Applications

From Edoc® To PeKITA: the evolution of a platform for e-learning and profiling technology[1]

Roberto Guarasci, Anna Rovella and Stefano Vuono

Università della Calabria

guarasci@unical.it; anna.rovella@unical.it; stefano.vuono@unical.it

Abstract

With the realization of the limits which have, so far, influenced the diffusion of E-Learning systems, above all, with regard to its elevated costs but also to its didactic rigidity, the idea has emerged of structuring a flexible system, with limited costs and which, by means of the semi-automatic indexing of teaching units, allows for significant adaptability with regard to the changing cognitive situations of individual students. The result obtained is edoc®, an E-Learning system with owner technology, which can easily be customised and adapted, and which has a high level of inter-operability with other systems. Its evolution in terms of market application is PeKITA (Personalized Knowledge In The Air) the result of collaboration between the University of Calabria, Siemens Italdata and the Institute of Computational Linguistics at the CNR (the Italian National Research Centre).

[1] Translation by Dr. Michael Cronin, Università della Calabria.

Edizioni Erickson – Trento
www.erickson.it

1. Start-up

For modern Knowledge oriented organisations, distance learning/training is a response to the continual need for growth in terms of one's competitive value, yet it presents the by now noted handicap of the inability to adapt, once in use, the didactic structure to audience variation in terms of its interests and desires. It is in this sense that there has been a strong emphasis towards mastery learning on the part of the second generation Distance Learning.

The aim of edoc© is precisely that of applying profiling technology, largely based on linguistic intelligence, to Distance Learning courseware and more general platforms of Knowledge and Content Management in an attempt to manage formal and informal learning phases by means of the integration of Knowledge Management and E-Learning and dynamic competency updating.

In the wake of the fast growth recorded in recent years, the market today offers an ample choice of products which make it possible to attain positive results both in companies and in Public Administration.[2] Some of the structural gaps, however, have not been overcome, gaps which have limited the spread of such systems over time. First among these is the elevated expense and second there is the didactic rigidity of the system which manifests extreme limitations in the presence of non-homogeneous classes, both in terms of entry level and learning capacity *in itinere*. As has been mentioned, it is from the awareness of these problems, that the idea of structuring a flexible system has emerged, a system with moderate costs which would manage to guarantee, by means of semi-automatic indexing of teaching units, adaptability with regard to the changing cognitive situations of individual students.

The result obtained is edoc®, an E-Learning system with owner technology, which can easily be customised and adapted, and which has a high level of inter-operability with other systems, so as to allow for the re-use of content,[3] realised by the Laboratory for Documentation at the University of Calabria.

2. Fields of Use

Edoc® has been conceived for use in a university context in the Course in Documentation at the University of Calabria as an attempt to cover the deficit in terms of study credits among students in the first year of their three year degree course. The entry tests had in fact revealed in particular a notable gap in regard to activities related to textual comprehension, in their consequent reduction into

[2] For an overview of problems related to E-Learning in Italian public administration, see: CNIPA, Quaderni, anno I, n. 7 (2004), (Number dedicated entirely to this argument).

[3] See La Monica (2002) and Guarasci & Rovella (2004).

primary informative elements and in their use both in terms of the personal enrichment of their cognitive patrimony and of their use in the various required activities.[4] These handicaps, which are pertinent to any discipline, were particularly evident in the field of documentary sciences, in which textual analysis is a central element, and they were further aggravated by the almost total ignorance of the basic methodology for bibliographical and documentary research combined with scarce knowledge of the use of technological instruments.[5]

Traditional classroom teaching, which is an obligatory part of the credits assigned to the discipline, did not lend itself well to the resolution of the problem, also because of the impossibility of weighing down students with integrative activities which would inevitably add to those of other institutional courses. The hypothesis was ,therefore, to realise a didactic structuring of the course in Documentation,[6] in which direct teaching could be supported by Distance Learning courses which would serve to contextualize and intensify study of the arguments treated, with particular care being paid to the strongly interdisciplinary aspect of the subject.

