Interactive Construction of EnCOrE (Learning by Building and Using an Encyclopedia)

Philippe Lemoisson, Edouard Untersteller, Maria Augusta Silveira Netto Nunes, Stefano A. Cerri, Alain Krief, Fabio Paraguaçu

Interactive Construction of EnCOrE
(Learning by building and using an Encyclopedia)

Philippe Lemoisson*, Edouard Untersteller***, Maria Augusta S. N. Nunes[1]*,
Stefano A. Cerri*, Alain Krief**, Fabio Paraguaçu ****

lemoisson@lirmm.fr ; edouard.untersteller@icsn.cnrs-gif.fr ; nunes@lirmm.fr
cerri@lirmm.fr ; akrief@fundp.ac.be ; fabioparagua2000@yahoo.com.br

* LIRMM, CNRS & University Montpellier II
161, Rue Ada 34392 Montpellier Cedex 5, France
** COS, FUNDP ; 61, Rue de Bruxelles, 5000 Namur, Belgique
*** ICSN, CNRS ; Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France
**** Universitad Federal de Alagoas, Maceio, Brazil

ABSTRACT

One of the major challenges of current ITS research is scaling up to real world learning scenarios. Another one, strictly interwoven, is to integrate human learning with other human activities, such as constructing theories or performing experiments, as it is the case in e-Science. The paper describes the results of the first phase of an ambitious project: EnCOrE for building and using an Encyclopedia of Organic Chemistry by virtual communities communicating on the Web. The current major result is the computational architecture, but perhaps more interesting is the chain of arguments for each of the architectural choices made and the emerging conceptual model supporting human learning within a socio-constructivist approach, consisting of cycles of deductive, inductive and abductive activities on facts - the shared reality - and concepts - their subjective interpretation submitted to negotiations and finally converging to a consensus - .

[1] PhD student supported by CAPES/1353-02-0 (Brazilian Government) and URI University

Introduction

This paper[2] does not describe ongoing research concerned with the modeling, conception, design, realization or evaluation of a system tutoring something to someone under the verified hypothesis that teaching those concepts to those students is the proper thing to do. Rather, the research described here focuses on the preconditions for building such a system – or perhaps not building it at all – in a real world, semantically rich domain of extraordinary complexity (Organic Chemistry).

The rational analysis of those preconditions occurs under a strong assumption that underlies our work, i.e.: that learning in humans is a side effect of the autonomous construction of knowledge in a social context (social constructivism; situated learning; learning by doing in a social, dialectic environment; negotiation, convergence of jointly agreed ontologies).

In spite of the multiple support offered by the current literature on the plausible adequateness of social constructivism [34] very few concrete guidelines and reusable experiments have been published that offer a roadmap to follow, or architectural and behavioral patters to imitate. One exception, quite known by us, consists of DaNobrega's thesis [8, 9, 10] that however limited the explanation of the conceptual foundations to the toy domain of archs and the experimentation to a reduced subset of concepts in the domain of law. Our current research, instead, has the ambition to play a

[2] The paper extends over more pages than the ITS Call would allow. We might have split into a suite of two, but we preferred not to do it before the reviewing phase, as it would not make much sense to adopt different reviewing criteria for each of the two parts. Authors are ready to split the paper in case the referees will accept it and the editor of the proceedings will prefer so.

pivotal role within a full scale e-Science project: EnCOrE, where both descriptive and procedural concepts are in large number, wide complexity, crucial practical relevance – e.g.: for the synthesis of new drugs or the understanding of germane issues such as those in the domain of biotechnologies – and finally are described by a quite contradictory, partially ambiguous, totally dispersed and uncontrollable, rapidly growing technical and scientific documentation available online.

The major challenge of ITS research being to model human knowledge acquisition (learning) in a domain, enhanced by artificial technologies, it cannot be even approached unless the domain to learn is reasonably well defined and stable. This is not the case of semantically rich and rapidly evolving domains as those that are subject of active scientific research. Therefore, we had to revisit the whole life cycle of ITS development in order to exploit the collaborative construction process for a synergic collaborative learning process of the domain itself, intertwining the acquisition of knowledge on the domain by experts with the learning of the partially defined knowledge previously acquired.

1. Understanding and Learning Organic Chemistry

Organic Chemistry[3], as a part of Chemistry, produces its own objects of study through chemical reactions. Chemists are able to synthesize most of the compounds, even those which possess complex chemical structures including unknown ones, using rules, similarities with known facts and their own perception. Senior chemists are also able to share their knowledge and to educate adequately the youngest both at the experimental and conceptual level.

Organic chemists perceive their subject as intellectually highly structured, with many interconnected ideas. There are many quite different ways of teaching/learning it, because there are many starting points, but the end result is often the same - a broad understanding and a shared but to outsiders opaque language. The sense of logic to the interconnected ideas disguises the fact that there is a serious problem in making the subject truly systematic.

This is probably why senior chemists recognize difficulties in sharing their comprehension, especially in case of knowledge dissemination using electronic media. In fact building connections between the scientific experimental level and chemical equations, textual explanations and models in order to get an integrated expertise have not yet found an explicit expression.

These difficulties find probably their origin in several factors which lead to a set of ambiguous statements hard to exemplify and ineffective for knowledge sharing:

- Huge number of known substances which can be involved in each single

[3] Organic chemistry is the domain of sciences and technology which deals with the compounds containing Carbon, Hydrogen, some hetero-atoms (Oxygen, Nitrogen, Sulphur ...) and metals (Lithium, Sodium, Magnesium …). Synthetic organic chemistry refers to the preparation of such compounds.

generic chemical reaction;

- Necessity to talk about substructures restraining the vocabulary size to a reasonable magnitude while schemes are the only representation of adequate intricacy;
- Gap between the internal complexity of reactions (type and roles of reagents, solvents, catalysts, their ratio and concentrations, physical conditions, etc.) and the apparent simplicity of the chemical equation;
- Important influence of the context on the meaning and the understanding.

Meanwhile, the generalization of computer-assisted interactions in Society has a positive lever effect on the paradigm shift from knowledge transfer (product oriented computing) to interactive knowledge construction (service oriented computing).

As a consequence of these two facts, a project was launched: EnCOrE[4] (Encyclopédie de Chimie Organique Electronique) as a result of an inter-disciplinary, international agreement involving several scientists and departments[5]. Further, the "human learning" aspects of EnCOrE have been considered at the core of a large integrated project approved and funded by the European Union under FP6: ELEGI[6] (European Learning GRID Infrastructure) for the next 4 years. A related EU project on the 5th FP:

---

[4] There exist a huge number of routes to any compound. The more complex the product, the more routes are imaginable. Computerised databases describing compounds exist, but have the same conceptual organization as the reference tomes from the 19th century, and are often misused. ENCORE (ENcyclopédie de Chimie Organique Eléctronique) aims to address this problem.

[5] http://www-encore.enscm.fr/

[6] Official ELEGI Project URL (http://dbs.cordis.lu/fep-cgi/srchidadb?ACTION=D&SESSION=295372004-3-18&DOC=55&TBL=EN_PROJ&RCN=EP_RPG:002205&CALLER=PROJ_IST)

LEGE-WG[7] offered us the scientific and technical background for the research we are describing, reported in the British Computer Society website[8].

EnCOrE will provide complex information retrieval as an educational service in Organic Chemistry for university students and trainee researchers. EnCOrE will contribute to the design of new ways of teaching/learning, especially of Organic Chemistry, which has complex problems of having to handle words, chemical structures, mathematical formulae and experimental techniques, all equally important. It is also expected that, by revealing connections that are not currently recognized, it will even expose areas of ignorance that chemists are not aware of. A collaborative network between well-known synthetic chemists, computer chemists and computer scientists has been initiated in order to develop a productive collaborative environment.

As Guzdial et al [16] said, collaboration in complex problem solving activities facilitates both successful performance and reflection for learning (to enhance the learning process).

