Collaborative and Distributed E-Research:

Innovations in Technologies, Strategies and Applications

Angel A. Juan
IN3 - Open University of Catalonia, Spain

Thanasis Daradoumis
University of the Aegean, Greece, & Open University of Catalonia, Spain

Meritxell Roca
IN3 - Open University of Catalonia, Spain

Scott E. Grasman
Rochester Institute of Technology, USA

Javier Faulin
Public University of Navarre, Spain

Information Science
REFERENCE

Managing Director:	Lindsay Johnston
Senior Editorial Director:	Heather Probst
Book Production Manager:	Sean Woznicki
Development Manager:	Joel Gamon
Development Editor:	Hannah Abelbeck
Acquisitions Editor:	Erika Gallagher
Typesetter:	Russell A. Spangler
Cover Design:	Nick Newcomer, Lisandro Gonzalez

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Collaborative and distributed e-research: innovations in technologies, strategies, and applications / Angel A. Juan ... [et al.], editors.
p. cm.
Includes bibliographical references and index.
Summary: "This book offers insight into practical and methodological issues related to collaborative e-research and furthers readers understanding of current and future trends in online research and the types of technologies involved"--Provided by publisher.
ISBN 978-1-4666-0125-3 (hardcover) -- ISBN 978-1-4666-0127-7 (print & perpetual access) 1. Internet research. 2. Group work in research. I. Juan, Angel A., 1972-
ZA4228.C65 2012
001.4'202854678--dc23
2011039614

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 6
Data Sharing in CSCR:
Towards In-Depth Long Term Collaboration

Christophe Reffay
Ecole Normale Supérieure de Cachan, France & Ecole Normale Supérieure de Lyon, France

Gregory Dyke
Carnegie Mellon University, USA

Marie-Laure Betbeder
Université de Franche-Comté, France

ABSTRACT

In this chapter, the authors show the importance of data in the research process and the potential benefit for communities to share research data. Although most of their references are taken from the fields of Computer Supported Collaborative Learning and Intelligent Tutoring Systems, they claim that their argument applies to any other field studying complex situations that need to be analyzed by different disciplines, methods, and instruments. The authors point out the evolution of scientific publication, especially its openness and the variety of its emerging forms. This leads them to propose corpora as boundary objects for various communities in the scientific sphere. Data release being itself a complex problem, the authors use the Mulce[1] *experience to show how sharable data can be built and made available. Once corpora are considered available, they discuss the potential of their reuse for multiple analyses or derivation. They focus on analytic representations and their combination with initial data or complementary analytic representations by presenting a tool named Tatiana. Finally, the authors propose their vision of data sharing in a world where scientists would use social network applications.*

DOI: 10.4018/978-1-4666-0125-3.ch006

INTRODUCTION

In the research process, data is crucial and often hard to collect. Researchers spend a lot of time designing studies and collecting, transforming, analyzing, or interpreting data. Once analyzed and communicated in some publication by a local research team, data is often lost and can't be re-used by anybody. This means that other researchers have no access to original data to deepen their understanding by replicating an analysis or comparing their own results on the same data with a slightly different analysis method.

In this chapter, we would like to draw the state of the art in data sharing among research communities and, in particular, to report the results of the Mulce project[1]. This project's main results are the design and creation of a data structure and a corresponding platform to share learning and teaching corpora. These results give the community a way to access, share, analyze and visualize learning and teaching corpora.

This work has been motivated by the lack of impact of research results in the real world of online learning. In the CSCL (Computer Supported Collaborative Learning) research for example, a very wide range of indicators on collaboration have been designed and prototyped in a particular context but almost none of them are reused in other situations or contexts. We argue in this chapter that our research community should be able to widen the validity of its results by sharing data, tools and analyses performed with these tools.

In their work on the coding and counting analysis methodology, Rourke, Anderson, Garrisson, and Archer (2001) have pointed out the weakness of our research domain. Replicability, reliability and objectivity need to be improved in our work. The main idea of research collaboration is well expressed by (Chan, et al., 2006) in the following terms:

"There is urgent need of putting together complementary strengths and contexts and combining our insights as rapidly as possible to make a greater impact and further elevate our research quality at the same time. Research generally has had a small voice in national educational outcomes; we can speak louder if we speak together." (Chan et al., 2006)

Considering e-Research as an efficient way to meet and collaborate, this chapter suggests that e-collaboration could provide emerging communities with tools and virtual places to actually share their data, analyses and results in order to improve their theories, knowledge and tools. Although the focus of our work is on CSCL, we argue that this proposal is not limited to this domain or even to its contributing disciplines, and that the core ideas and benefits of our proposal can be extrapolated to other fields of research.

In the remainder of the chapter, we first examine current trends in scientific publication and the central role played by data in the scientific process. We then highlight the particular problem posed by data collection and replication in learning-related research and examine the state of the art for data sharing within this context. The Mulce proposal for constructing and sharing learning and teaching corpora is presented in detail, followed by the Tatiana framework for creating and re-using analytic representations. We conclude by drawing up our vision of data sharing within the learning sciences field and describe how other fields can draw upon our experience to construct data and analysis sharing models of their own.

Evolution in Scientific Publication

Who is producing knowledge? Nowadays, this process is no longer limited to academic researchers and prestigious journals. Civil society including local, national and international organizations is bringing its truth in various areas like economy, education, environment science, etc. Forms of publication are also evolving from classical journal articles to virtual exhibitions, datasets, software tools, etc., or any combination of these elements allowing participants to find new spaces to further explain, demonstrate, or exemplify their theories with renewed modes of creativity.

For many reasons, open access is becoming the rule. Because of the unacceptable delay of release of articles in scientific journals, in comparison to the fast obsolescence of their results, physicists prefer the Open Archives Initiative. Gentil-Beccot, Mele, and Brooks (2009) shows that 97% of the publications used by the community of nuclear research scientists, were freely available as preprints. It also indicates that publications that are available as preprints are cited 5 times more than others, and that the citation peak occurs before the release of the journal publication. In the footprints of physicians, considering that knowledge, published in scientific journal, should be accessible for anybody in the world (including developing countries), many scientific communities have oriented part of their articles toward open access journals. The Directory of Open Access Journals (http://www.doaj.org/), created in 2002, counted (at the end of year 2002) 31 journals coming from 7 countries. In August 2011, the DOAJ counts 6920 indexed open journals coming from 112 countries. More recently, many publishers (either private or public) joined in the Open Access Scholarly Publishing Association (http://www.oaspa.org/), created in 2009 and organized their first conference in Lund (Sweden) in September 2009. In their discussions, we should mention the questions of economic models, transparency (scientific quality, reviewing process, metadata of their publications), impact factor, prestige, software tools (e.g. Open Journal Software) (PKP, 2010; Edgar & Willinsky, 2010), citations and references links persistence (e.g.: Digital Object Identifier) (Bilder, 2009) and the variety of publication types (exhibition spaces, datasets, books, articles, etc.). An ambitious project entitled “Sponsoring Consortium for Open Access Publishing in Particle Physics” (http://www.scoap3.org) presented by Mele (2009) is building a new economic model where publication costs would be endorsed by each country, according to the number of articles they submit. Considering the rate of scientific production in some domains (e.g. medicine), it is simply impossible for a single researcher or even for a well organized team, to keep up to date. Consequently, using semantic web techniques, a new format, namely “nano-publication” is suggested by Velterop (2009). Such a nano-publication would represent only the substrate of the published results in the form of RDF statements indexed in an open access database so that all researchers may be able to catch any new statement they are directly concerned with.

These few examples illustrate the speed, magnitude and depth at which the scientific publication process is evolving. We think that time has come for us to surf this wave in order to help our research community to share not only the articles and results, but also data and even analysis processes (methods or tools) that produce these results. The next two sections recall the central role of data in the scientific process and the importance for sciences that consider complex situations to share their data collections.

Data: At the Very Core of the Research Process

Most (if not all) areas of research involve a cycle with the following steps: define a research question, collect data, transform data in various ways, and produce statements which are the answer to the research question (Fisher & Sanderson, 1996). The epistemological framework within which research is conducted will define the kinds of question, data, analysis methodology and type of answer which are acceptable. For example, hypothetico-deductive research will typically require setting up a research question with competing hypotheses, collecting data in various conditions, performing statistical analysis and producing statements about robustness of results across conditions, or of causality/correlation between conditions and results. Certain fields of research (typically study of human behavior in authentic situations) have broad epistemological agreement that certain kinds of data can be collected with no research question in mind or that data collected with regard to a given research question is generic enough to be used to answer different questions (e.g. the Augmented Multi-Party Interaction [AMI] meeting Corpus [Carletta, 2007], telephone conversations [Godfrey, et al., 1992], etc.). Given that one of the major costs of research is data collection, it makes economical sense to exploit and share data as much as possible.