Arguments treated in classroom taught lessons[7] have been initially inserted into e-doc so as to offer the student an extensive reference grid, but it should be noted that these come from all the support materials which are traditionally used to promote deeper study of an argument. The learning structure is based on a reference scheme successively enriched by a «biography» embracing a wide range of forms of support (video, slides, tests, exercises) so as to offer more «technologies of thought», more means of explanation and therefore of thought. The structure of use was based on Cartesian coordinates.

The vertical line of buttons indicates the same argument within which different colours indicate the different media used: violet for video, grey for texts, sky-blue for practice/exercises.

The horizontal line, on the other hand, indicates a didactic pathway of sequential type similar to that followed in direct teaching, although there is no limitation on their use as single items, so as to leave complete freedom to the student in the construction of his/her course.

[4] In the Academic year 2002/2003, 37% of students at the Faculty of Lettere e Filosofia manifested difficulty at various levels in the analysis of written texts . This percentage rose to 42% the following year. The figure relate to the student entry tests for the first year of three year degree courses. For the use of a PC at a level equivalent to that necessary for the taking of the ECDL, the percentage was, for the two years examined, 19% and 22%.

[5] See Candalot (2003) and Coulon (1999).

[6] The course in Documentation is, in the University of Calabria, activated during the second year of the degree course in Mediazione Linguistica which foresees, among other things, a specific formative curriculum in languages aimed at an information and communication based society.

[7] The possibility of off-line use of class content was consistent with the necessity of guaranteeing learning to those unable to attend either for professional or health reasons.

The fundamental principle in the final evaluation has been that of avoiding that this be based exclusively on the short time period of the traditional exam and took into account, rather, the entire corpus of activities carried out during the learning period.

The evaluation has been based, therefore, on three addenda: written exam, oral exam and practical exercise. These three components, of which of which only the second is carried out «in presence» did not have the same weight as regards the final evaluation, but were valued respectively 50%, 24% and 26%.

On the basis of evaluations made by students both in response to questionnaires distributed by the *Nucleo di Valutazione di Ateneo* (the University Evaluation Team) and in response to specific questionnaires, it emerged that student approval is for the most part positive (88%) although there was unanimous complaint regarding the increase in teaching load. Use of the course on the part of students from different degree courses in which the teaching load in terms of credits was different, has, furthermore, has led to a modular structuring which has not always been combined with completeness of information and an exhaustive transmission of contents. An extremely positive aspect, however, has been the rise in the percentage of exam passes, which has shifted from 67 to 88% compared to the average of the two preceding years in which direct tradional teaching was used.

Subsequently, edoc® was tested, with positive results, on groups of users of different ages and educational levels. In particular, it was used as an instrument of Distance Training under the auspices of the Artnet project – Internationalisation of Calabrian artistic craftsmanship – financed by the Ministry for Foreign Affairs with a view to the establishment of stable import-export links. The use of edoc® has been fundamental in the training phase of the relative intermediaries in that it was possible to create web community composed of members who were extremely diverse in terms of conditions of entrance and geographical location of their activity. A further testing phases was carried out under the auspices of mis. 3.10 of POR (Progetto Organizzativo Regionale) Calabria, a project dedicated to the advanced training of managers in Public Administration based on the ICT themes. Also in this case, the platform succeeded in managing the diversities in a group of 250 students sub-divided into 10 groups distributed across the five provinces of Calabria and extremely non-homogenous in terms of age, basic training and working roles. In these applications, a rate of favourability to the platform emerged equal to 91%, yet the absence of a control group made a statistically assessable check on learning impossible.

