major tasks quality criteria. We instantiate a list of sub-tasks with, for each one, a type of measurement and a decision criteria. Table 4 presents instantiation of dashboard to monitor "Utility of SD & D" for PACES data. In this case, sub-tasks provided only subjective type of measurement. In the column remarks RA can notice interactions with data producer and encountered difficulties. Comprehensive dashboards are available into a report "Process of data analysis in TEL research"[1].

Table 4: Instantiation of dashboard to monitor utility of SD & D to analyze students' evolution with data PACES

Majors tasks	**List of Sub-tasks**	**Type of measures**	**Decision criteria**	*relevance of SD*	*relevance of D*	**Remarks**
Control that all information to pick up data files system are available and all variables to address student evolution progress are available too.	Having meta data to know data structure	subjective	if answer = yes then relevance is reached	yes		
	Data file system provided by responsible of PACES matchs with the description of this data structure	subjective		yes		2 interviews were necessary to understand data structure and some exchanges by mail
	Data structure contains all data files needful to adress research question	subjective		yes		
Contrat that all variables to address research question are available too.	Having coding scheme of data	subjective	if answer = yes then relevance is reached		yes	we identify into research question data useful to address it, we list variabls into data file, and we compare. To make this comparison we need description about several variables
	Having description of students (gender, axe, curricula)	subjective			yes	
	Having at least one variable that indicates time, in this case it is semester and cycle	subjective			yes	
	Having varibles which contains results of QCM	subjective			yes	
	etc.					
			Appraisal of Utility	All criteria must be equal to yes		
				Yes	**Yes**	=> Data structure and data reach "Utility"

This Table is the final version, but each sub-task could be iterative. Also dashboard to follow values of objectives were proposed. For example, descriptive statistics with thresholds for data usability.

This process allow us to follow and share this time consuming pre-process steps for a first research question. They was shared for a second linked research question with other context and others researchers.

To go beyond it is necessary refine criteria: "ease of interpretation". For instance, to study students' evolution and obviously: Iterative measures during time, Evaluation of students, Iterative measures for each topics, Time indicator, Topics and student indicator are useful. Future works on

[1] French title « Processus d'analyse de données pour la recherche en EIAH » L.Dupuis. 2015.

these criteria must be conducted to understand ecological data without ambiguity. How create useful data with a real semantic? Which sub-tasks can be made to enhance meaning of data?

4.2 Evaluation of criteria and of process by 36 data scientists

First, interviews administrated with data-scientists confirm process, four data properties and quality criteria. To evaluate our proposal with a more data-scientist and data-analyst, a survey was designed. Survey cannot present directly process and these elements; it is too hard to answer directly about properties and quality criteria. So, the survey measures level of agreement about useful elements to take over data, tasks and tools that we have proposed. Finally, we have evaluated DOP8_Qpp about data pre-processing steps (results are not presented here). Chosen methodology is an on-line questionnaire with Likert-scale. 51 answers are collected, (36 data scientist and 15 data analyst). Even if total numbers of answers are less than 100, results of questionnaire are presented in percentage, to ease reading.

Useful element to take over data in a comprehensive manner (Table 5). More than 75% of respondents agree the nine elements shown in table 5. The first five elements are crucial, more than 90% of respondent accept them. Then, the other four are important too, more than 75% of respondent accept them. Among these elements, the three most important are "list and description of variables" (77%), "coding scheme of data" (75%) and "description of data file contain" (73%).

Table 5: Agreement about useful elements to take over data.

%	Strongly agree	Agree	**Total**
description of data file contains	73	24	**97**
Name and contact of producer	47	47	**94**
Data production protocol	70	22	**92**
List and description of variables	77	14	**91**
Conceptual data model	59	31	**91**
Coding scheme of data	75	14	**89**
Data structure	64	24	**88**
Research question	68	18	**85**
Names of data files	57	20	**77**

Useful tasks to take over SD & D (Table 7). About data structure, we propose 4 majors tasks. More than 90% of respondents agree the 4 tasks. About data, we propose 9 majors tasks. More than 90% of respondents agree the 9 tasks. Controlling uniqueness of data and describing variables are crucial (100% of respondents agreed).

Table 7: Agreement about of set of tasks to take over data.

Major tasks to qualify data structure	Strongly Agree	Agree	**Total**
Compare data structure with information provided by producer	69	25	**94**
Make a process to transform data files	57	37	**94**
List of errors encoutered during process of data files transformation	74	19	**93**
Control if new structure is usable with automated software	69	20	**89**
Major tasks to qualify data	Strongly Agree	Agree	Total
Control uniqueness of data	81	19	**100**
Describe new variables	65	35	**100**
Describe process used to create new variables	58	39	**97**
Control completeness of occurrences	67	30	**97**
Calculate ratio of missing values	71	23	**94**
Verify if variables of research question are all available into data	79	15	**94**
Create variables to enhance interpretation	61	33	**94**
Control adequation between coding scheme and code of data provided	66	28	**93**
Calculate number of occurrences	62	26	**88**

Mostly experts, more than 85% percent, agree on elements to take over data, actions and tools that we have proposed. Thus, these elements are essential to take reuse data by data-scientist. So, these elements, tasks and tools can companion RA when he takes over data, then he can qualify data that come from educational field. So, he evaluates data quality and one of positive impact is quality of results.