Furthermore, within collaborations at various levels (grad students and faculty supervisors, in the context of a laboratory or of a project, etc.), not only should data be shared, but sharing the various analytic representations created during data analysis can also be beneficial: better reliability can be achieved in subjective analysis if identical independent analyses are performed (De Wever, Schellens, Valcke, & Van Keer, 2006), researchers can collaborate to extend the applicability of an analytical method to a new domain of application (e.g. Lund, Prudhomme, & Cassier, 2007), researchers can spread the workload (Goodman, et al., 2006), or can combine the insights of several analysts (e.g. Prudhomme, Pourroy, & Lund, 2007). Last, in situations where different epistemologies are united around shared data, the lack of commensurability between methodologies and between acceptable types of results can result in data being the only focal point from which productive discussion and mutual understanding is possible.

King (2007), involved in the *Dataverse* project (http://www.dataverse.org), found many benefits for the scientific community to make the data available. We can summarize some of them here:

- **Recognition** for the author: any other researcher that would reuse his/her data would of course cite the corresponding publication, increasing its value and then, the value of the collection it belongs to, i.e.: book, journal or conference proceeding. Articles in journals with replication policies that make data available are cited three times as frequently as otherwise equivalent articles without accessible data (Gleditsch, Metelits, & Strand 2003).
- **Transparency**: making the data verifiable, authorized and persistent should give more credit to results of a publication.
- **Replication**: in many complex situations, it is unfeasible to replicate the study in exactly the same conditions. In such situations, replication of analysis can be performed on the original data if they are shared.

Data Sharing to Face Complexity in Education Science

In education science and education technology, situations involving (several) human beings are far too complex to be replicable. For example, let us suppose that a publication shows a result R_1 in a situation S_1. If another team attempts to reproduce in S_2, the same situation as S_1 (same pedagogical scenario, same instructor, same school, same level, same age, etc.) for the next cohort of learners, it is not certain at all that result R_1 will be confirmed in situation S_2. There are a number of factors already pointed out in the literature, that may be the source of this: the higher experience of the instructor; the simple fact that learners are different; the world changes between S_1 and S_2 (news, crisis, art, technology, etc.) or even a slightly different timetable for learners that brings the observed sessions at an earlier or later time… Especially when we deal with collaboration or interaction analysis, we know that learners build their interaction on their own experience. In the constructivist theory, this is even considered as a basic hypothesis for learning in general. This experience being in constant evolution for everybody, a same learner cannot be in the same conditions for two sequential situations S_1 and S_2. Experience being unique for anybody, two different learners cannot be considered in the same conditions in a given situation. It is even more complex to replicate a given situation (S_1) for a group of learners because you have to take into account not only the (sum of) experience variation between these two groups of learners, but also the history of interaction between pairs of participants (learners and instructor).

Until now, the most frequent case in our scientific field is that we can read a publication that shows results on a given situation, but we have no access to the data collection from which these results have been derived. In most of the cases, we only have a general description of the situation, the data collection and the derivation processes. This state of facts prevents the scientific community from deeper discussion; better comparisons and understanding that could be obtained by the following new derivation processes in case of data sharing:

- Replication of the (same) analysis on the same data using the same analysis process either to understand in detail the analysis process or to verify it;
- Replication of the analysis on the same data but using a different analysis process for comparison of analysis processes;
- Derivation of a secondary analysis on the same data, for example trying to find other results on a different facet of the same data;
- Derivation of a complementary analysis on the same data that builds new results by using the previous outputs;
- Analysis of correlation between results (of different facets) of the same data collection.

When studying collaboration or interaction in learning groups, one way to face the complexity of such situations is to get various analyses with different points of view, in order to evaluate the conditions under which various results co-occur in the different derivations of a given dataset.

Beyond all these new possible derivations, the community would also gain in maturity by exchanging analysis process experiences, teaching and learning more consensual processes and having an available test bed for anybody who wants to perform such analysis processes. In addition, for the technological part of our research, these available data collections may also be useful to test or calibrate new instruments or indicators. Finally, for the most studied collection data, the corresponding situation gains in accuracy on different facets and becomes interesting (for tests) for its well known characteristics. For the most popular datasets, this can lead to well referenced datasets.

Instead of having hundreds of unclassified learning situations, where the data of each are available only to the researchers that built them, we argue that our communities would gain maturity and deepen their understanding by sharing some of the representative situations. Such data could be used as a test-bed for the variety of indicators or methods to analyze various facets of the collaboration.

The Potential of Links between Data, Analysis, and Results

For scientific fairness, data should be available for all discussants. Exposing data and analyses (and not only results) has great potential: deeper understanding and argumentation, articulation of various analyses on the same data. This can lead to in depth discussion of results or relation between complementary results brought by different disciplines. From a methodological point of view, as is the case for any multi-disciplinary project, again, this can also induce a wider spread of methods and tools among involved disciplines.

Considering the structural links between a dataset, its various derivative analysis processes and results, this can lead to a hierarchical representation (i.e. a tree) where the root would be the initial shared dataset and the nodes the derivative analyses, publication and results. As already mentioned by (King, 2007), the fact that a given publication is associated with a data collection makes its analysis process replicable. As well known properties of in- and out-degrees in Social Network Analysis, we can hypothesize that: the longer the list of results and publication that derive from a single dataset (high out-degree for the dataset), the higher the citation index for dataset and derivative publications will be.

DATA SHARING: AN OVERVIEW OF RELATED WORKS

For the Intelligent Tutoring System (ITS) field, the PSLC DataShop presented in (Koedinger, et al., 2008) provides a data repository including datasets and a set of associated visualization and analysis tools. These data can be uploaded as well-formed XML documents that conform to the Tutor_message schema. The goal is to improve the Intelligent Tutoring Systems (ITS) the data are logged from. The datasets are fine-grained, principally automatically generated by ITS and focus on action/feedback interaction between learners and (virtual) tutor tools.

In the CSCL community, a very interesting framework: DELFOS (Osuna, et al., 2001) provides similar proposals as the Mulce project. It defines an XML based data structure (Martinez, et al., 2003) for collaborative actions in order to promote interoperability (between analysis tools), readability (either for human analysts or automated tools) and adaptability to different analytic perspectives. Some of these authors joined the European research project on Interaction Analysis (JEIRP–IA) and reported in (Martinez, et al., 2005) a template describing Interaction Analysis tools and a common format. This common format should be automatically obtained from Learning Support Environments (by an XSL transformation) and either directly processed by new versions of Interaction Analysis tools, or automatically transformed in their original data source format to be processed by previous versions of theses tools. The resulting common format focuses more on technical interoperability than on learning context or human readability. The context is given for fine grain interaction.

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

For this last issue, impressive work has been done in the Dataverse Network project described in (King, 2007). We agree with the members of this project that datasets have to be made available, or at least identified and recorded in a fixed state in order to make sure that data used for a given publication are the same as those identified and (hopefully) made available for other researchers.

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

Considering today's available technology, Markauskaite and Reimann drew an ideal research world in (Markauskaite & Reimann, 2008) where grid computing, middleware services, tools managing remote resources, open access to publications and data repositories, open and interactive forms of peer review process, constitute great potential for e-research. We globally share the same vision for the future of research. Even if we consider that the path to reach this ideal vision is rather long, the main contribution of this chapter can be considered as a modest but concrete step in this direction by presenting an example of data structure for a teaching and learning corpus (Letec) as well as a framework for analytic representation and manipulation (Tatiana).

Availability of data should enable deeper scientific discussion on previously published results. Other researchers may be able to verify or replicate the methods proposed. It becomes possible to compare methods on the same data and then discuss the result or the efficiency of the methods. This way, different analyses can be done on the same set of interaction data. Data may also be connected or compared to other available data. As another example of benefits (for the data provider) of data sharing, let us conclude with these illustrating words:

"Everybody makes mistakes. And if you don't expose your raw data, nobody will find your mistakes." -Jean-Claude Bradley (Wald, 2010)

MAKING DATA SHARABLE

Even if the willingness to share is a necessary element, unfortunately, it is far from being enough to make a collected raw dataset sharable with other researchers. According to our experience (Reffay, et al., 2008; Reffay & Betbeder, 2009), we consider that a given dataset is sharable, if it verifies at least the following properties:

- The context (of the situation) is explicit;
- The structure (of the dataset) is explicit and data are saved in files in open formats;
- The data are free of sensitive and personal information. Rights of publication and use are stated;
- The dataset is referenced by a unique identifier.

These properties are described in more detail in the following subsections.

Make the Context Explicit

Let us refer to an *internal* researcher when dealing with a researcher who belongs to the team that built the situation (or study) or collected or structured the dataset. Conversely, we will refer to an *external* researcher when s/he did not take part of any of those processes. Then, the context elicitation of the situation (or study) is the process that collects all (implicit) information in the internal researcher's head or notes and organizes it in a document (or structure) so that external researchers can find all useful elements when interpreting the data themselves (interaction, learner's production, etc.). Even if this concept of context elicitation is easy to understand, it may be very difficult to achieve in the concrete acts and choices. In fact, the perimeter of *useful information* is different from one analysis to another, from one theory to another and even from one researcher to another. Moreover, some information that could be useful for a specific analysis (e.g. sociological/cultural/linguistic), may be undesirable for ethical reasons (e.g.: ethnic or geographical or cultural/linguistics origins). Now, even if some constraints are irreconcilable in very specific cases, we argue (1) that a lot of other analysis can be done without this undesirable information and (2) that special contracts between respectable research teams are often possible and may lead to an arrangement that may be of benefit to both.