In this article, we will sequentially:

- Sketch the main lines of the internal architecture of the encyclopedia, including the role of an ontological nucleus;
- Present a conceptual and a methodological frame intended to conceive relationships between rational agents in the interactive construction of

[7] Official LEGE-WG Project URL (http://dbs.cordis.lu/fep-cgi/srchidadb?ACTION=D&SESSION=27962004-3-18&DOC=7&TBL=EN_PROJ&RCN=EP_RCN:64019&CALLER=IST_UNIFIEDSRCH)
[8] http://www1.bcs.org.uk/

knowledge;

- Sketch the basis of a protocol for the interactive construction of the encyclopedia and show how – on the basis of previous experiences in other domains – one may attribute the function of “environment for human learning” to the system offering the EnCOrE functionalities, both for senior and for junior chemists interacting at a distance.

2. EnCOrE : an encyclopedia around an ontological nucleus

An Encyclopedia, when seen as a body of formally represented knowledge, relies on a conceptualization, in the meaning of an abstract, simplified view of the world wishing to represent something for some purpose [15]. Notice that any conceptualization relies on a purpose, therefore it depends from a viewpoint.

Gomez-Pérez [14] and Noy [24], define ontology as an explicit specification of the concepts in a domain and the relations between them, which provides a formal vocabulary for information exchange. Crubézy [7] and Knublauch [19], agree on this when saying that an ontology provides a structured framework for modeling the concepts and relationships of some domain of expertise.

The idea of a “shared understanding of information, formally expressed“ is clearly highlighted in Noy [24]; and Fensel [11] insists on the point that it can be re-used, cf also [23].

In the particular case of the EnCOrE project, relevant work has already been done inside the community of Chemical Scientists for establishing (i) a shared formal vocabulary [20, 17] and (ii) relations between the concepts (Reactions involve substances in the role of starting materials or products). In a way, ontologies already exist, are more or less formalized and may be available, but are not sufficient from a semantical point of view as they lack consistency and completeness, in other words they are a useful starting point but not the satisfactory result.

In order to render this diagnosis more precise, let us consider the traditional difficulties encountered in ontology building.

As Klein and Fensel underline [18], the pieces of knowledge included in an ontology are often not static, evolve over time: "ontologies are dynamic networks of meaning, in which consensus is achieved in a social process of exchanging information and meaning". Considering this evolution [25], modifications in ontologies are caused by either: (i) changes in the domain, (ii) changes in the conceptualization, or (iii) changes in the explicit specification.

- Changes in the domain are very common from a theoretical point of view, but when addressing Organic Chemistry, we may hope that the domain is rather stable;
- Changes in the conceptualization can result from a changing view of the world; for instance the chemical notion of "functional group" may happen to characterize NOT the molecules themselves, but the relations between molecules: this has practical consequences in the ontology's structure.

According to Staab in Stojanovic [32] this variety of causes and consequences of the ontology modifications make ontology evolution a very complex operation that should be considered as both, an organizational and a technical process;

- Changes in the explicit specification occur when an ontology is translated from one knowledge-representation language to another; in the case of Organic Chemistry, we may hope that this language (the name of which is the same as the domain name) is strongly unified and is not subject to translations.

Difficulties arise because a modification in one part of the ontology may generate subtle inconsistencies in other parts of the same ontology; in ontology-based instances as well as in depending ontologies and applications (Fensel in [32]).

Therefore, it seems to us that out of the three kinds of instabilities, we are due to consider mainly one: conceptualization. However, we are aware that we have to control this in a very accurate way. Changes in conceptualization have not been administered in an "ontology evolutionary process" for years, and one of the hard tasks of EnCOrE is to unify "a posteriori" a great number of independent local evolutions in conceptualization, the expression of which may only be found in the scientific literature. We shall see that, in order to deal with this major difficulty, the conceptors of EnCOrE intend to:

- explicitly take into account <u>viewpoints</u> (both SEMANTIC AND PRAGMATIC VIEWPOINTS), as it will be explained further in this chapter;
- rely on the strong assumption that each semantic point of view is based on chemical experiments, and therefore will always remain relevant; what is to be

expected is the addition of new points of view, not the deletion of old ones. This seems to simplify the scenario;

- Focus on an interaction protocol, which will enable and simplify the emergence of new, duly argumented points of view, as it will be explained in chapter 4.

2.1. Components

We consider three principal components to construct an encyclopedia of Organic Chemistry:

- An ONTOLOGICAL NUCLEUS containing domain-specific scientific terms defined and organized with explicit relations between them. On a consensual basis, terms will describe shared concepts[9] (e.g.: pure substance, chemical element, structure, functional group, vessel reaction, chemical equation, named reaction, retro-synthesis, etc.), using typical SEMANTIC VIEWPOINTS aimed at classifying theoretical situations, and therefore at contextualizing the meaning. In the same time, a collection of PRAGMATIC VIEWPOINTS will be stored and managed in order to take into account the questions raised during the knowledge acquisition process;
- An experimental corpus containing chemical reactions stored with their original description, represented using terms and viewpoints from the ONTOLOGICAL NUCLEUS;
- A theoretical corpus made of scientific articles conceptualizing the experimental corpus. Article content will be "indexed" by the terms and viewpoints of the

[9] Few rough definitions are proposed in the glossary.

ONTOLOGICAL NUCLEUS.

2.1.1 Ontological nucleus

The roles of the ONTOLOGICAL NUCLEUS are:

- To gather the shared vocabulary of the community, to explicit formal representation modes;
- To achieve a coherent integration of contents for their reusability and their recovery by artificial agents;
- To serve as a reference for those willing to construct new ontologies around it, by providing a reusable shared conceptual framework to describe meaning.

The preliminary construction of a consensual ONTOLOGICAL NUCLEUS will be a critical task. We foresee a collaborative construction process among a few established Organic Chemists, specialists in the area of synthesis, in order to secure from the beginning both the birth and the growth of the nucleus. The ONTOLOGICAL NUCLEUS will play the role of a valuable model for more effective work involving a larger number of actors. Iterative control of the consequences of conceptual choices at a rational level will be essential. The challenge of this initial work will consist of the choices on its granularity and the convergence of the consensus among experts.

The construction of the ONTOLOGICAL NUCLEUS will involve collaborative interactions over a shared dual reality based on the collection of experimental facts (Vessel Reactions) and pieces of scientific literature, as it will be exposed in Chapter 4. Shared concepts will be identified and named in reference to the shared reality, maintaining the semantic track and the consensus constraint. The identification of SEMANTIC VIEWPOINTS will help resolving conflicts over meaning of terms, by allowing the contextualization of term definition. The meaning of terms will be defined in a dictionary, using the SEMANTIC VIEWPOINTS, which will be identified and selected in the definition process.

Addressing generic problems and solutions related with the process of knowledge acquisition, the PRAGMATIC VIEWPOINTS will play an important role in the writing as well as in the navigation through articles. As a complement, SEMANTIC VIEWPOINTS will be used in semi-automatic or automatic mark-up of article content *in the theoretical corpus*. Therefore, the organization of the domain into formal structures will occur on a bi-dimensional way based on PRAGMATIC VIEWPOINTS and SEMANTIC VIEWPOINTS, corresponding respectively to the encyclopedia and dictionary modality of utilization of EnCOrE, both in writing and reading activities.

A preliminary proposition to organize the ONTOLOGICAL NUCLEUS content is depicted below:

Ontology of Concepts

e.g. Chemical structure , Element, Functional group, Connectors ..

Ontology of Techniques

e.g. Equipment, Methods, Document types, Document organization …

Ontology of pragmatic viewpoints

PRAGMATIC VIEWPOINTS will be identified and named while classifying GENERIC QUESTIONS. These questions will allow authors to organize their writing and to label their texts and schemes with explicit tags that will be recognized by searching (artificial) Agents. We claim that the explicit classification of pragmatic viewpoints by authors will play a key role in their learning process: not only learning Chemistry, but also learning how to perform meta-reasoning on Chemistry by using the suite of tools available in EnCOrE.

