Empirical methods for ontology engineering in bone marrow transplantation

Brigitte Endres-Niggemeyer
Fachhochschule Hannover
University for Applied Sciences
Department of Information and Communication
Ricklinger Stadtweg 120
D-30459 Hanover, Germany
phone +49 511 92 96 606
fax +49 511 92 96 610
Brigitte.Endres-Niggemeyer@ik.fh-hannover.de

Summary
I report here on a small preparatory study in ontology construction. It explored T-cell depletion, a special issue in bone marrow transplantation (BMT). The target ontology aims at supporting WWW summarizing in BMT. As a result I propose a case-specific research design backed up by thesaurus construction methods and a systematic field research grounding. It integrates domain experts / future users and system developers. At conceptualization and formalization time, the research procedure joins current ontology engineering methods and formats.

1. Ontology engineering seen empirically

The methods papers in ontology engineering (Blazquez et al. 1998, Fernandez et al. 1997, Uschold und Gruninger 1996) are helpful wherever knowledge is to be conceptualized, formalized and stored, but they do not place much emphasis on knowledge acquisition. This may not be necessary in domains where all knowledge is at hand or can be obtained from standard sources, where this knowledge is stable and not compartmentalized with respect to tasks.

In BMT this is not the case. It therefore seems wise to envisage a structured empirical investigation. Good empirical research practice prepares a field study by means of a pre-study. This is just as advisable in ontology engineering. Discovering the concepts of a field and their relationships is a sort of inductive modeling, no matter whether we represent them in an ontology or not. Inadequate empirical methods risk compromising the results of knowledge acquisition before we can include them in an ontology. Furthermore, the success of ontology engineering depends upon the cooperation of the field subjects, also called domain experts, who know and define their concepts. In Hanover as elsewhere, the BMT domain experts work in their labs, on wards and in offices. There, they are willing to share their knowledge, and there they want to use the target system and its ontology.

These points should motivate readers to follow an empirical researcher in planning ontology design.

In many respects, ontologies seem akin to thesauri and library classifications, such that principles of thesaurus and classification construction apply (observation by Vickery 1997). During my pre-study in ontology engineering, I used the guidelines for building library classifications or thesauri (described by Buchanan 1979; Aitchison and Gilchrist 1997). First, two papers explaining T-cell depletion (Hertenstein et al. 1998; Kernan 1994) were exploited for relevant concepts. After that, five user questions about T-cell depletion were answered by consulting Hertenstein (1998). The processing of the intended summarizing system was simulated by hand. Whenever knowledge gaps showed up, I entered new concepts, facts and

inferences. At the end of the test, the would-be ontology had some 600 plus concept records. It is stored as a relational database. I noted the agents needed for text interpretation and summarizing, be they already existing (from earlier research – see below) or required in the new domain.

2. Some stock-taking

Let us now inspect the environment, or, ecologically speaking, the habitat (for more background in ecological theories see Clancey 1997) into which we intend to put the ontology: the task and structure of the summarizing system, special characteristics of the domain, the domain experts at their workplaces, the research history of the approach, and the scientific possibilities of the researcher. This prepares the discussion of an empirical research design that adapts to the ontology engineering task at hand. There, we first sketch the target ontology, and after that the procedure of ontology construction.

The task

The target ontology is to be used in a WWW summarizer for physicians in BMT. They are interested in fast knowledge support from outside when making therapy decisions. Conventional information retrieval is too slow for their purposes. The summaries of the proposed system should be correct, because doctors have no spare time to work their way through unreliable or irrelevant information. A high-quality knowledge processing approach inspired by human cognitive strategies (Endres-Niggemeyer 1998) seems appropriate.

Fig.1 illustrates that the ontology is a key resource in the summarizing system. A user starts a summarizing process by formulating a search scenario with terms of the ontology. The scenario is mapped onto a WWW search form. It is passed to a (meta)search engine and to Medline retrieval. Where the retrieved Medline records refer to online journal articles, these are included. As soon as the results arrive, a text retrieval component screens them and highlights promising passages. They are interpreted in a restricted way and summarized with respect to the question scenario. Relevant source text clips are organized and put into the question-answering scenario. They are all linked to their home document passage. Now the user interprets the summary, possibly digging into the original papers. The system is ready for the next summarization round.

The domain

BMT has a key function in many cancer therapies. The domain is small, but it is evolving quickly. By looking (mostly in vain) for BMT concepts in Medical Subject Headings (MeSH), a big and popular medical thesaurus, readers can experience themselves how specialized the domain is.

Besides their core know-how, physicians in BMT refer to a wide range of specialties in medicine, for instance when treating infections, which are a major risk during transplantations. The main avenues of knowledge distribution in BMT are scientific journals and conferences, whereas stable and codified traditional knowledge sources, such as up-to-date textbooks, are rare.