3. Methodology and Development of edoc®

The basic idea of edoc® and its strong point is in its integration of Knowledge Management and E-learning and the use of an ontology builder engine for the construction of dominions based on the extraction of ontologies from the metadata and the indexing of conents.[8]

The edoc® system is able to shape contents on the basis of user requests evaluating their level of preparation and allowing for the re-modelling of the teaching/training pathway in real time, and supplying collaborative work tools with which it is possible to create a virtual class with other users, with a tutor who verifies the learning phase. In its present configuration, it is in conformity with the criteria and general definition for distance learning supplied by the Ministry for Education, University and Research with the Decree dated 17 April, 2003.[9]

The core of the system which effectively represents it innovative nature, especially in relation to its low implementation costs, is represented by the profiling methodologies based on an initial personal declaration of competencies upon entry and on the progressive modification of the so defined profile by means of evaluation tests *in itinere*. The didactic contents reside in a Knowledge Base and are composed of a combination of Learning Objects structured according to the Dublin Core[10] standard and represented in XML language. Upon this basis, edoc® respects the principle of inter-operability, guaranteeing the re-use of contents within other E-learning systems which adopt Dublin Core as a standard of reference.

The choice of Dublin Core as a standard for Learning Objects within edoc® was influenced by two factors: on the one hand, the opportunity of using a standard which was compatible with the majority of web resources and therefore endow the system with a high grade of inter-operability, while on the other there was its use of clustering software.

Each Learning Object is described (or indexed) by a maximum of 15 descriptive elements (metatag Dublin Core, version 1.1) divisible in 3 groups:

- content of document (title, subject and key-word, description, source, language, relation to other documents, spatial or temporal cover);
- intellectual property of document (name of creator, editor, contributor, juridical source);
- exemplarity, or intrinsic characteristics, of document (date, type, format, identifier).

[8] See Garro et al. (2003, pp. 36-45).

[9] Decree 17 April, 2003 – Criteria and procedures for the accreditation of distance learning courses in state and non- state universities and in university institutions qualified to confer academic qualifications according to art. 3 of the decree of 3 November, 1999, n. 509 – Official Gazzette (Gazzetta Ufficiale) n. 98, 29/04/2003.

[10] Dublin Core is an ISO standard in that the ISO commission on «Information and Documentation» (TC46 SC4) has defined the standard ANSI/NISO Z39.85 - 2001 Dublin Core Metadata Element Set, which has been officially approved as ISO standard ISO 15836:2003. The latest version is Dublin Core Metadata Element Set, Version 1.1 Cfr. Metadata for e-Communities: Supporting Diversità and Convergence, Atti del decimo International Dublin Core Workshop, Firenze, 13-17 ottobre 2002, Firenze University Press, Firenze (2002).

The indexing of the Learning Objects is carried out in a semi-automatic fashion in respect of the rules of ISO standard 5963/85.[11] A textual analyser[12] carries out textual analysis which allows for the generation of indexes and word lists, proceeding to the calculation of occurrences, to the analysis of keywords and the research of phrases and idioms contained.

The resulting report contains the necessary indications for the next phase of conceptual analysis, for the identification of the index terms and the relative Dublin Core categories. In all cases in which it is not possible to determine a term for a category, the latter is omitted.

The aggregation of Learning Objects on the basis of user request comes about by means of the clustering[13] of these last, which reside in the Knowledge Base.

Clustering is carried out with the Insight Discoverer Cluster[14] – an application capable of generating a single XML file containing all the documents in the Knowledge Base, opportunely grouped in a hierarchical fashion according to argument.

4. From edoc® to PeKITA

The most recent evolution of edoc® is PeKITA (PersonalisedKnowledge in the Air), the result of collaboration with CNR – ILC and Siemens – Italdata S.p.A.[15]

[11] UNI ISO 5963/1985: Methods for the analysis of documents, the determination of their subject and the selection of index terms. For an analysis of the problems related to the application of this regulation, see Jolion (2001).

[12] Concordance was conceived by R. J. C. Watt M.A. (Aberdeen), M.Litt (Oxford), FEA Senior Lecturer in English, University of Dundee. Fellow of the English Association (elected 2002); it was chosen and used because, though freeware, it allows for the importation of specialist language (see note 13).

