Abstract – We present an approach to the production of guidelines on cancer biomarkers based on the CommonKADS methodology. This approach aims to define a conceptual information retrieval system that can be accessed in Internet and Intranet, where an author can test the conceptual content of his own work and where experts charged to define the guidelines can find a valuable help. A software tool based on this approach has been set up and may be found at http://bartolomeo.medinfo.dist.unige.it.

I. INTRODUCTION

Clinical practise guidelines (GLs) are more and more developing in order to reduce the delay in transferring research findings into practise for similar pathology conditions. This trend is particularly felt in cancer domain, where the decrease in mortality can be deeply influenced linking state-of-art knowledge and best medical practice.

Literature review in oncology is a very complex task, as well as the boot for guidelines development. The central problem originates by the fact that present day growth of cancer literature is of exponential nature (breast cancer is a typical example) [1]. Thousands of experimental and clinical researchers are actively engaged in cancer research. Specifically, one among the many aspects, which are being increasingly investigated, is the identification of prognostic factors for breast cancer. Specialised journals continuously report a flood of interesting results sometimes conflicting; the evaluation of these results including the identification of possible lacks in several aspects of the research is a prerequisite for the production of a prognostic guideline.

Although knowledge modelling has been rapidly changing in the last few years because of the introduction of hypertext and multimedia techniques, the knowledge resulting from a research activity is likely to be still presented and recorded by the researchers in natural language (usually English) for many years. This does not exclude that an electronic format of the paper that is of the journal in which the paper has been published is also made available for many journals, in a short time.

The first steps of a guideline production are the reading of the literature on a specific topic, the conceptualisation and the understanding of it, and the assessment of whether the concepts contained in it should be taken into account to modify the guidelines which concern this topic. The key point of this process is the knowledge conceptualisation and the result understanding (fig.1, fig.2). With respect to this, a system which guides the conceptualisation of the knowledge contained in a paper should be a useful tool in the production of a guideline. In this work, we address the definition of a conceptual information retrieval system for the production of guidelines on prognostic factors of breast cancer; below, we present our approach to the system design, the global system architecture, the conceptual representation and the part of the system already implemented.

Several medical models may be related to guidelines production, addressing different aspects, such as, for example:

- UMLS (Unified Medical Language System) [2-4],
- SNOMED (Systematized Nomenclature of MEDicine) [5];
- ICD9 (International Classification of Diseases with Clinical Modifications) [6];
- GMN (Gabriel Medical Nomenclature) [7];
- GALEN (Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine) [8].
II. METHODS

Global system architecture

The architecture of the global system we propose is shown in fig.3 and it is based - to some extent - on the general architecture of DR-LINK [9]. We propose a four layered architecture. The text processing layer processes the information source (IS) database in order to select the papers which contains the topics addressed by the guideline. This can be performed by a traditional information retrieval (IR) system which can select from IS the papers which address a specific topic (modelled in an IS model, in the form of boolean expressions of keywords). The service given to the assessment layer (fig.3) is a queue of papers which satisfies the information needs of the addressed GL modelled in the IS model.

Figure 1. Abstracting the knowledge about a specific topic of interest (i.e. cancer biomarkers). An information developer may access a conceptual bibliographic research and produce a conceptual representation of papers concerning a topic and to produce a conceptual representation of a paper and its quality self-verification.

Figure 2. The domain knowledge of a guideline database represented in a entity-relationship syntax (based for example on CML).
The assessment layer performs the conceptualisation of the papers selected at the text processing layer. Each selected paper is abstracted and compared with an IS conceptual model, which defines the syntax, the relations and the constraints of the concepts which should be contained in the current selected paper in an abstract formalism. The services provided to the upper layer is the result of this comparison, which is assessed in a decision which contains the abstracted structure of the selected paper as well as parameters indicating its compliance with quality criteria.

The learning layer stores the decisions given from the assessment layer. The service of this layer is to update the guideline. An assessment task compares the GL database with each element of the queue; the result of the assessing may be either an updating of the GL database with the results presented in the paper or a decision of not taking into account those results.

The GL database is the real core of the system; specifically its formalism as well as its level of abstraction has to be defined in order to comply a trade-off between natural language and computer language structure. The need of a representation which is not distant from the natural language is due to the very fact that a computer guideline should always have a parallel one in natural language (a sort of guideline «book»). Besides, it is feasible that in the next few years multimedia and/or networking techniques will be used to describe guidelines; in this case, although guidelines are still expressed in natural language, a greater effort in structuring the GL contents will be made. Finally, guidelines might become part of knowledge based systems (KBS) as well as of other applications. So in this case GLs should be expressed in a language, which is highly structured in order to make inferences and to allow a fast access to a specific content; moreover a highly structured language will be necessary to automate the guideline building process. In our opinion, a formalism that may meet these needs is a conceptual representation in term of modelling language.

The CommonKADS Library for Expertise Modelling

Our approach to the conceptualisation of the problem follows the CommonKADS (Knowledge Acquisition and Design Structuring) Library for Expertise Modelling [10]. KADS is a comprehensive methodology for the development of KBS. KADS was set up during European research projects (ESPRIT P1098 and P5248) and is becoming a «de facto» standard for KBS development in Europe. KADS methodology is currently under evolution and CommonKADS is the name of the last version, which was available at the beginning of this work.