BMT knowledge is bustling and neither stable nor integrated. It is mostly made up of detailed facts and relationships and must be mapped to the ontology in its existing form. There are general rules that sweep wide ranges of knowledge, but their contribution remains limited. Knowledge of today will be superseded soon by newer and better approaches. The intended ontology must keep up with its domain, i.e. it must be flexible, modular and easy to maintain. We need a teleological ontology that is rich in terminological detail (Hovy 1997).

Users and domain experts at work
The pilot users of this system and ontology are Hanover haematologists. They are willing to share the development effort, taking over the responsibility for medically correct system results. Typically, the physicians work in teams of experts, but they are aware that not all knowledge is locally available. What knowledge they need may differ widely as they go through specific situations. No matter whether they encounter a problem at the lab, on a ward, or when advising an out-patient, they deserve an ecologically useful ontology for their conceptual support.

The knowledge acquisition procedure has to comply with their demands by considering a good "representative" choice of situations and picking up the conceptual material for dealing with them. Since future users will cooperate in shaping their system, user-centered design methods are required.

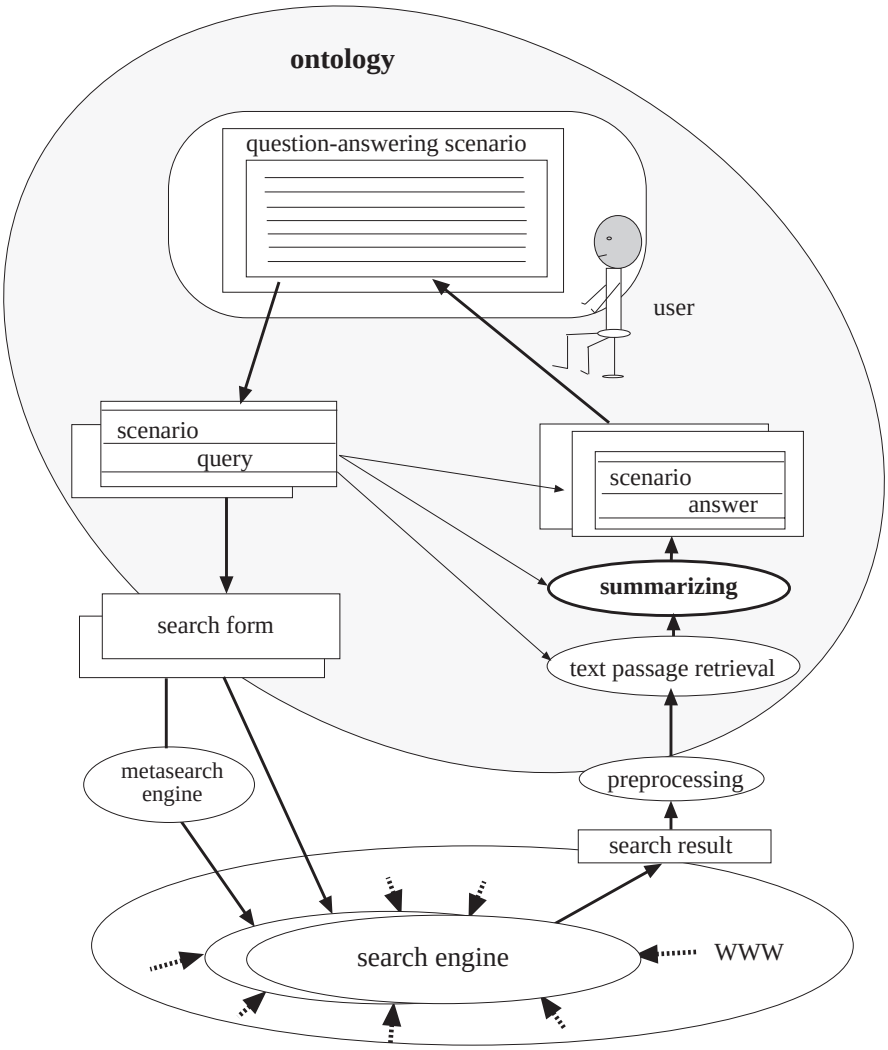

Fig.1. Ontology-based summarizing in the WWW

Research pre-history
The summarization agents at the core of the intended system are known in principle from an earlier empirical cognitive model of expert summarizing (Endres-Niggemeyer 1998). It in-

cludes an implemented simulation of four expert summarization sequences. 552 expert summarization strategies were described, some 40 of them were implemented in great detail as knowledge-based agents. Now this work is to be applied to a real-world domain. The empirical model comes with its scientific history, bringing in a background of qualitative field research or grounded theory development (Mayring 1990, Glaser and Strauss 1980). Since some conceptual modeling must go on in BMT, it is advisable to remain coherent in empirical modeling techniques.