As a positive side effect, even without making the data available for other teams, the simple fact of making the context explicit for any other research also serves the internal research team itself and confers longevity (or even immortality) to the dataset. Newcomers (newly recruited researchers) in the team will be able to re-use this dataset and discuss it with their colleagues despite the fact they are *external* researchers for this dataset. This can help in building a common ground/culture in a research team.

Make the Structure Explicit

We showed in the previous section that explicit context makes data readable (interpretable) for humans. In this section, we argue that explicit structure renders them human and machine readable (usable, computable). The main advantage of a well organized dataset is that any information (contained in the dataset) is easy to find for a human and possible to retrieve for an automated process. The very important corollary is that, if you can't find fixed information, it means this information does not exist in the dataset. In other words: every piece of information may have a single position in the dataset or may be duplicated in (or linked to) any other possible (logical) places.

Being related to computation tools used by the target scientific community, the data may be prepared for these tools in the corresponding formats. Otherwise, the structure should (at least) enable an automated translation to transform the selected format in the target tool's format. Note that it is not the responsibility of the dataset holder to produce the automated translation process. But if such a process already exists, it may be interesting (for the rest of the community) that the selected format (in which data have been actually stored) work as a direct input of this process. The various automated transformation processes may flourish afterwards and be built by some external researchers, interested in transforming the data into a specific (new) format. For the dataset, the ease of use may increase its value. In the CSCL Community, the XML based data structure proposed in (Martinez, et al., 2003) is a common format that enables centralized interoperability.

In order to enable such automated transformations, it is very important (for independence and longevity) to save original data in open formats (txt, rtf, csv, xml, bmp, mpeg, etc.) or very widely used formats (pdf, xls, sql, jpeg, avi, or mov).

Data being often heterogeneous and split into several files, it may be convenient to consider several levels of structure: a global level making clear where metadata, data, information, complementary resources, are and how they are related to one another, and a local or specific level, where we can find individual pieces of data (typically a text, a data table, an XML structure, …) whose organization must be explicit so that readers can take advantage of each of their information pieces. In the Mulce project we adopted the IMS-CP standard integration content package. Such a package is basically composed of a manifest (XML file) structured according to the corresponding schema (XSD) and a "content" folder containing any type and number of heterogeneous files. The last part of the manifest (list of referenced resources) describes and locates each of these files. The first part is generally dedicated to metadata; an arbitrary number of internal parts may be used to describe more specific data or information. We found XML particularly interesting for different reason: (1) it's increasingly widely adopted by different research communities, (2) it is simultaneously formal and malleable and (3) local tag names and parameters make the structure explicit in the innermost parts of wide and long lists repeating headers in each element description. This means that parts of an XML structure can be cut and paste maintaining their comprehensibility. Moreover, identifiers and references in XML structures may avoid repetition of key information or heavy blocks. These are good means to ensure coherence in the data structure.

Consider Ethical Perspectives

It is of particular importance to take into account rights and ethical aspects of data when dealing with long term conservation and widespread dissemination. Data may be free of sensitive (religion, ethnic, health specificities, etc.) and personal information (name, addresses, etc.). Either the dataset depositor received the appropriate permission from participants for all videos, photos and documents where they can be identified, or all documents have been anonymized (photos or videos blurred, names replaced by pseudonyms in text documents). Anonymization may be difficult and time consuming to achieve for some datasets. This may discourage some researchers from sharing their data. We argue that more support may be given to researchers to help them in this task. Efficient anonymization process and tools for each type of data may be developed to support researchers. Now, in case of interaction data in learning session, when participants are aware that they are taking part in a research experiment, sensitive data may not appear. Moreover, ethical committees may take into account the context of the experiment (data genesis) that should influence the risk (for participants) to release these data. Sometimes, ethical committees request destruction of data after a fixed period of time and put some restrictions on data diffusion.

Currently, research foundations such as NSF (that modified its policy in January 2011), are requesting for researchers who want to be funded to provide an explicit plan for data management and a justification in the case they don't share their data. Another initiative: the Panton Principles (www.pantonprinciples.org) were publicly launched in February of 2010. Four months later, about 100 individuals and organizations had endorsed the Principles. According to Pollock:

> *"It's commonplace that we advance by building on the work of colleagues and predecessors standing on the shoulders of giants. In a digital age, to build on the work of others we need something very concrete: access to the data of others and the freedom to use and reuse it. That's what the Panton Principles are about."*

STRUCTURE OF THE CORPUS: THE MULCE PROPOSAL

The Mulce Project

We know how hard it is to build authentic learning situations. When we launched the Mulce project (Mulce, 2010), we thought it would be useful to add some more work on a data collection to structure and document it so that it could be reused later by other researchers or even be ourselves. As we had no ready-made structures to pack a Learning and Teaching Corpus (LETEC), we decided to define one, tentatively reusing standard bricks. One of the basic principles was to pack together a general description of the corpus with an arbitrary number of heterogeneous files. This brought us to use the IMS-CP (2011) specification to build our general package structure presented in the next section. We reused IMS-LD (2011) for context description (Learning Design and Research Protocol) and IMS-MD (2011) for general metadata of the corpus. However, we built a new XML schema for the core component (namely Instantiation) that can contain the data (production and interaction) of the learning situation.

Description of the Package Structure

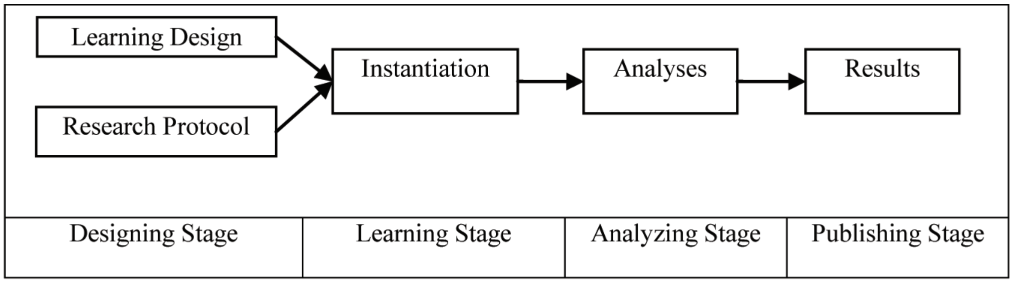
In this section, we first present the main phases involved in this methodological process. Then, we give the derived definition of a "learning and teaching corpus" and explore the structure of its main components.

Building and Recording Interaction in an Online Course

A general organization for an online study is illustrated in Figure 1.

In a first stage, the educational scenario is described at an abstract level, by defining the educational prerequisites and objectives, the abstract roles (learner, tutor, etc.), the learning ac-

Figure 1. Building a research study for an online course: chronology



tivities and the support activities with their respective environments (abstract tools, e.g. chat, forum, etc.) When the course has to be observed for a research study, the researchers define on the one hand the research questions and objectives and on the other hand the list of observable events to be logged. This documentation makes explicit the research protocol or context of the study: i.e. what will be evaluated, are there pre- or post- tests, or participant interviews or questionnaires? In the second stage, the learning situation actually takes place. The abstract roles (designed in both parts of the first stage) are endorsed by real participants, and abstract environments have been implemented in particular platforms including identified tools and virtual (instantiated) spaces. This is the instantiation phase where embodied learners and tutors actually run the activities and identified processes or researchers collect their observable actions (interactions and productions). Specific activities designed in the research protocol may also take place during this period: e.g. pre- or post- tests, interviews, etc. At the end of the learning stage, i.e. when learners and tutors are gone, the collected data can be structured and analyzed by researchers. These analyses hopefully lead to research publications that summarize the context and the methodology and emphasize the results. The data collection is generally not disseminated.

Both documentations of the design phase describe the context of the experimentation. The instantiation phase produces the core data collection that is analyzed in the third stage. In order to make this data collection sharable with external researchers, we show how the various phases presented above become the main components of the corpus defined hereafter.

Learning and Teaching Corpus (Letec): Definition

We define a Learning and Teaching Corpus as a structured entity containing all the elements resulting from an on-line learning situation, whose context is described by an educational scenario and a research protocol. The core data collection includes all the interaction data, the course participants' production, and the tracks, resulting from the participants' actions in the learning environment and stored according to the research protocol. In order to be sharable, and to respect participant privacy, these data should be anonymized and a license for their use be provided in the corpus. A derived analysis can be linked to the set of data actually considered, used or computed for this analysis. An analysis consisting in a data annotation/transcription/transformation, properly connected to its original data, can be merged into the corpus itself, in order for other researchers to compare their own results with a concurrent analysis or to build their complementary analysis upon these previous shared results.

The definition of a Learning and Teaching Corpus as a whole entity comes from the need of explicit links, between interaction data, context and analyses. This explicit context is crucial for an external researcher to interpret the data and to perform their own analyses.

The general idea of this definition intends to grasp the context of the data stemming from the course to allow a researcher to look for, understand and connect this information even though he was not present during the learning (data collection) stage.