The relation between GENERIC QUESTIONS and PRAGMATIC VIEWPOINTS is exemplified in the following table 1:

Table 1 – Generic questions & Pragmatic viewpoints

Generic question	**Pragmatic viewpoint**
What makes the chemical reaction XXX so useful in synthesis?	Strategic
What is the stereo chemical outcome of the reaction XXX?	Stereo chemical
What is the reactivity of XXX?	Reactivity
How can one obtain XXX?	Synthetic
What are the conditions to transform XXX into YYY?	Conditions
…	…

Ontology of semantic viewpoints

Constructing meaning by relating the verbal and graphical facets of the chemical language represents an area of discovery and learning for organic chemists. Meaning of texts and schemes will be managed under SEMANTIC WIEWPOINTS, providing a conceptual frame supporting (and contextualizing) the interaction with the user by means of the integration of the verbal and graphical modes of expression. As a starting point, six SEMANTIC VIEWPOINTS are proposed, three to consider primary objects, under either Structure/Substructure or Function/Element or Substance/Molecule interrelationships, and three to consider secondary objects: Equation/Redistribution, Taxonomy/Examples, and Strategy/Practice.

2.1.2. Experimental Corpus

The EXPERIMENTAL CORPUS is a collection of facts in their experimental description. For the experimental description of a chemical reaction we introduce the term *Vessel Reaction*. The value of the EXPERIMENTAL CORPUS will reside in the selection of reaction examples following a chart defined by EnCOrE editors, and in its data structure refereeing to the ontological nucleus. Experimental facts (Vessel Reactions) represent the "hard" reality less prone to controversy, and are a good basis to construct a stable ontological nucleus that will be reused in the theoretical corpus.

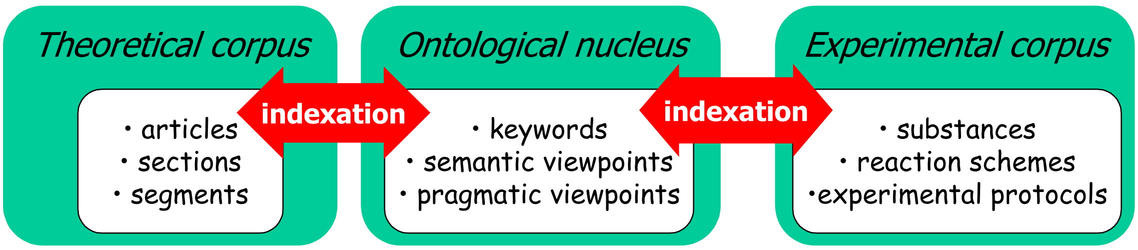
Even if "hard facts" are to be considered somehow "true", a technically profound dialectic process on their interpretation by experts is considered to be the major source of advancements in the chemical science itself offering an opportunity to established scientists participating to EnCOrE to profit from the process of organizational learning that will be necessary in order to reach a consensus and consolidate its results on the evolving Encyclopedia.

2.1.3 Theoretical Corpus

Building block of the THEORETICAL CORPUS is the encyclopedia article, written in reference to the EXPERIMENTAL CORPUS, and in relation with the ONTOLOGICAL NUCLEUS. Based on the hard reality of the experimental corpus, the THEORETICAL CORPUS will be the ground to construct a soft reality, in line with the Popperian [28] approach of conjecture/refutation. The semantic viewpoints and pragmatic viewpoints from the ontological nucleus will help formalizing the Synthetic Organic Chemistry theories.

An editorial ontology will control the article structure, including chapters and segments. A segment is defined as an editorial "atom", to which pragmatic and semantic viewpoints are attached.

The following illustration (Figure 1) shows the three parts of the encyclopedia, and the



relations between them:

***Figure 1 Indexation relationship between the** ONTOLOGICAL NUCLEUS **and the** THEORETICAL **and** EXPERIMENTAL CORPUSES**.** Each article and section has a title whereas each segment is indexed by both pragmatic and semantic viewpoints. They all include some terms of the ontological nucleus and this is* *also the case of each experimental protocol. For each of those terms, the link is established with the definition of the term in accordance with the relevant semantic viewpoint, whenever there is a choice.*

2.2. Building the encyclopedia

Building the EnCOrE encyclopedia means to deal with two intricate views:

- EnCOrE is aimed to become a "repository" for shared scientific knowledge; and therefore we must deal with architectural aspects linking "ONTOLOGICAL NUCLEUS", "EXPERIMENTAL CORPUS" and "THEORETICAL CORPUS";
- At the same time, we must deal with the dynamic aspects of its interactive construction, and define a protocol addressing senior chemists, each one with his/her own experimental background and his/her own semantic viewpoints.

2.3. Learning from the encyclopedia

Any scientific construction makes sense only as much as it may be communicated, shared, learned, used by others; as much as the concepts and processes described and exemplified are reusable by others thanks to their understandability and learnability by other humans, both experts and beginners. In this sense, EnCOrE is aimed to become an "intelligent tutoring system" or, more precisely, an "intelligent learning environment", where chemists may engage in constructive interactions, thus "learn by doing, in the socially dialectic context of pairs".

For instance, according to Crozat and Trigano [6] : 'Our ontology, even if only a restrictive and partial representation of the domain, can bring standardization and methodology, providing to the actors of the domain ; common language and agree in on what is or should be a tutoring system'.

Therefore from the beginning we foresee learnability by peers to be crucial in EnCOrE. We adopt in EnCOrE a clean "design for learnability" approach. It consists firstly of the support for the dynamic aspects of its interactive construction which will imply both senior and junior chemists[10]. The approach of learning is here a synergic approach, sometimes called "human centered design", very different from the classical, applicative approach. Considering the junior chemist at the center, learning is clearly a social, constructive phenomenon. It occurs as a side effect of interactions, conversations and enhanced presence in the dynamic "Virtual Community" of the Encyclopedia authors and users. Therefore we must also define protocols for the dynamic learning aspects: from a modeling and computational viewpoint, the major challenge is not in controlling versions of the consolidated knowledge within the encyclopedia, but rather controlling and facilitating the conversations among agents on the network. This challenge fits with the corresponding one adopted by the "semantic GRID" proclamation: the transition from a product view to the service view of the network supporting global interactions among human and artificial agents [21, 3, 4, 5].

According to Bravo et al [2], scientific discovery learning is an approach which considers the learner as an active agent in the knowledge acquisition process. According to VanGlasersfeld [33], in discovery environments learners are engaged in constructivist activities, and they acquire knowledge through a process of subjective

[10] By senior and junior chemists we denote the variable degree of competence of chemists.

construction starting from the experience rather than through the discovery of an ontological reality.

In order to deal with the interactive aspects of both construction and learning, we need a framework. Next chapter is dedicated to this point and will disclose not only a conceptual framework for the description of rational agents and of their interactions but also will focus on a brief description of the model MIDES, which is being developed [26], in order to establish a methodological framework for this interaction.

3. Interaction of Rational Agents

We wish to build a "framework" where formally defined protocols can be defined, and where the following questions find an answer:

- How can "senior" chemists belonging to remote laboratories build together an "ONTOLOGICAL NUCLEUS", powerful enough to express a deep understanding of the "chemical reality"?
- How can they share an EXPERIMENTAL CORPUS, and build a THEORETICAL CORPUS on top of it?
- Will "junior" chemists be able to learn from the encyclopedia, and how?

To answer these three questions, we shall start with a very general conceptual framework, which can be understood as an extension of the multi-agents paradigm [12]. We will then introduce a methodological framework, which is being developed in a close collaboration with EnCOrE.

3.1. A conceptual framework

We must assume, before starting, that:

- There is a "physical world", involving real "objects" in which real experiences take place. Therefore all chemists are looking at the same REAL objects and events, although they may see and therefore describe those in different ways;
- There are "intelligent entities", which allow to put these experiences in order, to link some experiences to other ones, and to build "meaning" upon all this. We will call them "agents" independently from their human or artificial nature[11].