[13] In statistical applications clustering techniques allow for the gathering of single units belonging to a sample or to a population on the basis of the similitude of the variables that characterise them. A hierarchical structure is thus created from which important information can be gathered regarding the aspect of the phenomenon under examination. If we consider a Learning Object as a statistical unit, characterised by a series of variables (metadata), the result of clustering will be a hierarchical structure in which the Learning Objects are sub-divided in terms of semantic/linguistic similitude, and therefore by subject class. While in statistics the rules of aggregation of single clusters are composed by mathematical formulae, in the case of Learning Objects, it is on the basis of terms or groups of terms (features) that they are included or excluded from the cluster.

[14] Insight Discoverer Clusterer is a Java application, based on Remote Method Invocation technology, with a public Application Programmatic Interface. In order to function, Insight Discoverer Clusterer requires that documents be described by means of arrays of their characteristics or in XML format. The clustering engine allows for the grouping of documents which are similar from the point of view of their content. To measure inter-documentary similarities, it needs to know the characteristics of the documents, specialist dictionaries and the rules of extraction which have been defined and implemented in collaboration with the laboratory of phonetics and phonology at the University of Calabria.

[15] Project no. 5370 using Muir funds according to Ministerial Decree 593/2000.

The necessity for further development is related to the desire to pass from an object conceived to satisfy institutional didactic requirements and realised through the integration of software and technical solutions for which the level of ownership was limited to the didactic sphere, to project hypothesis which was entirely owner-based with the aim of commercialisation on the part of the industrial partner, Siemens Italdata.

Conceived in the same sense is the collaboration with the institute of Computational Linguistics at the CNR for the realisation of four modules: *one for advanced linguistic analysis, one for term-clustering, one for terminological indexing and one for conceptual indexing.*[16]

The linguistic analysis module subdivides into a battery of instruments for the automatic treatment of the written text, which include:

- a segmenter for the automatic identification of base textual units (words, punctuation marks, dates, proper names etc.);
- a morphological analyser which associates a morphological analysis to each textual unit in terms of one or more lexical exponents, and one or more groups of morpho-syntactic features;
- a builder for non recursive syntactic groups, which groups continuous sequences of words which are labelled morpho-syntactically into elementary syntactic constituents (known as «chunks»);
- an analyser of the relationships of syntactic dependency between chunks.

Many of these sub-components have already been developed by ILC in the course of its institutional activities. In some cases, already existing components will be updated or modified as required with a view to their integration.

An *ex-novo* development phase is foreseen for the checkout stage for the rules and the output format for the relationships analyser, given that this sub-component represents the natural interface between the linguistic analysis engine and the other functions of the system. It is therefore essential that this be capable of producing results which are perfectly aligned with the functional specifics of the integrated system and with the general requirements of the project. In particular, in order to guarantee wide possibility of use and versatility for the system, there is a plan to develop a package of rules which are interpretable according to a syntax and a format which are pre-defined, which can be easily modified/integrated by the user.

The Module for term clustering will make use of input from the linguistic analysis module in order to produce groups of technical terms acquired from texts and relevant according to a pre-selected dominion, with a view to their

[16] See Peters and Picchi (2003). The synopsis of the contribution on the part of CNR-ILC to Pekita is taken from the project document, edited by V. Pirrelli.

lexical-semantic classification. The module will consist of an integrated battery of «machine learning» techniques for:

- the identification of candidate terms;
- the calculation of semantic similarity within the terminological repertory acquired during the preceding phase.

These techniques will be partly developed *ex novo* within ILC and in part adapted from software resources which are already available.

The role of indexing modules is that of identifying portions of text in a document (continuous or discontinuous) which are identifiable as terms or conceptual nodes of the classification resulting from the term-clustering phase. The mark up of the text with high level categories annotated in the form of metadata (XML) is a function of the possibility of indexing, searching and visualising the documents according to content category rather than in terms of simple word sequences.