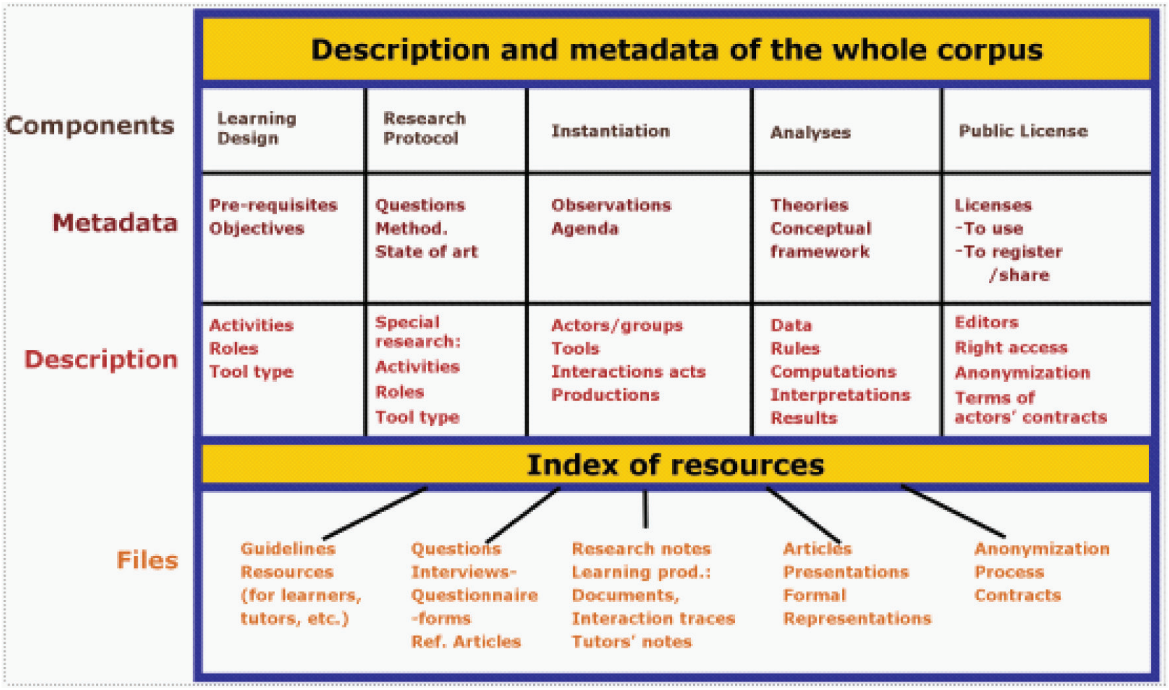
Corpus Composition and Structure

The main components of a Letec: Learning and Teaching Corpus (see Figure 2) are:

- The Instantiation component, the heart of the corpus, which includes all the interaction data, productions of the on-line course participants, completed by some system logs as well as information characterizing participants' profiles.
- The Context concerns the educational scenario and the optional research protocol.
- The License component specifies both corpus publisher's (editor) and users' rights and the ethical elements toward the participants of the course. A part of the license component is private, held only by the person in charge of the corpus. Only this private part may contain some personal information regarding the participants of the course.
- The Analysis component contains global or partial analyses of the corpus as well as possible transcriptions.

The Mulce structure aims at organizing the components of the corpus in a way that enables linking subparts of components together. For example a researcher, while reading a chat session (in the instantiation component), must be able to read the objectives of the activity in which this chat session took place (the activity is described in the pedagogical context, i.e. Learning Design component).

Figure 2. Teaching and learning corpus: the main components in a content package



Description and metadata of the whole corpus

Components	Learning Design	Research Protocol	Instantiation	Analyses	Public License
Metadata	Pre-requisites Objectives	Questions Method. State of art	Observations Agenda	Theories Conceptual framework	Licenses -To use -To register /share
Description	Activities Roles Tool type	Special research: Activities Roles Tool type	Actors/groups Tools Interactions acts Productions	Data Rules Computations Interpretations Results	Editors Right access Anonymization Terms of actors' contracts

Index of resources

Files	Guidelines Resources (for learners, tutors, etc.)	Questions Interviews-Questionnaire -forms Ref. Articles	Research notes Learning prod.: Documents, Interaction traces Tutors' notes	Articles Presentations Formal Representations	Anonymization Process Contracts

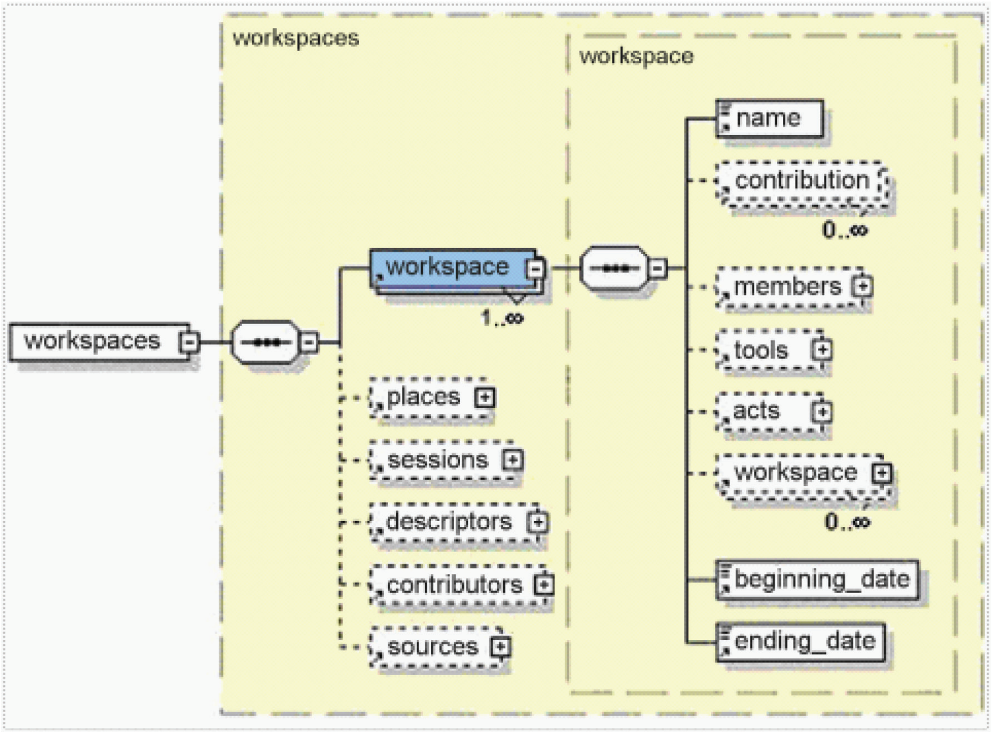
A standard exchange format is also required to download the whole corpus.

Considering these constraints, we chose the IMS-CP formalism (2011) as the global container. This XML formalism fits these constraints by expressing metadata, different levels of description, and an index pointing to the set of heterogeneous resources. In this container, each component is described as an "organization" element of the IMS-CP structure. Each of these can be structured either as basic IMS-CP organization or a more specific one. For example, Learning Design and Research protocol components can use IMS-LD structure as their organization model. If they are only described by a simple text, this text can be defined as an "item" element of the basic IMS-CP organization. For the Analyses component, we generally use a standard IMS-CP organization model. However, the Instantiation component is more specific and has to capture and organize the collected tracks of the situation, played out by the participants. It is the central component where all interaction data and logs may be recorded. In the Mulce project, we decided to define a special XML scheme for this organization: the Structured Interaction Data model (mce_sid, 2011). The next section gives the most important concepts of this model.

Description of the Core Component: Instantiation

The hierarchical structure of the learning stage is captured in the *workspaces* element that contains a sequence of *workspace* elements (see Figure 3). The *workspaces* element may also define: a list of *places* that organizes the space (i.e. each *place* element defines a reference and description of a virtual or physical place like chat room or classroom), a list of *sessions* that splits the time into meaningful periods (i.e. each *session* element defines a reference and description for a dedicated

Figure 3. Extract of the XML schema: the workspaces element



period of time like a chat session or any other [mainly] synchronous activity), a list of *descriptors* or tags that may be used by researchers in their analysis (by associating interaction acts to a set of these descriptors) in order to categorize or count units for each category, the complete list of *contributors* (researchers, developers, compilers, recorders, inputers, etc.) for the corpus and the list of *sources* (i.e. a *source* element is generally a reference to an audio or video record).

A *workspace* is generally linked to a learning activity (of the pedagogical scenario). It encompasses all the events observed during this activity, in the tool spaces provided for this activity, for a given (instantiated) group of participants. As shown in Figure 3, a *workspace* description includes its *members* (references to the participants registered in the learning activity), starting and ending dates, the provided *tools* and the tracks of interaction (*acts*) that occurred in these tools. In order to fit the hierarchical structure of learning and support activities, a *workspace* can recursively contain one or more *workspace* elements.