In the simplified framework we have outlined above, a RATIONAL AGENT is able to:

- Interact with the real world through interfaces (eyes, mouth, keyboard, monitor ...) according to rules expressed by experimental protocols;
- Describe the objects of the real world, as well as interact inside groups through messages, using languages at the syntax[12] level;
- Integrate private experiences, develop abstractions, and give a meaning to received messages. This is the semantic level: semantics is primarily a result of choices made by Agents (human or artificial), more than an intrinsic property of syntactic structures. One point we wish to outline here is that the agents have

---

[11]We are going to adopt this habit, not because we have an anthropomorphic view of software (or a mechanistic view of human thought), but rather for simplicity.

[12] SYNTAX can be described :

- in "social sciences" as the "grammatical relationships among signs, independently of their interpretation or meaning" ;
- in "formal sciences (Informatics, AI, ...)" as the level where "well formed expressions" are built and recognized. For instance, when a program is seen as a set of expressions, the syntax is checked by the interpreter or the compiler.

NO DIRECT ACCESS to another's semantic level, they are just allowed to guess with the help of protocols relying on real experience and syntax. The consequences of this assumption are very important, as we will see in the following.

The basic "operations" in the building of ontologies are classifying and naming: they imply both the semantic and the syntactic level (Figure 2).

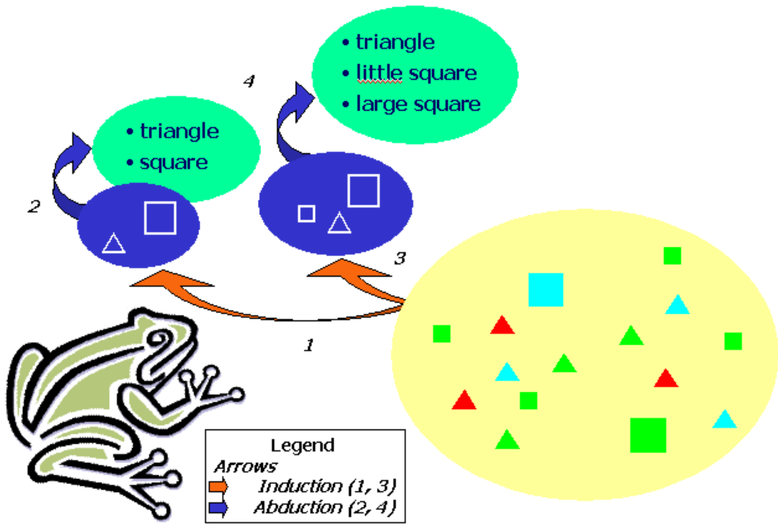


Figure 2. Induction/Abduction cycle. *As a simplified model for any rational agent, our frog is able to perceive the shape of "real" objects from her "local horizon", and then to classify and name them by means of induction and abduction*

Classifying happens in biological brains through the cross-activation of neural networks [13]; it happens in the case of symbolic machine learning, for instance, through algorithmical analysis of Galois lattices [22]. The input for classification is a set of experiences / examples which constitute the local private memory of the agent who classifies; the logicians call this operation **induction**, defined as "generalizing from peculiar examples". In the figure below, it is first induced that 2 types of geometrical figures are to be considered (triangle, square), then with more accurate classification 3 types. This happens at the local semantic level (inside the "brain" of the agent).

Naming makes a classification visible to other agents; otherwise it would remain confined inside each agent's semantic level. And of course, naming is subject to consensus since a community (and not a single agent) is implied. We shall characterize this operation as an **abduction** [27] defined as the "emission of an hypothesis which will have to be validated". In the figure below, each induction is followed by an abduction: the complex process of "agreeing on words" is not represented.

Jean Sallantin and his team [8, 9, 10] have built a protocol in which the basic cycle connecting the private sphere of semantics to the public sphere of syntax is the induction/abduction cycle. With the help of this framework, we wish to focus on a few points:

- <u>Horizon</u> (local experimental reality): Does each agent have a local horizon (access to only partial information) or a global horizon (access to complete information)? In the context of Organic Chemistry, each agent's horizon consists in his/her own bench, or in his/her own readings and conversations, for instance when attending conferences but also when chatting on the Web;
- <u>Memory</u>: Are all the examples simultaneously given or are they sequentially introduced? In the second case, do they have to be stored by the agent, and for how long? In the context of Organic Chemistry, each new experience has to be added to a huge experimental corpus;

- Starting point: Do the agents of the group already share a language, with or without derivation rules? In the context of Organic Chemistry, the wish to build a language on the basic "chemical equations", that is assumed to be a starting point.

We are now able to define constructive interactions [21] as situations in which a given protocol allows a group to develop a common knowledge. This will allow building and stabilizing a new syntactic corpus through a conversational process involving each agent's internal semantics.

In the following example, two agents are co-building a theory, following a protocol defined by:

- *Local horizons* (each one only perceives the objects included in the corresponding rectangle);
- *Simultaneous examples* (we suppose that all the objects are present when the agents speak);
- *A shared ontology and a formal language* with the expressive power of first order logics (allowing IF … THEN statements), therefore allowing **deduction.**

In a first step, each one induces, from his own experience, and makes a hypothesis, which is submitted, to approval or refutation (Figure 3).

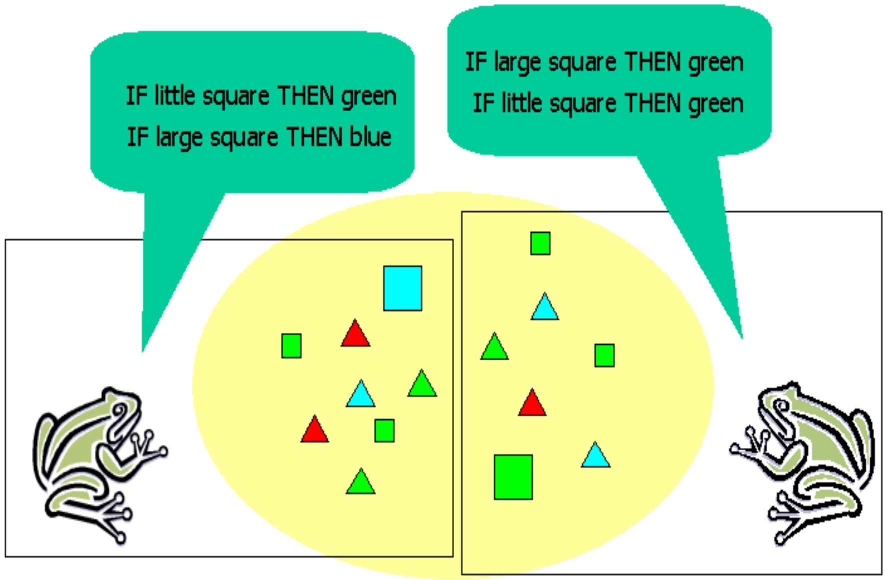


Figure 3 Agent local horizon. Each agent makes his own abduction, according to his local horizon.

In a second step, the theory has to be revised, according to the robustness of each agent's hypothesis applied inside the other's horizon (Figure 4).