5. Conclusions

Realization – still *in itinere* – foresees, therefore, the gathering and selection of a corpus of documents as a representative sample of the input of the Ontoly Builder and the specialisation of tools for automatic treatment. The analysis will include segmentation, morpho-syntactic labelling and analysis of the text in terms of non-recursive syntactic constituents (chunks) as well as the automatic extraction of dominion terminology by means of hybrid symbolic-statistical techniques, which will use, at the input stage, annotated linguistic data for the semi-automatic creation of ontologies of structured metadata, useful for the classification of documents at the input stage. The creation will take place the «cooperative pattern»: On the basis of clustering techniques applied to the extracted terms, the Ontology Builder will propose to the human user (ontology manager) new «concepts» together with their relations (eventually as potential candidates for the updating of an starting ontology specified by the ontology manager. The ontology manager will intervene in the final selection of concepts and the confirmation of modifications to the ontology proposed by the system. Finally, a «Knowledge Markup» module will be developed, which, starting with an analysis of the grammatical dependencies of the text in input, will annoatate it, with metadata in XML format, supplied by the extracted ontologies. Thus, a single integrated platform will be created, totally owned, strongly adaptable, graduated and capable of being dimensioned according to target users and the result of the synergic implementation of components which can also be used autonomously in differing contexts of application.

BIBLIOGRAPHY

Candalot, C. (2003). *La Formation aux competences informationnelles à l'Université: une voie ouverte pour le developpement des sciences de l'information et de la communication?* Bucarest, CIFSIC.

Cardinali, F. & Sampson, D. (2001). *The KOD knowledge-on-demand packaging toolkit: Sharable adaptive content objects packaging, re-usability and brokerage for learning society.* Proceedings of the Conference: European Multimedia, Embedded Systems and Electronic Commerce Conference (EMMSEC 001), Venice.

Coulon, A. (1999). *Penser, classer, catégoriser: l'efficacité de l'einsegnement de la méthodologie documentaire dans les premiers cycles universitaires.* Paris, Laboratoires de Recherches ethnométhodologiques Université de Paris.

European Commission (2000). *A Memorandum on Lifelong Learning.* Recovered on 13th April 2004 from: http://europa.eu.int/comm/education/life

Garro, A., Leone, N., Ricca, F. (2003). *Logic Based Agents for E-learning.* Proceedings of the Conference: IJCAI (International Joint Conference on Artificial Intelligence) Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems, Acapulco.

Guarasci, R. & Rovella, A. (2004). E-doc: documentazione ed e-learning, *Culture del Testo e del Documento* (13): 19-27.

Guarino, N. & Giaretta, P. (1995). *Ontologies and Knowledge bases: towards a terminological clarification.* Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing: 25-32.

Hartley, D.E. (2000). *On-demand learning: Training in the new millennium.* Amherst MA, HRD Press.

Jolion, J.M. (2001). *Indexation.* Paris, Hermes.

Karagiannidis, C. & Sampson, D. (2003). *Re-using Adaptation Logics for Personalised Access to Educational e-Content.* Proceedings of the Conference: Workshop on Adaptive Systems for Web-Based Education. 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga.

La Monica, E. (2002). *La Formazione a distanza: un modello di didattica universitaria.* Palermo.

Peters, C. & Picchi, E. (2003) Bilingual lexicons, parallel and comparable corpora: creating the basis for crosslanguage information retrival. In: Zampolli, A., Calzolari, N., Linguistica Computazionale, I.L.C. and Computational Linguistics, special issue. vol. XVIII-IXX, pp. 573-596, Pisa-Roma, I.E.P.I, 2003.

Peters, C. & Picchi, E. (2003). Bilingual lexicons, parallel and comparable corpora: creating the basis for crosslanguage information retrival. *Linguistica Computazionale, I.L.C. and Computational Linguistics special issue.* vol. XVIII-IXX:573-596, Pisa-Roma, I.E.P.I.

V.A. (2002). *Metadata for e-Communities: Supporting Diversità and Convergence.* Proceedings of the Conference: 10th International Dublin Core Workshop, Florence.