The lists of *places*, *sessions*, *descriptors*, *contributors* and *sources* defined in the *workspaces* element can be referenced by *workspace*, *contribution*, or *act* elements. For example, *descriptors* may list identified categories so that each act of the acts element list could refer to one or more of these categories. This principle enables browsing of the interaction data in many different ways, independent of the concrete storage organization in the XML document.

Our specification describes communication tools and their features with a great level of precision. The corpus builder can specialize/particularize the schema (i.e., restrict it) to fit the specific tools and features proposed to the learners in a specific learning environment. In the meantime, if a tool cannot be described with the specification, one can augment the schema by adding new elements, in order to take into account the tool's specificities. Restriction and extension are the mechanisms we offer to corpus builders, to adapt our specification to their specific tools or analysis needs.

Moreover, recursive *workspace* descriptions enable the corpus compiler to choose the grain at which he needs to describe the environment. Thus, a workspace can be used to describe a complete curriculum, a semester, a module, a single activity or a work session (a concept generally related to synchronous learning activities). The workspace concept represents the space and time location where we can find interaction with identified tools. This concept has the same modularity as the EML learning units (Koper, 2001; Mce_sid, 2011).

Devices and tools within which interaction occurs can be as different as a forum, a blog, a chat or collaborative production tools (e.g., a conceptual map editor, a collaborative word processor, a collaborative drawing tool).

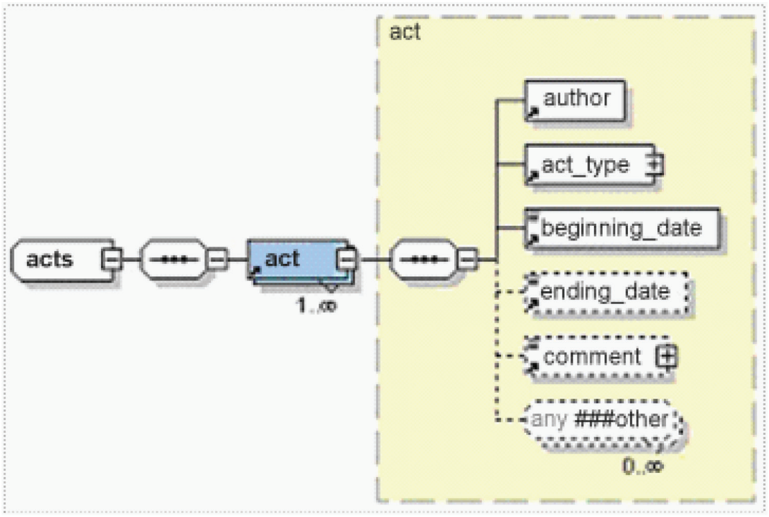
Interaction tracks are stored according to the *act*'s structure presented in Figure 4. All actions, wherever they come from, are described by an *act* element. An *act* necessarily refers to its *author* identifier (defined in the members list—Figure 3), and a *beginning_date*. Depending on the nature of the act (*act_type*), an optional ending_date can be specified. The *act_type* element is a selector. The actual content (or value) of the act depending on its type, is stored in the appropriate structure.

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

Optional element *comment* (Figure 4) contains a sequence typed text of any type and can be used to store researchers' annotations. The last optional element of the *act*'s structure (*any*) leads to any extension not provided in our schema.

This XML Schema defines the storage structure for many act types, e.g.: forum message, chat act,

Figure 4. Extract from the XML schema: the act concept



transcribed voice act, blogs and more. This chapter only gives some of the main ideas of this schema, but the complete schema for structured information data is available online (Mce_sid, 2011).

The definition, composition and structure of a Learning & Teaching Corpus have been presented in the sections above. The next one explains how these data structures can be available throughout Open Archives and specific platforms.

MAKING YOUR DATA AVAILABLE

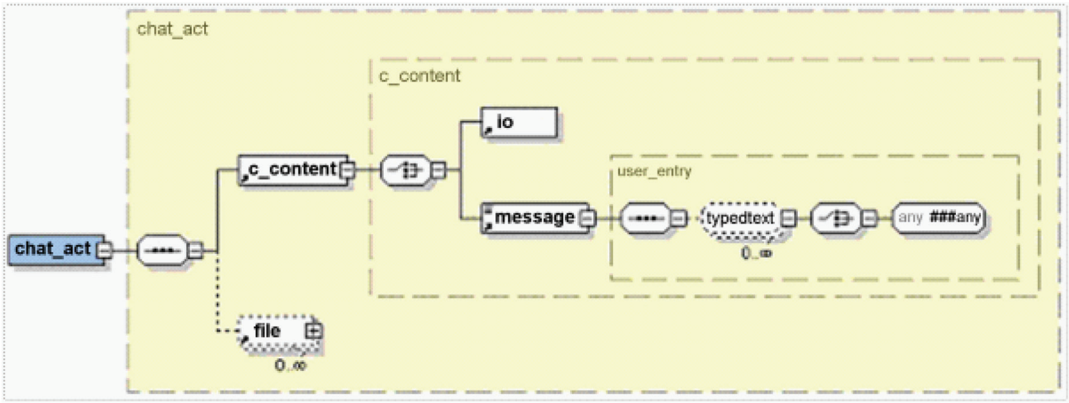
At this point, we can consider that data structure is explicit and context has been documented. When data are correctly aggregated in a formalized package, we have to specify some metadata according to the search need users may have. In our case of learning and teaching corpora (in Language learning field) we decided to use the following metadatasets:

- IMS-MD (2011): general metadata for a content package, including Dublin Core specifications (Powell, Nilsson, Naeve, Johnston & Baker, 2008);
- LOM (2002): Learning Object Meta-data;
- OLAC (2007) meta-data: Open Language Archive Community: collections of data in various languages or concerning languages.

Because metadata characterize data, they can serve several objectives: an object description summary, detailed characterization of the object in various classifications, specific description for referencing, etc. Even if a given corpus is not intended to be widely spread on the network, it could be important (for authorship and precedence reasons) to define its metadata and make them widely available.

In the Mulce project, the 34 objects named corpora (currently registered) can be entirely downloaded by any registered user (Mulce Platform, 2011). Each of these objects encapsulates its own metadata. An exhaustive list of all (general) metadata of all registered objects is stored and maintained in the static repository (i.e. a simple XML file available at a specified URL [Mulce-SR, 2011]). This XML file is harvested daily by the OLAC server that makes them widely visible and responds to any OAI request concerning the Mulce collection. This way, we can be sure that our objects are widely visible and searchable on the web. Moreover, each of the corpora has a unique

Figure 5. Extract from the XML schema: the chat act concept



identifier that can be cited by any researcher who may reuse its content for: scientific discussion, comparative or complementary analysis.

LINKING DATA AND ANALYSES

Once the infrastructure for sharing data has been considered and has been put in place, the natural question of what can be done with these data arises. Already, in the Mulce structure of a corpus, we consider that transcriptions might form part of the primarily shared data and assign a component for analyses on that corpus. In various fields of research (e.g. Conversational Analysis vs. Content Analysis), different standards are expected of transcription and different "authority" is given to the transcription. In conversational analysis (Sacks, Schegloff, & Jefferson, 1974), Jeffersonian notation is used for a transcription which includes pauses, vowel lengths, overlaps between speakers and voice intonation. Such a transcription is kept as objective as possible but is nevertheless not generally considered an acceptable substitute for the original audio or video. On the other hand, for Content Analysis (De Wever, Schellens, Valcke, & Van Keer, 2006), a transcription can be more subjective, including the editorialisation of sentences (which do not naturally exist in spoken language), and the omission of false starts. The subjectivity of the transcription is embraced and subsequent analyses usually trust the transcription as a proxy for the video.

Such practices hint at the idea that secondary artifacts and representations, constructed from the primary data during the analytic process, have different roles in different scientific communities. They are also judged "valid" by a variety of standards. For example in Content Analysis, utterances are typically coded according to a coding scheme laid out by the researcher. A second researcher codes a subset of the data, using the same scheme. The Kappa statistic (Cohen, 1968) is then used to assess the extent to which agreement amongst the researchers is better than chance.

In both the examples of transcription and coding, the researcher constructs an object which is later to be used as a substitute for, or in combination with, the original data. In this section, we examine such analytic artifacts in the field of CSCL. We first examine some of the roles they could play if they were shared in the same way that we suggest data be shared throughout this chapter. We then present a case study describing our experience in sharing data as the focus of a series of workshops in multivocal CSCL analysis. Last we describe the Tatiana framework, which lays out the requirements and a proposed solution for describing and sharing a subset of these "secondary" analytic artifacts.

Why Share Analytic Representations?

The arguments and advantages for sharing and re-using analytic representations are similar for those (King, 2007) for sharing data on which these representations are based. First, a tremendous amount of effort is spent analyzing data. Fisher and Sanderson (1996) report analysis time to sequence time (the duration of the data source being analyzed) ratios of 5:1 up to 100:1 and higher. Second, many (if not all) analysis methods (e.g. Content Analysis) insert steps between the primary data and the final result. How is a reviewer to evaluate an analysis if only the primary data and not the coded data are made available to them? Third, new analyses can be performed more quickly and find results that are more profound by taking a previous analysis as a starting point, rather than as a competing analysis. Such a hermeneutic view is common in fields which analyze corpora that will no longer be extended (e.g. the Bible for theological research, Corpus Iuris Civilis for legal research, etc.).