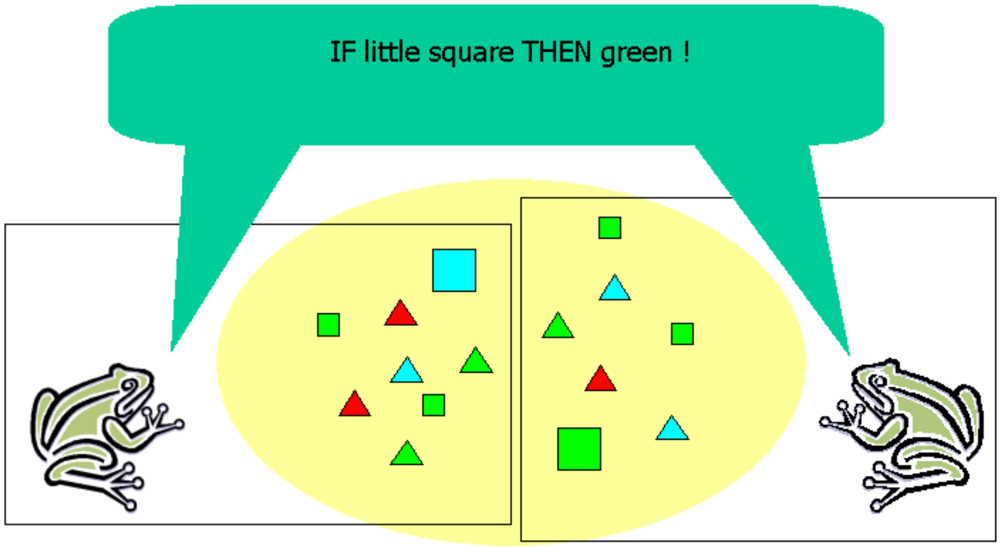


Figure 4. Theory revision. *Each agent takes into account the other's abduction and proceeds to a revision of one's theory*

3.2. A methodological framework: MIDES

The design of scientific theories is the objective of MIDES (Model Interactif de DEcouverte Scientifique) as a methodology to connect scientists using their computers as a P2P communication tool as well as a structuring context to build their models. The MIDES conception [26] provided a conceptual frame to help scientist in the initial step towards the discovery of underlying information structures in their domain. MIDES offers four concurrent modes of expression in a four-sector screen visualization. Three modes of expression are typical of knowledge modeling systems; the fourth one is specific to MIDES:

1. Organizing entities in a *hierarchy;*
2. Enunciating *relations* between entities;
3. Making *causal hypotheses*. In our architecture they can prefigure logical modeling of reactions ;
4. *Generic questioning schemata* nourish the model with generically expressed questions and answers using local variables that are instantiated on the model. These generic questions, emerging from conversational processes, will help to structure the theoretical corpus. They will be organized in PRAGMATIC VIEWPOINTS as shown in section 2.1.1.

At the symbol level (Figure 5 from [26]), MIDES is a methodological toolkit integrating the four parallel modes onto the same syntactic frame; it has been tested with success in a collaborative interdisciplinary context, allowing to make explicit viewpoints and highlighting their importance when giving definitions to terms. With the help of this toolkit, the Rational Agent manages the memory of action and the memory of problems.

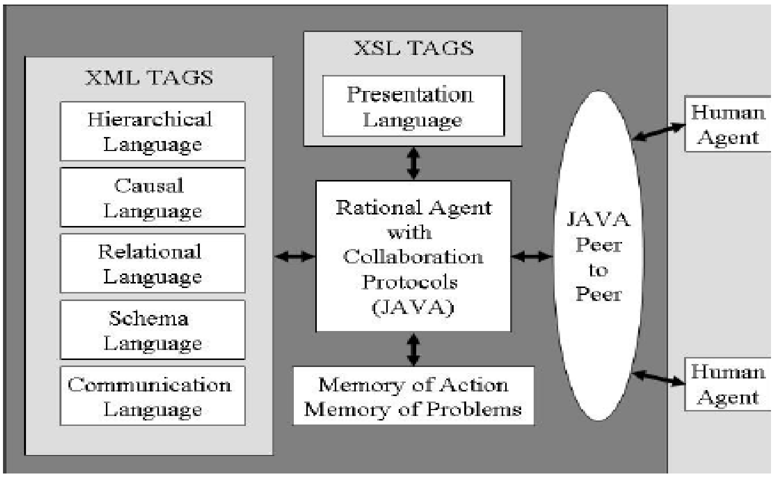


Figure 5. ***MIDES symbol level.***

MIDES keeps record of the whole conversational process. This is of special importance in the collaborative building of ontologies in order to avoid endless iterations in arbitration.

4. A basis for an experimental protocol

We are now ready to reformulate our questions and give clues for answering them:

- What kind of protocol and tools do "senior" chemists belonging to remote laboratories need in order to build together an "ONTOLOGICAL NUCLEUS", powerful enough to express a deep understanding of the "chemical reality"?
- What kind of protocol and tools do they need in order to build together a THEORETICAL CORPUS on top of an EXPERIMENATL CORPUS?
- Will "junior" chemists be able to learn from the encyclopedia, and how?

Before answering these questions, we must sketch the main architectural choices we made for EnCOrE. Then we are going to explain the protocols implied in the interactions, briefly evocate the tools, which will support the conversational processes, and finally explain and argument how the processes guided by this protocol entail learning.

4.1. Main architectural choices

We have proposed a conceptual architecture based on a clear distinction between:

- a hard kernel, the ONTOLOGICAL NUCLEUS, structuring the whole encyclopedia, rarely modified because very closely connected to scientific experimentation, and containing the consensual and stable community representations;
- a peripheral soft layer pertaining to specific areas (articles, sections, segments) containing the controversial and variable representations based on the core ones: controversies may be due to group's assumptions (schools of Chemistry sharing a yet unexpressed viewpoint) or individual's assumptions.

We will implement this nucleus in a database, and use the terms of this database in order to structure and index the whole encyclopedia. An encyclopedia as a closed universe offers a shared reference and can be checked for consistency; this consistency will be easier to maintain when each “key word” corresponds to a unique record in a database.

The peripheral layer, although “soft”, has the vocation to be accessible to the greatest number of human agents, and eventually to be computable by artificial agents; therefore we need commonly accepted editorial norms. XML appears today as the best (if not as the only) candidate. Using XML, we shall structure “articles” in “sections” and “segments”, and index all of those by “key words” belonging to the ontological nucleus.

4.2. Protocols

First of all, let us come back to the basic “abduction/induction” cycle (figure 6), which we shall refine in the rest of the paragraph:

- arrow 1: from the observation of facts, scientists induce “regularities”;
- arrow 2: these “regularities” are expressed in the literature, as hypothesis of “general properties” submitted to the community of chemists, which will accept or reject them;
- arrow 3: from the study of literature, scientists induce new “properties”;
- arrow 4: in order to test these theoretical properties, they abduce “experiments” that will be accepted or rejected as proofs.

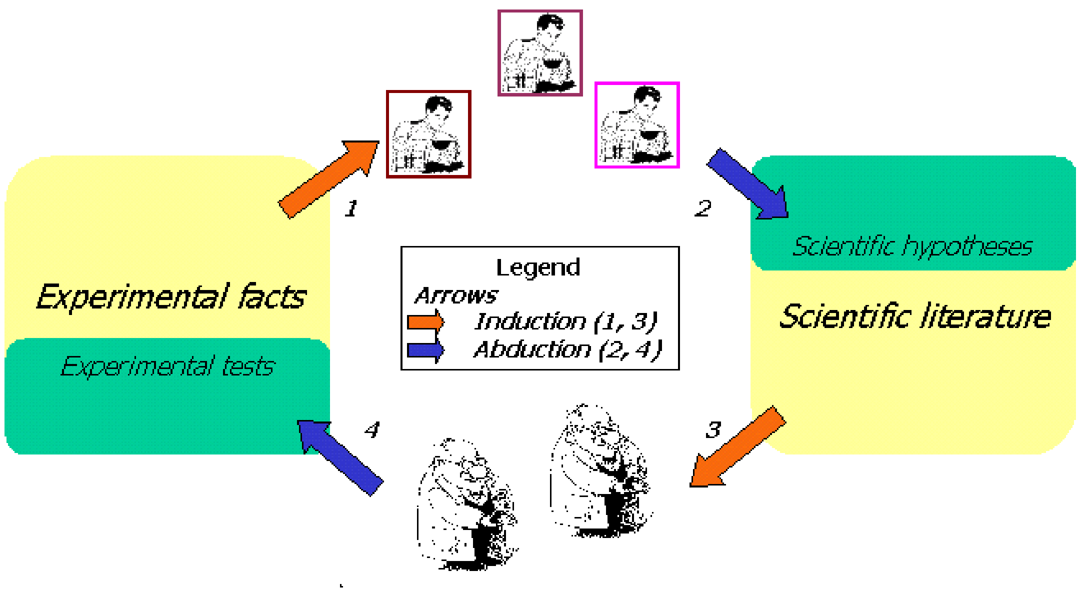


Figure 6- the basic "abduction /induction" cycle.