5. CONCLUSION

Using data produced by and for stakeholders such as teachers or students ensures relevance of the research results. Although, these data are valuable, it is difficult to reuse them without pre-processing steps and without properties and indicators to control data quality. Data quality research and data life cycle are combined here to target this problem and give means to refine the crucial pre-processing step. Thus important step more accessible to TEL domain. With respect of related work, our research brings process and concepts to take over ecological data. Add value of our proposal is four data properties Utility, Compliance, Usability and Acceptability of ecological data. Data quality indicators ensure that four dimensions are reached and then if data are reusable by other users, in other contexts. To elaborate data

quality criteria a list of majors tasks are provided. RA uses this list and defines news subtask to take into account ecological data, which he wants used. Also, our research shows need to monitor data validation and transformation. To do this we proposed a set of dashboards. Finally, our process help RA to generate sets of data that have been qualified with four properties and quality criteria. Once, educational data field have been validated they can be reused with data mining tools.

One instantiation shows the use of the process with data produced in a flipped classroom pedagogical approach. (1800 students, 12 scores for each students, and 12 MCQ tests). Right now, process and data properties are used with two other cases study (data come from MOOC: MOOCAZ [16], Serious games: TAMAGOCOURS [17]).

Positive impact of our proposal is the qualification of data with four data properties: Utility, Compliance, Usability to reach Acceptability of ecological data to address and resolve research questions. Another add values are majors' tasks and dashboards to target problem and assist this activity.

Beside, presented evaluation show the necessity to refine quality criteria and go beyond to support their application. Use sematic of data is one of the research tracks. For example, if we know that data represented "errors of students in several courses" we can apply reasoning methods to evaluate the utility of the data for a particular research question.

Also, when ecological data are qualified with four dimensions, we can ask question of positive impact of these data quality on quality of results and how to measure this quality? With which criteria?

Finally, in our proposal acceptability does not take into account social dimension, such as defined by [10]. In the case of student's evolution studies, one perspective is adding ethical criteria to control anonymization of data.

6. REFERENCES

[1] Berti-Equille, L. 2007. *Quality awareness for managing and mining data*. University Rennes 1.

[2] Bishop, L. 2012. Archiving your data: planning and managing the process. (2012).

[3] Brasseur, C. 2005. *Data Management: Qualité Des Données et Compétivité*. Hermes Science Publications.

[4] Di Ruocco, N., Scheiwiler, jean-M. and Sotnykova, A. 2012. La qualité des données : concepts de base et techniques d'amélioration. *La qualité et la gouvernance des données*. Lavoisier. 25–55.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. 1996. Advances in knowledge discovery and data mining. *AI Magazine*. 17, 3 (1996), 37–53.

[6] Haug, A., Zachariassen, F. and Van Liempd, D. 2011. The costs of the poor data quality. *Journal of Industrial Engineering And Management*.

[7] Mandran, N., Ortega, M., Luengo, V. and Bouhineau, D. 2015. DOP8: merging both data and analysis operators life cycles for technology enhanced learning. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (2015), 213–217.

[8] Márquez-Vera, C., Cano, A., Romero, C. and Ventura, S. 2013. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*. 38, 3 (2013), 315–330.

[9] MikSovskỳ, P., Matousek, K. and Kouba, Z. 2002. Data pre-processing support for data mining. *Systems, Man and Cybernetics, 2002 IEEE International Conference on* (2002), 4–pp.

[10] Nielsen, J. 1994. *Usability Engineering*. Elsevier.

[11] Polańska, J. and Zyznarski, M. 2009. *Elaboration of a method for comparison of Business Intelligence Systems which support data mining process*. School of Engineering. Box 520. Ronneby. Sweden.

[12] Romero, C., Romero, J.R. and Ventura, S. 2014. A Survey on Pre-Processing Educational Data. *Educational Data Mining*. Springer. 29–64.

[13] Sokovic, M., Pavletic, D. and Pipan, K.K. 2010. Quality improvement methodologies–PDCA cycle, RADAR matrix, DMAIC and DFSS. *Journal of Achievements in Materials and Manufacturing Engineering*. 43, 1 (2010), 476–483.

[14] Stamper, J.C., Koedinger, K.R., Baker, R.S.J. d, Skogsholm, A., Leber, B., Demi, S., Yu, S. and Spencer, D. 2011. Managing the Educational Dataset Lifecycle with DataShop. *Artificial Intelligence in Education*. G. Biswas, S. Bull, J. Kay, and A. Mitrovic, eds. Springer Berlin Heidelberg. 557–559.

[15] Wang, F. and Hannafin, M.J. 2005. Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*. 53, 4 (2005), 5–23.

[16] https://www.fun-mooc.fr/courses/ENSCachan/20002S04/session04/about

[17] http://eductice.ens-lyon.fr/EducTice/recherche/jeux/tamagocours