Furthermore, there are many advantages to analyzing within a team. As exemplified by Content Analysis, several analytic methodologies use inter-rater reliability to validate a new analytic artifact. In this method, two researchers independently perform the same subjective analytic manipulation and compare their resulting artifacts for which a certain amount of agreement must be met. Well modeled analytic representations can easily allow such agreement to be computed automatically. Goodman et al. (2006) describe an analytic tool which they use to distribute the workload among several analysts. Making it easy to piece together the resulting analytic fragments removes the reluctance a coordinator might otherwise feel. Prudhomme, Pourroy, and Lund (2007) describe an analysis in which the multiple perspectives of argumentation and collaborative design iteratively come together to show how multiple criteria are leveraged to evaluate solutions, much in the same way that multiple justifications are used to defend a statement. Last, it is common within projects both in small (student-advisor) and large (multiple laboratory/institution) groups to refine and criticize analyses and methodologies.

For each of these purposes, the ability to share, transform, edit and compare analytic representations is essential in order to avoid duplication of efforts and to allow multiple analysts to examine the data in parallel without having to formally "hand off" the analysis to another person.

Building Inter-Disciplinarity: An Experience

In the field of CSCL, there is a common agreement in the positive value of collaborative learning. However, there is some disagreement as to exactly what constitutes collaborative learning and why it should be educationally effective (Suthers, 2006). The diversity of epistemological beliefs and research methodologies has led several researchers to question how such a plurality could become a source of productive scientific discourse rather than disagreement or balkanization. A series of 5 workshops were conducted, from 2008 to 2011, to address this issue. Some preliminary results of this effort were reported in Suthers et al. (2011). Initial workshops focused on examining five analytic dimensions: epistemological assumptions, purpose of analysis, units of interaction that are taken as the basis of analysis, representations of data and analytic interpretations, and analytic manipulations taken on these representations. Discussions about these dimensions proved productive but did not indicate how or whether different approaches could be reconciled. Sharing of datasets and multiple analyses on these datasets was then investigated, but it was frequently hard to determine to what differences in interpretation were due without being able to return to the primary data. In the final two workshops, to as great an extent as possible, analytic representations were shared and used to combine multiple viewpoints.

In one case study based on these workshops, Dyke et al. (2011) describe three analyses, which originally appeared to be so different as not to be able to inform each other. It was shown how a single visualization combining the viewpoint of two analysts directly on the primary data could illustrate the differences in belief the analysts had about collaborative learning.

Creating Reusable Analytic Representations: Tatiana

In this last case study, the Tatiana analysis tool (Dyke, Lund, & Girardot, 2009; Dyke, Lund, & Girardot, 2010) was used both to perform the initial analyses and to combine these analyses into a single representation. Tatiana is based on a framework which answers several requirements for analytic representation creation and sharing: the necessity of being able to use analytic representations both in combination with each other and with the primary data; the ability to create and share analytic representations based on a previously shared corpus; and the ability to combine and re-use existing analytic representations.

Tatiana (Trace Analysis Tool for Interaction ANAlysts) is an environment (and an underlying conceptual framework) designed for manipulating various kinds of analytic representations, in particular those that present a view on event-based data. We call these representations replayables, because they can be replayed in a similar fashion to a video. They are one of the major kinds of representations that researchers construct to analyze computer-mediated interaction.

Tatiana is built on a number of core concepts and components. Tatiana replayables can be created either automatically (through import) or by hand. Once created, all replayables in Tatiana benefit from Tatiana's four core functionalities: transformation, enrichment, visualization and synchronization.

Transformations

Replayables can be transformed (again, automatically or manually) and exported. As replayables are made up of events (with each event having a set of facets or properties), a transformation results in the creation of a replayable containing new events or a new sequence of existing events. Automated import, transformation and export works through the application of what we call filters. These are objects that combine scripts into a workflow. Scripts are small programs written in XQuery to perform a specific operation, such as transforming a file in the corpus into data Tatiana can understand, excluding certain kinds of events from a replayable, finding certain kinds of events in a replayable, combining multiple replayables, etc. A filter might combine a new script for data import from the interaction log data produced by a new kind of tool with an existing script that only shows the actions of a particular subset of students. Manual transformations include the ability to delete, reorder, re-group and split events.

Enrichment

All replayables within Tatiana can be enriched by analysis generated by the researcher. Such enrichment is the equivalent of adding new columns in a table or, in other words of adding new facets to previously existing events. There are currently three kinds of enrichments supported by Tatiana: free-form annotation, categorization, and graphs. Categorization is simply a way of annotating the verbal transcripts from a restricted list of words and can be used for coding, labeling and adding keywords. The list of categories available can be edited at any time thus allowing for an evolving analysis scheme. Graphs allow researchers to explicitly mark relationships between events. As enrichments annotate the data in a standoff notation, they can be shared separately from the original corpus and can also be opened in concert on the same representations, much in the same way as multiple map overlays can be placed on top of a single map to add in multiple geographical features.

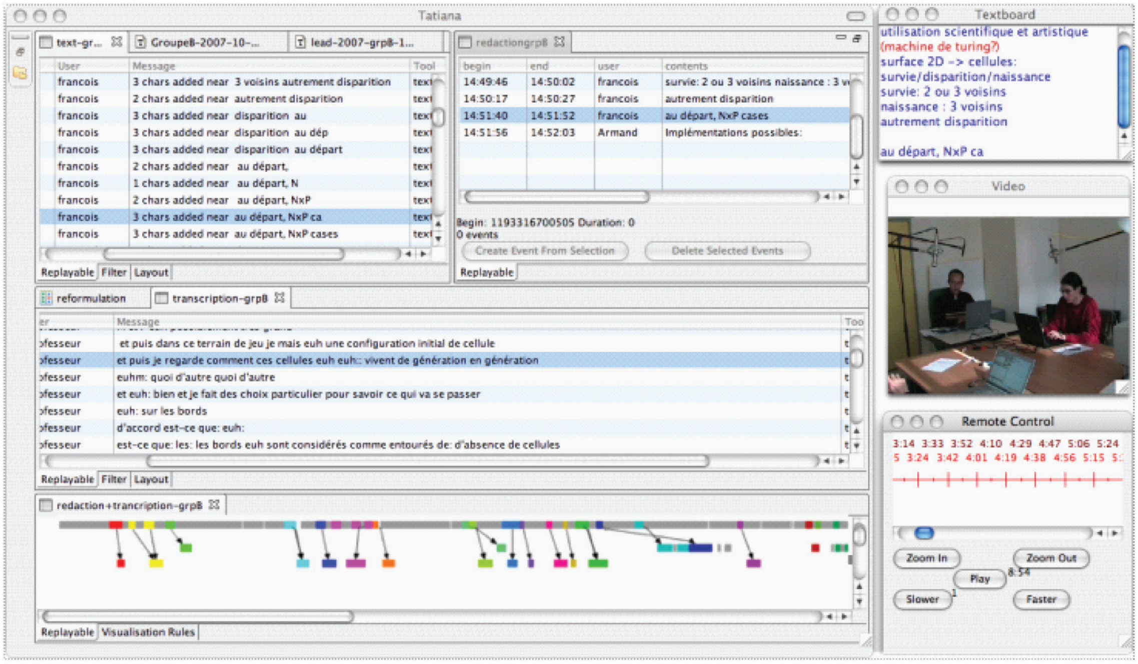
Visualization

All replayables within Tatiana can be visualized in different viewers, which do not modify the underlying abstract nature of the replayable but merely style it appropriately for examination by a human. There currently exist two kinds of viewers: a table view, with one row per event and columns for each of the event's properties and a graphical timeline. The graphical timeline is a first attempt at assisting the automated creation of visualizations. It presents each event as a graphical object whose graphical properties (color, shape, size, position, etc.) can be set according to the properties of the event (user, tool, timestamp, analysis category, etc.). Tatiana is extensible, allowing new kinds of views to be created, affording new ways of visualizing data. The ability to create and configure multiple visualizations, in concert with transformation and enrichment contributes to the ability to re-use and combine previously existing analytic artifacts.