The crucial question about the adequateness for theory formation of this kind of cycle is of course the question of convergence : whether the accumulation of experimental facts lead to an increasing consensual theoretical corpus or not.

Our answer to this question, which will hopefully be experimentally verified, is that convergence relies mainly on three factors :

1. intersection of *horizons;*
2. shared access to *memory;*
3. expressive language as a *starting point.*

In the following paragraph we intend:

- to look closer into the global cycle, and to split it in two according the architecture;
- to see how ontologies will be associated with the reality of dialogues between their designers and with examples, in order to fulfill the requirements 1. and 2.

The first part of the cycle consists in building the ONTOLOGICAL NUCLEUS:

As it has been underlined before, the main key to a shared expressive language is building an ontology of SEMANTIC VIEWPOINTS. This task is the highest priority in the ONTOLOGICAL NUCLEUS (Figure 7); it will:

- allow the contextualization of definition terms;
- greatly help a homogenous writing of the theoretical corpus.

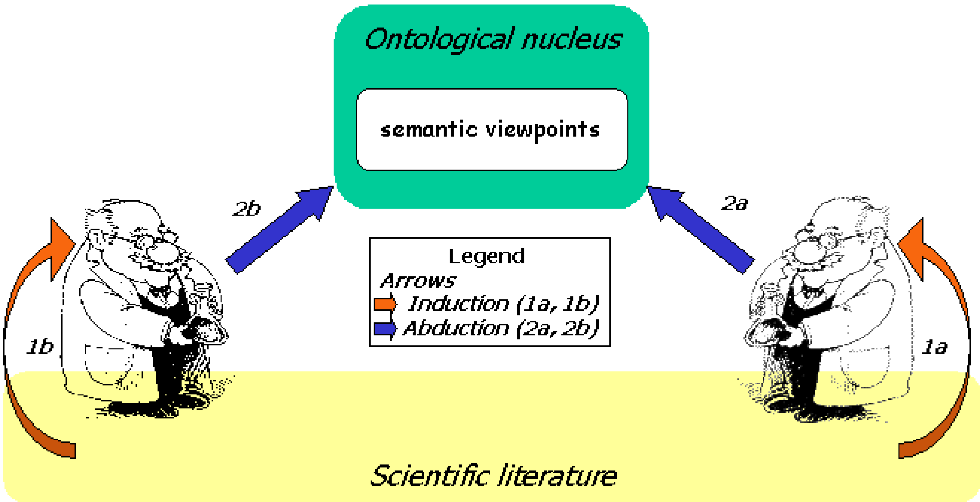


Figure 7. Building the ontology of semantic viewpoints.

The second part of the cycle consists in building the EXPERIMENTAL and the THEORETICAL CORPUS:

- experimental corpus: Juniors build INDUCTIVE reasoning (arrow 1a, Figure 8) about experimental facts, ABDUCE new experiments aimed at infirming or confirming their interpretation, and as soon as this local cycle is stabilized, ABDUCE descriptions of these experiments at the syntactic level (arrow 2a, Figure 8) in the vessel reaction database, using terms from the ONTOLOGICAL NUCLEUS. They also ABDUCE descriptions of these experiments (arrow 2c, Figure 8) in scientific literature. In the same time, they will elaborate part of the ontological nucleus under supervision by seniors (for example, a format to represent chemical reaction in the experimental corpus);

- theoretical corpus: On the basis of their respective reading of the scientific literature[13], Seniors build INDUCTIVE reasoning about what has been written in the past by themselves or others, and submit (ABDUCTION) theoretical hypothesis inside articles structured in reference to the pragmatic points of view listed in the ONTOLOGICAL NUCLEUS.

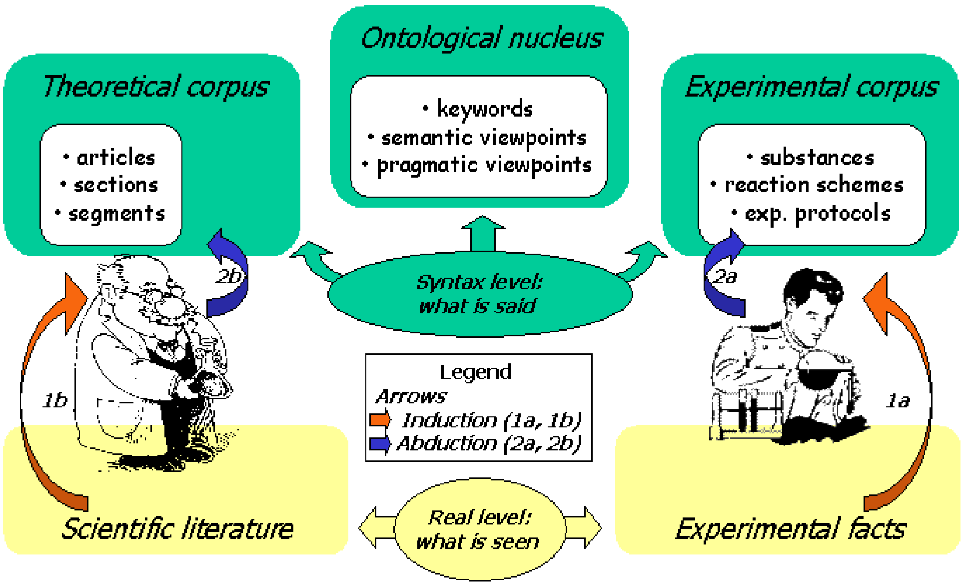


Figure 8. Experimental facts and scientific literature.

4.3. Tools

The protocols we have described rely on strong interaction between a great numbers of contributors. If "juniors" may appear willingly involved in such constructive interactions, seniors are expected to have little time to spend…

That is why this kind of protocols would remain purely theoretical without the help of tools allowing:

[13] Although it is of syntactic nature, the scientific literature automatically acquires the status of "shared reality": we do not mean that different chemists interpret what they read in the same way (if so, there would be no problem at all), we just mean that they read the same sentences.

- a robust capitalization upon asynchronous conversations;
- distant interaction based on enhanced presence.

According to Bourdeau and Mizoguchi [1], shared intelligence among the designer/author and the authoring environment is the dynamic component for the interactive construction of instructional scenarios and learning environments. Therefore, declarative knowledge about instruction could be the "beating heart" of an ITS. In the context of the EleGI / EnCOrE project, the operational principle of building an instructional system that sustains and fosters the learning process will be in our focus during the analysis of conversational processes, and the designing of tools for enhanced presence.

Our intention will not be to investigate speech acts produced by "teachers" with the intent of examining the effects they may have on students learning, as it is done in [29], in reference to a traditional "teaching paradigm".

It will be to investigate dialogue, through dialectic processes exemplified by the tutor and internalized by the learner , because such processes guide the learner's reasoning and accelerate the transition from naïve to scientific conceptualizations. Therefore our work will be inspired, among other ones, by [31].

4.4. Learning

Learning first happens when building the encyclopedia (Figure 9):

- Juniors are given experimental facts, and are asked to describe them in the terms of the ontological nucleus (indexation by key words);

- Juniors are given scientific articles, and are asked to structure them according to the ontological nucleus (splitting according to pragmatic viewpoints, indexation by key words);
- Seniors observe these "indexation processes" and interfere.