A CommonKADS product model is a set of interrelated models from the CommonKADS model set. This model set specifies six model templates to describe different views on, or models of, a problem solving context. The six models are:

1. the Organisation Model, which describes the organisational context in which the knowledge based activities related to the problem;
2. the Task Model, which describes the tasks and activities that are executed to realise the organisational function;
3. the Agent Model, which collects the relevant properties of the different agents performing the previous defined tasks;
4. the Communication model, which describes transactions among agents;
5. the Expertise Model, which describes the knowledge of an agent relevant to a particular task;
6. the Design Model, which describes the realisation of the problem solving behaviours described in the Expertise and Communication model in computational and representational terms.

In this work, the CommonKADS Expertise Model is addressed. In CommonKADS the different roles are captured in three basic independent knowledge categories:

1. the domain knowledge specifies form, structure and contents of the specific domain which is relevant for an application (structures made by concepts and relations);
2. the inference knowledge specifies the primitive steps in reasoning (inferences) in an application and the knowledge roles that refer to classes of domain knowledge statements manipulated by the inferences;
3. the task knowledge describes a recursive decomposition of a top-level task in sub-tasks, specifies what it means to specify these tasks, and describes when the sub-tasks are to be executed in order to achieve their parent task (control).

These descriptions are made by the CommonKADS Conceptual Modelling Language (CML) [11]. For many applications, these three layers are the GL text processing layer are sufficient to build the system. Moreover, the CommonKADS modelling libraries capture knowledge engineering expertise in terms of reusable elements that are potentially modelled in a project.
There has already been some work [12] that has used KADS for cancer treatment modelling (specifically of Northern California Group Protocol 2083-1 for small-cell lung carcinoma treatment in ONCOCIN). In that work, the results of the application of the KADS methodology were not completely satisfactory. In another oncology approach always in oncology, CommonKADS has been taken into account as a good methodology for knowledge acquisition [13]. In this work, as in [13], the modelling aim was quite different from the work by Linster et al [12]. For example: we were not biased by preconceptions of an already existing system; the application to be modelled is quite different, we started conceptual modelling by a knowledge acquisition process; and we needed a methodology to guide the work. Thus in our opinion, it was worthwhile to use CommonKADS. The model presented here has been built following CommonKADS methodology; at this stage, we have not used any specific CommonKADS tools.

Figure 3. The architecture of the proposed system for the production of guidelines.
III. RESULTS

We have focused on the GL text processing and the assessment layers, which, in our opinion are the «bottle-neck» in guideline production. A prototype environment for these two layers has been implemented.

In the GL text processing layer, we have chosen to implement «ad hoc» software for the conceptualisation which is completely guided by the user; the software uses a hierarchical database of concepts expressed following the CML by CommonKADS. The conceptualisation is guided by form and common gateway interfaces (CGI) in an Internet Intranet environment. Besides, the application is still guided by the user, since, even if there are many recent good results in automated concept managing from natural text, this is still a distance objective to be reached. We have so defined a prototype environment. The larger window on the left contains the document that has to be processed for conceptual extraction from papers, which can be either searched in Internet or a local file.

An exhaustive stand-alone assessment layer could solve many quality problems at the source of the information: for example, the very author may test his work and verify the conceptual content, as well providing a conceptual «summary» of the work, easier to be handled by expert. Abstracting and conceptualising the information is a crucial point. If an author believes that her/his paper is worthwhile to contribute to a guideline, the following procedures (from manual to automated) can solve the abstraction: apart from the paper, she/he can fill a previously defined «quality form» resembling her/his research findings; she/he should process the product with an «ad hoc» software which is able to conceptualise it; the conceptualisation can be guided by the user or automated. We have also implemented a rule based environment in Clips [14] allowing an automated comparison of the different conceptualisations.

Fig. 4 shows the frame oriented interface of the text layer prototype. The larger frame on the left-hand side is meant to show the text of the current paper which is to be conceptualised or searched on the Internet. The larger frame on the right-hand side is meant to show and to guide the conceptualisation of the paper. The other frame on the right-hand side (bottom) is meant to show a glossary of concepts and of their relationships which can help conceptualisation. The frame on the left-hand side (bottom) shows some control links which change according to the current task. More specifically, in a typical conceptualisation work, a knowledge engineer or an expert working at guideline definition will have the possibility of analysing the results from a scientific paper (if available in electronic format) on the left-hand side of the window. She/he will be interactively guided in the conceptualisation by a tree of concepts whose main leaves are written on the right-end side (top). Then, she/he will be asked to fill some forms to conceptualise the paper and its results and at the end the results will be saved in CML. This CML code is ready for its evaluation in the CLIPS based assessment layer. A software tool based on this approach has been set up and may be found at http://bartolomeo.medinfo.dist.unige.it.

Figure 4. The HTML forms used for the text processing layer.
IV. FUTURE DEVELOPMENTS

We are studying at the integration of the rule based automated system (assessment layer) using the Jess environment [15] (a Java version of Clips environment) in order to obtain an agreeable user friendly multimedia interface of the system.

ACKNOWLEDGEMENTS

The present work has been carried out within the ECOLE/GRIP (European Collaboration in Oncology Literature Evaluation) project of Telematics Application, grants n. HC1013.

We thank Dr. M.Gion (Ospedale Civile, Venice, Italy) for providing documentation on biomarkers and Dr. Daidone (National Institute for Research on Cancer, Milan, Italy) for providing published papers on biomarkers for the validation of the system.

REFERENCES