Synchronization

Finally, all visualizations of replayables in Tatiana can be synchronized with each other (cf. Figure 6) and also with data viewed in external replayers such as media players. Synchronized replay means that when a timestamp is selected in the "remote control," the video players (and other external replayers) are instantly navigated to that timestamp, and the events matching that timestamp in the currently visualized replayables are highlighted. Furthermore, selecting an event in a visualized replayable will again navigate all the other views to that moment in time. For example, during analysis of a video and its transcription in

Figure 6. Various replayables visualized in Tatiana: traces of a shared text editor (top left), transcription (middle left), writing units (top center), visualization of reformulation (bottom left), synchronized with external tools, DREW replayer (top right), video player (middle right), remote control (bottom right)



User	Message	Tool
francois	3 chars added near 3 voisins autrement disparition	text
francois	2 chars added near autrement disparition	text
francois	3 chars added near disparition au	text
francois	3 chars added near disparition au dép	text
francois	3 chars added near disparition au départ	text
francois	2 chars added near au départ,	text
francois	1 chars added near au départ, N	text
francois	2 chars added near au départ, NxP	text
francois	3 chars added near au départ, NxP ca	text
francois	3 chars added near au départ, NxP cases	text

begin	end	user	contents
14:49:46	14:50:02	francois	survie: 2 ou 3 voisins naissance : 3 v
14:50:17	14:50:27	francois	autrement disparition
14:51:40	14:51:52	francois	au départ, NxP cases
14:51:56	14:52:03	Armand	Implémentations possibles:

Begin: 1193316700505 Duration: 0
0 events

er	Message	Too
ofesseur	et puis dans ce terrain de jeu je mais euh une configuration initial de cellule	t
ofesseur	et puis je regarde comment ces cellules euh euh:: vivent de génération en génération	t
ofesseur	euhm: quoi d'autre quoi d'autre	t
ofesseur	et euh: bien et je fait des choix particulier pour savoir ce qui va se passer	t
ofesseur	euh: sur les bords	t
ofesseur	d'accord est-ce que: euh:	t
ofesseur	est-ce que: les: les bords euh sont considérés comme entourés de: d'absence de cellules	t

utilisation scientifique et artistique
(machine de turing?)
surface 2D -> cellules:
survie/disparition/naissance
survie: 2 ou 3 voisins
naissance : 3 voisins
autrement disparition

au départ, NxP ca

Tatiana, if a researcher clicks on a time stamped utterance in the table view, this action causes the replayer to bring the video to this point. Information on the dynamics of the interaction in thus provided, which is oftentimes difficult to discern in static log traces. Zooming in on particular episodes becomes possible. In general, such linking between replayables is very useful for limiting the amount of information displayed in a single visualization, with the knowledge that further information is available in other visualizations on demand. Synchronization is the main answer to the necessity of being able to use multiple analytic artifacts in concert.

Beyond Tatiana

The framework presented above is limited to analytic representations which preserve the notions of time and ordered events. It excludes notions such as aggregations (e.g. number of utterances by each speaker), experimental conditions, social networks, etc. It does, however provide a model for how reusable analytic representations can be created. This model, and Tatiana (or any other tool) could be extended to encompass new kinds of representations while considering how they could be integrated with the existing notion of replayables.

Already, some of the corpora available on the Mulce platform include analyses in the Tatiana format. However, while the analytic representations are reusable in concert with each other, they do not currently interoperate with other parts of the corpus (learning context, research context, etc.). Nevertheless, because of the commitment to open standards throughout, once it is more clearly understood what purpose such interoperability might serve, it should be relatively straightforward to implement new tools to facilitate combined use.

While Tatiana may superficially appear to be an analysis tool and its associated data format, we believe that the underlying concept of a replayable and its associated operations provide not only the means for static interoperability, where one analysis tool uses another tool's data as input, but also a means for dynamic interoperability, where multiple tools showing multiple analytic representations are open together in real time and coordinate to enable a better understanding of the underlying data.

It is not only interesting to reuse a shared corpus to add a new analysis of its data; in (Reffay, et al., 2011) we show how a new tool may offer new possibilities to analyze interaction data in terms of social cohesion. But we also show that the research questions (being new for a given tool) may lead to interesting adaptation of these tools that should make them able to support new analyses or extend data that they can use as entry.

CONCLUSION

In this chapter, we recalled the importance of data in the research process and showed the implication of sharing this data. Thus, the road for data (and analysis) sharing is long and our work only represents an initial step, which we hope can serve as an example to researchers in other fields who feel the need for data sharing within their community. In the Mulce project findings recalled in this chapter, we have detailed that: explicit context (both of the learning situation and the research situation), explicit structure (both of the data and subsequent analyses), and a regard for ethical issues are prerequisites for successful data sharing. As an explicit example, we presented in details the general structure of a Mulce corpus as a package that can be referenced and downloaded.

We have also described the Tatiana tool, designed for manipulating various kinds of analytic representations. This tool enables both to perform the initial analysis and to combine these analyses into a single representation by producing synchronized replayable and analytic representations. More generally, with the Tatiana framework, we argue for a thorough understanding of the analytic representations common in a field, of the operations for moving from one representation to another (starting at the original data), and of the means for understanding how multiple analytic representations can inform each other. The current state of the framework only integrates a limited number of kinds of analytic representations (those in which time is a dimension) and is as yet not fully able to work with statistics, aggregations, experimental conditions, etc.

Within our field, at least, there nevertheless remain many obstacles to encouraging researchers to share their data. In particular, it is not immediately obvious that publication of a dataset will pay off in terms of recognition, especially on the part of institutions. Furthermore, journals may be reluctant to impose data sharing, both with regard to submitters who might see it as to high a barrier and to reviewers who already perform a time-consuming task which is not of direct benefit to them. Finally, we are conscious that structuring the data can be a heavy workload as different data types need to be formatted according to the schema. Similarly, adopting new analysis tools for creating interoperable analytic representations requires learning to use them and that they adequately replace existing tools. Although we have only few comparative analyses on the same dataset, we think that this work is an encouraging step towards sharing research data.

As with any system where the benefits are indirect and long-term, while the profitability of a functioning solution in the long term is plain to see, the means of bootstrapping the system are less obvious. As data are increasingly shared, we will be better able to find new ways to manage, describe and combine analytic viewpoints on it.

OUR VISION: THE ROLE OF DATA SHARING IN CSCR

From the contents of this chapter, our vision for the role of data sharing is clear, and may at first glance appear simplistic. Because it promotes recognition, transparency, and replication, we see it as a means by which our field (indeed, many fields) can move forward, with stronger results, built on a multi-faceted understanding. The particular cost of data collection and difficulty of replication in technology enhanced learning, not to mention the variety of epistemologies for subsequent analyses renders the question of data and analysis sharing all the more important and timely for us.

Considering this arising need and emerging service of data sharing, the most efficient way to share is still to be determined. On the one hand, institutions are putting pressure on researchers to release their results, publications and data, especially for publicly funded research. The type of repository targeted by these institutions is rather formal and the benefits for the researchers are not clear. On the other hand, social network applications, both for personal or professional purposes, are evolving quickly and tend to encompass a variety of services. In these social network applications, every type of relationship may lead to a specific sociograms for the research community that can help newcomers to socialize or leaders to better build or manage projects. These relationships may be as diverse as: “attended a given scientific event,” “share common research topics/tag,” “co-authored a publication,” “co-worked in the same project/team/laboratory,” etc. These applications being centered either on individuals or on specific objects shared by individuals (interests, conferences, journals, projects, team, laboratories, etc.), we see at least two reasons for these communities to adopt data sharing: the first would to consider a shared corpus as a scientific publication and the second to use it as an attracter (intermediary object) between researchers who worked on it either as contributor or as analysts. In the SNA theory, we can hope that two scientists sharing the same corpus as contributors are strongly related to each other and the link between a contributor and an analyst may show a concrete involvement of the analyst to engage scientific discussion and collaboration on related or complementary topics (based on the same data as boundary object). Our idea is that such a link is extremely useful to build bridges between communities (complementary analysis) and may lead to in-depth long-term collaboration.

Following this vision that considers shared data as attracters in a widely connected social network, we can expect that widely reused data, providing a variety of analyses and results, would become more and more attractive and play constructive roles for the communities as a boundary object disseminating methods, tools, epistemologies and results. We think this is a way towards better adoption of our research tools and results and finally contributing a more significant impact to society. However, we are not expecting that thousands of researchers will rapidly join as earlier contributors. Because of the cost of data (and methods and tools) adoption, and those of analysis contribution, we are rather expecting that these costs may work as efficient filters preventing noise of superficial communication in the scientific relationships. It should result in a small heavily concerned and engaged community that may easily consider co-publication and project building. In a sense, it would promote a new way of networking where ties between researchers are stronger, communities' doors are more widely opened and visible, so that any contributor is welcome but where the entry ticket (the first contribution for a newcomer) is rather expensive in terms of work. The strength of this approach stems from the openness of data and transparency of analysis processes that should attract newcomers and lead them to better understanding situations, methods and results. Researchers are not the only ones who would be able to reuse these data and analysis processes; this would also be of interest to programmers that would like to test their tools for robustness by attempting to treat the data of the corpus, or implementing alternative processes among those already used in the derived analyses. Their contribution may be either a new tool proposal for the research community, or an application tool that may have real impact to society.

REFERENCES

Bilder, G. W. (2009). Video presentation of CrossRef by its boss: Geoff W. Bilder. In *Proceedings of the 1st Conference on Open Access Scholarly Publishing*. Lund, Sweden. Retrieved from http:// river-valley.tv/ tag/ geoff-bilder/.