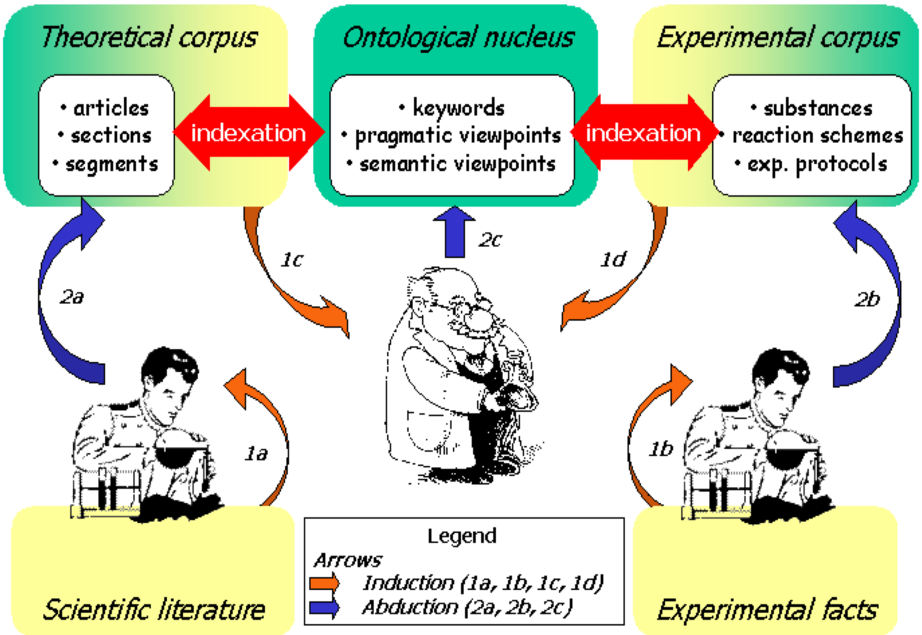


Figure 9. Learning while building the encyclopedia

Learning also happens when using the encyclopedia and questioning the theoretical corpus:

- Seniors provide articles in the theoretical corpus, which are structured according to pragmatic viewpoints in order to assist searching and reading of the encyclopedia (Figure 10);
- Juniors ask questions about the theoretical corpus and therefore ask for answers corresponding to specific pragmatic viewpoints (Figure 10).

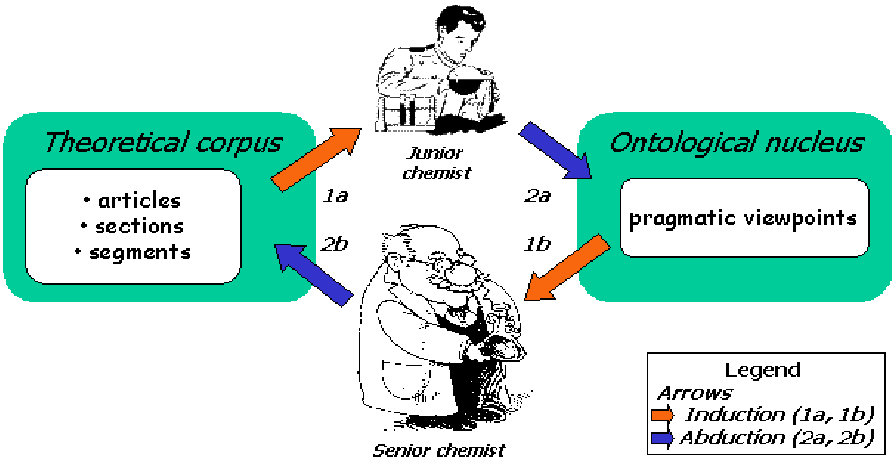


Figure 10. Learning while using the encyclopedia

5. Conclusions and perspectives

Human learning is a hard subject, to understand and to foster. Technologies enhancing human learning have often adopted a classical paradigm emerging from the recent history of teaching, more than on profound considerations - such as those reported in the "Dialogues" of Plato's conversations with his teacher Socrates - that would have required to enter into the subtleties of human behavior, before producing any technical aid.

We have described above the first phase of an ambitious research work that aims at a deeper exploration of human learning behavior when humans are engaged in a common challenge: build collaboratively an Encyclopedia for Organic Chemistry. Human learning, in that endeavor, is mixed with human understanding and human scientific work.

The paper has shown that it is possible, with the currently available technologies, to foresee a generic yet quite neatly defined architecture supporting understanding, doing and learning within remotely connected virtual Communities interacting on the Web (on the GRID). The lessons we have learned from this study, and that we present and exemplify in the paper are mainly two:

1. understanding, doing and learning occurs as chains of deductions, inductions and abductions that reflect quite directly our common experience in reasoning, abstracting and generalizing on and from examples, on and from conceptualizations. The difference with many other attempts is that these operations are put into the social perspective of conversations for a common goal. This rational aspect of the processes within constructive interactions may be today represented, controlled and enhanced by suitably designed computational structures organized within an architecture equipped with conversational protocols. Previous experiences performed also by some of us in different depth and domains have shown the feasibility and the adequateness of the approach. What will be challenging for us now, is the leap in scale and the genericity of the model, and its corresponding realizations.
2. understanding, doing and learning are not purely rational activities, but require emotional aspects, such as motivation, trust, confidence etc. The conversational approach we have described above, together with the enhanced presence tools we just have cited in passing, and the analysis of conversational traces, argumentations, contradictions and

dynamic human help available online seem to us very promising for becoming a crucial support offered today by technologies.

We did not enter these last issues in depth in the paper, as the first one - the architecture - was already dense and required to perform and justify quite selective choices. However, we are perfectly aware that if there will be one bottleneck in the EnCOrE project, this will be due to emotional aspects not sufficiently cared for by the initial versions of the system. Therefore, we will address these issues as the most important ones deserving our priority in future activities in the EnCOrE project.

Acknowledgements

This work has been possible with the support of CAPES (Brazil), the European Union (Projects LEGE-WG and EleGI), and the respective institutions of the authors.

REFERENCES

1. Bourdeau, J. ; Mizoguchi, R. Collaborative Ontological Engineering of Instructional Design Knowledge for an ITS Authoring Environment. ITS'2002. Proceedings… 2002.pp: 399-409.

2. Bravo, C. ; Redondo, M. A. ; Ortega, M. ; Verdejo, F. Collaborative Discovery Learning of Model Design. ITS' 2002. Proceedings…2002Crescencio Bravo, Miguel A. Redondo, Manuel Ortega, Felisa Verdejo: Collaborative Discovery Learning of Model Design. Intelligent Tutoring Systems 2002: 671-680 671-680.

3. Cerri, S. A. ; Eisenstadt, M. ; Jonquet, C. Dynamic Learning Agents and Enhanced Presence on the Grid. 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning Berlin : 2003. (Disponible in http://ewic.bcs.org/conferences/2003/3rdlege/session2/paper2.htm).

4. Cerri, S. A. Open Learning Service Scenarios on GRIDs. 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning Berlin : 2003. (Disponible in http://ewic.bcs.org/conferences/2003/3rdlege/session2/paper1.htm).

5. Colaux-Castillo, C. ; Krief, A. EnCOrE (Encyclopédie de Chimie Organique Electronique): an Original Way to Represent and Transfer Knowledge from Freshmen to Researchers in Organic Chemistry. 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning Berlin : 2003. (Disponible in http://ewic.bcs.org/conferences/2003/3rdlege/session1/paper5.htm).

6. Crozat, S. ; Trigano, P. An ontological approach for design and evaluation of tutoring systems. ITS'2000. Proccedings… Montréal, Canada, 2000.

7. Crubézy, M. and Musen, M.A. Ontologies in Support of Problem Solving. Handbook on Ontologies*. S. Staab and R. Studer*, Springer(2003). Pp: 321-341.

8. DaNóbrega, G. M.; Cerri, S. A.; Sallantin, J. On the Social Rational Mirror: learning e-commerce in a Web-served Learning Environment. In S.A. Cerri, G. Gouardères, and F. Paraguaçu, editors, Intelligent Tutoring Systems 6th International Conference - ITS 2002, Biarritz (France), June 5-8 2002. LNCS 2363, pp. 41-50. Springer-Verlag Berlin Heidelberg 2002.