The researcher
Besides bringing in her research history, the investigator also learns from her teaching of thesaurus construction. Since thesauri and library classifications are less formalized than ontologies, their construction guidelines focus on the acquisition of domain concepts, a good representation of the domain, and its conceptual organization. Where they have their strong points, they may evidently supplement an ontology development methodology.

3. Methods planning

After looking through the affordances for ontology construction in BMT, let us examine the exploratory pre-test. It started from the already mentioned thesaurus construction principles (Buchanan 1979; Aitchison and Gilchrist 1997) and at conceptualization and formalization time would dock computer-oriented ontology engineering techniques. It approximated the research procedure described below.

A sketch of the target ontology
The structure of the domain suggests modularizing the ontology. This can be done by referring to Penman. According to Penman (Penman 1989) the ontology comprises an upper model and any number of embedded lower models. The local expert group proposes the issues they want to possess a lower model. Some examples of current issues are:

- T-cell depletion
- high-dose therapy for breast cancer
- qualitative polymerase chain reaction (PCR)
- stimulation of donors with G-CSF (granulocyte colony-stimulating factor)

The issues are not liable to keep to their lower models. Instead, many concepts from surrounding areas of medicine and common knowledge are used for dealing with an issue, as illustrated for T-cell depletion by Fig. 2. Co-occurring concepts may, for instance, be at home in biochemistry, T-cell depletion itself, in its surrounding discipline of haematology, or they may come from the realm of general scientific argumentation, let us say from statistics. It would be hard to defend a flat knowledge organization with all lower models at the same level, and just as difficult to believe in a strict hierarchy of scientific disciplines. More realistic is some hierarchical embedding structure as shown in Fig. 2. It reflects a standard view on medical disciplines. Concepts can be allocated to any suitable set of lower models.

The BMT ontology links up to WordNet (Miller 1995), because frequently, general concepts of English are involved. Even *remission*, a key concept of oncology and haematology, has a useful general reading in WordNet (cf. Fig. 3). The MeSH descriptor helps to access general medical knowledge from Medline.

The *remission* record is simple, but it illustrates this key feature of the planned ontology. It supports for instance:

1. the program that prepares a form for Medline search. The program finds the MeSH descriptor and tree number in the last two slots.
2. an inferencing agent. The agent sees from the FactsAndRules slot that remission duration implies remission.

3. a speaker of German. (S)he finds the German term in the GermanTerm slot.
4. a text interpretation agent. The agent sees from the LexicalEquivalents slot that two readings of *remission* in source texts are equivalent.
5. a human user. (S)he would learn from the Equivalents slot more or less synonymous expressions for the concept *remission.* Currently, we have none.
6. a text interpretation agent that uses the Definition slot. There it finds a formalized description of features that must be considered when interpreting occurrences of a concept, such as contexts to which the concept belongs. The slot is still empty.

Lower Models						
General						
	Tech					
	Science	Life Sciences	BioChem			
			Genetics			
			Medicine	Onco		
				Haema	BMT	TCdeplet
				Immun		
				Cyto		

Fig. 2. Lower models needed for T-cell depletion

ConceptName	remission
Sort	physiological state
Equivalents	
GermanTerm	Remission
Definition	
SuperConcept	physiological state; clinical result
SubConcept	complete remission; first remission; second remission
FactsAndRules	remission duration -> remission remission rate -> remission
LexicalEquivalents	remission duration <-> duration of remission
LowerModel	Haema; Onco; General
WordNetEntry	remission noun 1
MeSHTerm	remission induction
MeSHTreeNumber	E2.860

Fig. 3. A sample record: *remission*

The ontology provides system participants such as physicians, search machines and retrieval systems, summarization agents, and others, with conceptual knowledge. These very different players need knowledge tailored to their use. In order to make them cooperate, in spite of approaching a piece of knowledge under different task-oriented views, the ontology must keep all knowledge about a concept together and integrated.

An empirical research plan for ontology engineering in BMT

The empirical research plan for ontology engineering in BMT is inspired by methods of thesaurus and classification construction. They insist on grounding the thesaurus on data from

the literature, and on testing it as early as possible in an application setting, normally in retrieval. We supplement them with more sophisticated empirical research techniques. This permits the desired formative evaluation of the ontology under construction.

Empirical methods must definitely give way to knowledge representation techniques when representations are formalized for machine agents. Around this point, ontology construction may be supported by a technical tool such as ODE (Blázquez et al. 1998), and join a current format, for instance provided by Ontolingua (Farquhar et al. 1996) or OntoSaurus (Knight and Luk 1994, Swartout et al. 1996).