Carletta, J. (2007). Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, *41*(2), 181–190. doi:10.1007/ s10579-007-9040-x

Chan, T., Roschelle, J., Hsi, S., Kinshuk, , Sharples, M., Brown, T., & Hoppe, U. (2006). One-to-one technology-enhanced learning: An opportunity for global research collaboration. *Research and Practice in Technology Enhanced Learning*, *1*(1), 3–29. doi:10.1142/S1793206806000032

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. doi:10.1037/h0026256

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyse transcripts of online asynchronous discussion groups: A review. *Computers & Education*, *46*(1), 6–28. doi:10.1016/j.compedu.2005.04.005

Dyke, G., Lund, K., & Girardot, J.-J. (2009). Tatiana: An environment to support the CSCL analysis process. In *Proceedings of the International Conference on Computer Supported Collaborative Learning,* (pp. 58-67). Rhodes, Greece: ACM.

Dyke, G., Lund, K., & Girardot, J.-J. (2010). Tatiana: Un environnement d'aide à l'analyse de traces d'interactions humaines. *Technique et Science Informatiques*, *29*(10), 1179–1205. doi:10.3166/tsi.29.1179-1205

Dyke, G., Lund, K., Jeong, H., Medina, R., Suthers, D. D., van Aalst, J., et al. (2011). Technological affordances for productive multivocality in analysis. In *Proceedings of the International Computer Supported Collaborative Learning,* (pp. 454-461). Hong Kong, China: ACM.

Edgar, B. D., & Willinsky, J. (2010). A survey of the scholarly journals using open journal systems. *Journal Scholarly and Research Communication*. Retrieved June 27, 2011, from http:// pkp.sfu.ca/ files/ OJS Journal Survey.pdf.

Fisher, C., & Sanderson, P. (1996). Exploratory sequential data analysis: Exploring continuous observational data. *Interaction*, *3*(2), 25–34. doi:10.1145/227181.227185

Gentil-Beccot, A., Mele, S., & Brooks, T. (2009). Citing and Reading behaviours in high-energy physics: How a community stopped worrying about journals and learned to love repositories. *Scientometrics*, *84*(2), 345–355. doi:10.1007/s11192-009-0111-1

Gleditsch, N. P., Metelits, C., & Strand, H. (2003). Posting your data: Will you be scooped or will you be famous? *International Studies Perspectives*, *4*(1), 89–97.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, *1992*, 517–520.

Goodman, B. A., Drury, J., Gaimari, R. D., Kurland, L., & Zarrella, J. (2006). *Applying user models to improve team decision making*. Retrieved April 10, 2008 from http:// mitre.org/ work/ tech_papers/ tech_papers_07/ 06_1351/.

IMS-CP. (2011). Instructional management system content package, version 1.2. *Public Draft 2 Specification*. Retrieved from http:// www.imsglobal.org/ content/ packaging/.

IMS-LD. (2011). Instructional management system learning design, version 1. *Specification*. Retrieved from http:// www.imsglobal.org/ learningdesign/.

IMS-MD. (2011). *Instructional management system meta-data. version 1.3, final specification*. Retrieved from http:// www.imsglobal.org/ metadata/.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, *36*(2), 173–199. doi:10.1177/0049124107306660

Koedinger, K. R., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In *Proceedings of the First International Conference on Educational Data Mining,* (pp. 157-166). ACM.

Koper, R. (2001). Modelling units of study from a pedagogical perspective: The pedagogical metamodel behind EML. *Technical Report OUNL June*. Retrieved from http:// dspace.ou.nl/ bitstream/ 1820/ 36/ 1/ Pedagogical%20 metamodel%20 behind%20 EMLv2.pdf.

LOM. (2002). Learning technology standards committee of the IEEE. *Draft Standard for Learning Object Metadata*. Retrieved from http:// ltsc.ieee.org/ wg12/ files/ LOM_1484_ 12_1_v1_ Final_Draft.pdf.

Lund, K., Prudhomme, G., & Cassier, J.-L. (2007). Using analysis of computer-mediated synchronous interactions to understand co-designers' activities and reasoning. In *Proceedings of the International Conference on Engineering Design*. Paris, France: IEEE Press.

Markauskaite, L., & Reimann, P. (2008). Enhancing and scaling-up design-based research: The potential of e-research. In *Proceedings of International Conference for the Learning Sciences,* (pp. 27-34). Utrecht, The Netherlands: ACM.

Martinez, A., De la Fuente, P., & Dimitriadis, Y. (2003). Towards an XML-based representation of collaborative action. In *Proceedings of International Conference on Computer Supported Collaborative Learning Conference,* (pp. 14-18). Bergen, Norway: ACM.

Martinez, A., Harrer, A., & Barros, B. (2005). Library of interaction analysis tools. *Deliverable D.31.2 of the JEIRP IA*. New York, NY: KaleidoScope. Mce_sid. (2011). *Full schema for the structured information data (instantiation component) of a Mulce corpus*. Retrieved from http:// lrl-diffusion.univ-bpclermont.fr/ mulce/ metadata/ mce-schemas/ mce_sid.xsd.

Mele, S. (2009). *SCOAP3: Sponsoring consortium for open access publishing in particle physics*. Paper presented at the First Conference on Open Access Scholarly Publishing. Lund, Sweden. Retrieved from http:// river-valley.tv/ media/ conferences/ oaspa2009/ 0301-Salvatore_Mele/.

Mulce. (2010). *French national research project 2006-2010*. Retrieved from http:// mulce.org.

Mulce Platform. (2011). *Multimodal learning and teaching corpora exchange*. Retrieved from http:// mulce.univ-bpclermont.fr:8080/ PlateFormeMulce/.

Mulce-SR. (2011). *Static repository for the Mulce collection*. Retrieved from http://lrl-diffusion. univ-bpclermont.fr/ mulce/ metadata/ repository/ mulce-sr.xml.

OLAC. (2007). Open language archives community. *University of Pennsylvania*. Retrieved from http:// www.language-archives.org/.

Osuna, C., Dimitriadis, Y., & Martínez, A. (2001). Using a theoretical framework for the evaluation of ksequentiability, reusability and complexity of development in CSCL applications. In *Proceedings of the European Computer Supported Collaborative Learning Conference*. Maastricht, The Netherlands: ACM.

PKP. (2010). *Public knowledge project*. Retrieved from http:// pkp.sfu.ca/ about.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2008). DCMI abstract model. *Dublin Core Metadata Initiative*. Retrieved from http:// dublincore.org/ documents/ abstract-model/.

Prudhomme, G., Pourroy, F., & Lund, K. (2007). An empirical study of engineering knowledge dynamics in a design situation. *Journal of Desert Research*, *3*, 333–358. doi:10.1504/JDR.2007.016388

Reffay, C., & Betbeder, M.-L. (2009). Sharing corpora and tools to improve interaction analysis. In *Proceedings of the 4th European Conference on Technology Enhanced Learning,* (pp. 196-210). Springer.

Reffay, C., Chanier, T., Noras, M., & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, *15*, 185–219.

Reffay, C., Teplovs, C., & Blondel, F.-M. (2011). Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. In *Proceedings of International Conference* on *Computer Supported Collaborative Learning*. Hong Kong, China: ACM.

Rourke, L., Anderson, T., Garrisson, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education*, *12*, 8–22.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, *50*, 696–735. doi:10.2307/412243

Suthers, D. D. (2006). Technology affordances for intersubjective meaning-making: A research agenda for CSCL. *International Journal of Computer-Supported Collaborative Learning*, *1*(3), 315–337. doi:10.1007/s11412-006-9660-y

Suthers, D. D., Lund, K., Rosé, C., Dyke, G., Law, N., & Teplovs, C. (2011). Towards productive multivocality in the analysis of collaborative learning. In *Proceedings of CSCL 2011*, (pp. 1015-1022). Hong Kong, China: ACM.

Velterop, J. (2009). *Nano publications*. Paper presented at the first Conference on Open Access Scholarly Publishing. Lund, Sweden. Retrieved from http://river-valley.tv/ media/ conferences/ oaspa2009/ 0201-Jan_ Velterop/.

Wald, C. (2010). *Scientists embrace openness*. Retrieved from http:// sciencecareers.sciencemag. org/ career_magazine/ previous_issues/ articles/ 2010_04_09/ caredit. a1000036.

KEY TERMS AND DEFINITIONS

Analytic Representation: Intermediary representation resulting from a transformation or analysis process.

Data Sharing: making a data set (issued from a research project) available, understandable and re-usable for researchers not involved in the project that collect the data.

Learning and Teaching Corpus: a Learning & Teaching Corpus is a structured entity containing all the elements resulting from an on-line learning situation.

LETEC: Learning and Teaching Corpus.

Mulce Structure: Conceptual organization of a Learning and Teaching Corpus including learning design, research protocol, structured interaction data, analytic representation and license. The corresponding XML schema is available here: http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/mce_sid.xsd

Replayable: Core concept of the Tatiana software: analytic representations issued from event-based data.

Replication: Run either the same analysis method or comparable one on the same dataset for training or verification.

Tatiana: Trace Analysis Tool for Interaction ANAlysts. (http://code.google.com/p/tatiana/).

ENDNOTE

[1] Mulce is a French 3-year project (funded by the French National Research Agency), led by T. Chanier. More information can be found here: http://mulce.org.