9. DaNóbrega, G. M. ; Cerri, S. A.; Sallantin, J. A Contradiction-driven Approach of Learning in Discovery Learning Environments Proceedings of SBIE 2003, 12 - 14 November 2003, Universidade Federal do Rio de Janeiro. Best paper award (out of 330). Journal of the Brazilian Computer Society.

10. DaNóbrega. Une Approce Dialectique à la Formation de Théories : Aspects Conceptuels; Formels et Pragmatiques dans le Cadre de l'Apprendissage Humain. Thése de Doctorat. Université Montpellier II. 2002.

11. Fensel, D.; Harmelen, F. van,; Ding,Y.; Klein, M.; Akkermans, H.; Broekstra, J.; Kampman, A.; Meer, J. van der; Sure, Y.; Studer, R.; Krohn, U.; Davies, J.; Engels, R.; Iosif, V.; Kiryakov, A.; Lau, T.; Reimer, U. On-To-Knowledge: Semantic Web Enabled Knowledge Management. *In IEEE Computer*.(2003).

12. Ferber, Jacques. "Multi Agents Systems: an introduction to distributed artificial intelligence". Addison-Wesley.1999.

13. Gerald M. Edelman. "Neural Darwinism: The Theory of Neuronal Group Selection". Basic Books, 1988.

14. Gómez-Pérez, Asunción. Ontological Engineering: A State Of The Art . *Expert UPDATE, 2(3)*: 33-43, 1999.

15. Gruber, Thomas R.. A. Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5(2*), 1993, pp. 199-220.

16. Guzdial, M. ; Kolodner, J. ; Hmelo, C. ; Narayanan, H. ; Carlson, D. ; Rappin, N. ; Hübscher, R. ; Turns, J. ; Newstetter, W. Computer support for learning through complex problem solving Communications of the ACM. Volume 39 , Issue 4 : 1996. pp: 43 – 45.

17. IUPAC {International Union of Pure and Applied Chemistry} Compendium of Chemical Terminology, Second edition, 1997, Edited by A D McNaught and A Wilkinson. http://www.iupac.org/publications/compendium/index.html

18. Klein, M.; Fensel, D. Ontology versioning for the Semantic Web. *In Proceedings of the First International Semantic Web Working Symposium (SWWS).*Stanford University, California, USA.2001. pp: 75-91.

19. Knublauch, H. An AI tool for the real world: Knowledge modeling with Protégé. *Java World Journal*, june 2003. (Disponible in http://www.javaworld.com/javaworld/jw-06-2003/jw-0620-protege.html on 20/01/2004).

20. Lavoisier, Antoine-Laurent. Traité Élémentaire de Chimie, Cuchet, 1789.

21. Lemoisson, P.; Cerri, S. A.; Mahe, S. A.; Sallantin, J. Constructive Interactions. 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning. Berlin.2003. (Disponible in http://ewic.bcs.org/conferences/2003/3rdlege/session1/paper2.htm).

22. Liquière, M.; Sallantin, J. Structural machine learning with gallois lattice and graphs. In 5th International Conference on Machine Learning, pages 305–313, Madison, Wisconsin (USA), 1998. Morgan Kaufmann.

23. Motta, E. Reusable Components for Knowledge Modelling, IOS Press, 1999

24. Noy, N.F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R., Musen, M.A .Creating Semantic Web Contents with Protégé-2000 *In IEEE Intelligent Systems*, Vol. 16, No. 2, March/April 2001, special issue on Semantic Web, pp. 60-71.

25. Noy, N. F. and Klein, M.. Ontology evolution: Not the same as schema evolution. In Knowledge and Information Systems, 5. in press. 2003.

26. Paraguaçu, F. ; Cerri, S. A. ; Costa, C.; Untersteller, E. "A peer-to-peer Architecture for the Collaborative Construction of Scientific Theories by Virtual Communities" IEEE/ITRE2003, August 10-13, 2003, Newark, New Jersey, USA.: ITRE/IEEE, 2003. v.1. p.140 – 145.

27. Peirce, Charles Sanders. Draft E: On Arguments from Memory 19, 1902.

28. Popper, K. The Logic of Scientific Discovery. Routledge, 1977.

29. Porayska-Pomsta, K.; Mellish, C.; Pain, H. Aspects of Speech Act Categorisation: Towards Generating Teachersв Language. International Journal of Artificial Intelligence in Education. 2000.Edition 11, 254-272.

30. Procédé et système de conception interactive d'une base de connaissance libre d'ambiguïtés ainsi que l'outil informatique pour la mise en oeuvre du procédé et du système. FIDAL/CNRS, Patent deposed on November, 21 2000. No. 0014999.

31. Ravenscroft, A., Pilkington, R. M. Investigating by design: developing dialogue models to support reasoning and conceptual change. International Journal of Artificial Intelligence in

Education: Part I of the Special Issue on Analysing Educational Dialogue Analysis. 2000. Edition, 11, 3, 273-299.

32. Stojanovic, L.; Maedche, A.; Motik, B.; Stojanovic. N. User-driven Ontology Evolution Management. *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management EKAW*, Madrid, Spain, 2002.

33. VonGlasersfeld, E. (1978). Radical constructivism and Piaget's concept of knowledge. In F. B. Murray (Ed.), *The impact of Piagetian theory* (pp. 109-122). Baltimore, MD: University Park Press.

34. Vygotsky L..S. Mind in Society. The Development of Higher Psychological Processes. Edited by M. Cole, V. John-Steiner, S. Scribner and E. Souberman.Harvard University Press. Cambridge, Massachussets. 1978.

GLOSSARY

Chemical equation

Symbolic representation of a chemical reaction where entities are separated by a '+', with stoichiometric numbers indicated if they differs from 1, where reactants are on the left and products are on the right, both sides being separated by an arrow or =. In our definition, the two sides of the chemical equation must be in a stoichiometric relation. This constraint frequently disqualifies chemical equations to represent reactions with appropriate richness; removing the stoichiometric constraint gives reaction schemes.

Chemical structure

Made of atoms connected through bonds, organic compounds all have a structure that can be described at several levels (isotopomerism, connectivity, bonding, aromaticity, stereochemistry, tautomerism, epimerism, etc.); structures are not necessarily related to a known compound, but must integrally represent an entity in opposition to substructures.

Element

Found in the periodic table as a set of isotopes, each element belongs to a categorization that is specific to Synthetic Organic Chemistry. Based on the chemical properties their presence confers to organic compounds, Carbon, Hydrogen, Metals, and Heteroatoms are the four main classes

Functional group

Considering reacting chemical structures, taking the saturated linear hydrocarbon backbone as an inert reference, frequently occurring reactive substructures are classified under representative classes that do not generally have in return an exact structural definition.

Named reaction

In contrast to substances, known reactions do not have systematic names, but they do receive one, usually given in acknowledgment to those who first identified them, or in reference to a transformational facet of the reaction.

Pure substance

Ideal representations of matter at the macroscopic level, pure substances are the relata in a network of chemical transformations that connect them.

Retro-synthesis scheme

Inverted synthetic scheme used in planning a synthesis or in discussion of strategic issues about a known one, it shows strategic bonds to disconnect, key intermediates, and starting materials required to reach a target.

Segment

Editorial atoms explicitly labeled into encyclopedia articles, containing usually several words including keywords, delimiting a topic, a question, a claim, etc, that can be classified in relation with semantic and pragmatic viewpoints of the ontological nucleus.

Vessel reaction

In a vessel under given physical conditions, when contacting chemical substances induces a chemical transformation, if a product can be isolated or if it is possible to show chemically the presence of an intermediate (by inducing an alternative reaction, orthogonally to the synthetic scheme), then a vessel reaction has been performed. Within one experiment, zero, one or several successive vessel reactions lead to isolation of created or recovered compounds.