For every issue defined by the expert group, the knowledge is represented in 12 working steps. In the initial phase (steps 1-4) a basic stock of concepts is built up. Then the research procedure loops through an incremental phase with formative evaluation (steps 5-12). In step 13, the module is integrated into the overall ontology:

1	2 – 3 current relevant papers or book chapters are exploited to obtain an initial stock of concepts
2	concepts are supplemented with WordNet knowledge
3	if available, MeSH descriptors are added
4	the meaning of the concepts is made explicit and they are formalized and represented for the use of different players
5	users set up search scenarios
6	from user search scenarios queries are derived, the search engines are started
7	the found documents are summarized
8	the summarization results are integrated into the question/answer scenarios
9	summaries are checked for failures by physicians and technical team members
10	the knowledge representation is improved
11	agents are adapted or created
12	back to step 5 as often as needed
13	a new partial ontology is integrated into the existing one

By going through the field issue by issue, the concepts needed from neighboring and more comprehensive areas of knowledge will accumulate as well.

4. Conclusion

I propose an empirical ontology engineering procedure that attempts to ease the ontology construction process under specific conditions. Adaptation to the given research situation will not compromise the results, for instance in terms of transfer to other BMT groups, or of integration into a larger ontology. The inspiration from thesaurus construction guidelines and systematic empirical methods may, however, contribute to a better representation of the domain, speed up the research process, and integrate users from the early beginnings of system design, with obvious consequences for acceptance. It may be difficult to prove that empirical methods perform better than standard ontology engineering approaches, but in any case there is nothing bad in methods that stand up against principles of thesaurus building and empirical research methods as well as against those of ontology engineering. This can strengthen their theoretical backup. Qualitative field research and user-centered design methods agree in making research a cooperative venture of field subjects and investigators. Even if they had no

other advantages than improving quality of life during ontology engineering, they are worth considering.

5. References

Aitchison, J.; Gilchrist, A. (1997): Thesaurus Construction and Use: A Practical Manual. 3rd edn. London: Aslib.

Blázquez, M.; Fernández, M.; García-Pinar, J.M.; Gómez-Pérez, A. (1998): Building Ontologies at the Knowledge Level using the Ontology Design Environment. KAW'98, Banff, Canada. http://delicias.dia.fi.upm.es/miembros/ASUN/kaw98.ps.zip.

Buchanan, B. (1979): Theory of Library Classification. London.

Clancey, W.J. (1997): Situated Cognition. Cambridge: Cambridge University Press.

Endres-Niggemeyer, B. (1998): Summarizing Information. Berlin: Springer.

Farquhar, A.; Fikes, R. ; Rice, J. (1996): The Ontolingua Server: A Tool for Collaborative Ontology Construction. http://ksi.cpsc.ucalgary.ca/KAW/KAW96/farquhar/farquhar.html.

Fernández, M.; Goméz-Pérez, A.; Juristo, N. (1997): METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Workshop on Ontological Engineering. AAAI 1997 Spring Symposium, Stanford, USA. http://delicias.dia.fi.upm.es/miembros/ASUN/SSS97.ps.

Glaser, B.G.; Strauss, A.L. (1980): The Discovery of Grounded Theory: Strategies for Qualitative Research. 11th edn. New York: Aldine Atherton.

Hertenstein, B.; Arseniev, L.; Novotny, J.; Ganser, A. (1998): A Comparative Review of Methods for T Cell Depletion in the Prophylaxis of Graft-versus-Host Disease. BioDrugs 9:2, 105-123.

Hovy, E. (1997): What would it Mean to Measure an Ontology? Internal Paper, ISI Marina del Rey.

Kernan, N. A. (1994): T-cell Depletion for Prevention of Graft-versus-Host Disease. 124-135 in Forman, S.J.; Blume, K.G.; Donnel Thomas, E. eds. Bone Marrow Transplantation. Boston: Blackwell.

Knight, K.; Luk, S. (1994): Building a Large Knowledge Base for Machine Translation. 773-778 in AAAI-94. 12th National Conference on Artificial Intelligence, Seattle, WA.

Mayring, P. (1990): Einführung in die qualitative Sozialforschung [Introduction to Qualitative Social Research]. München: Psychologie-Verlags-Union.

Miller, G. (1995): WordNet: A Lexical Database for English. Comm. ACM 38:11,39-41.

Penman Project (1989): PENMAN Documentation: The Primer, the User Guide, the Reference Manual and the Nigel Manual. Technical Report. ISI Marina del Rey CA.

Swartout, B.; Patil, R.; Knight , K.; Russ, T. (1996): Toward Distributed Use of Large-Scale Ontologies. http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html.

Uschold, M.; Gruninger, M. (1996): Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11:2, 93-136.

Vickery, B.C. (1997): Ontologies. Journal of Information Science 23:4, 277-286.