Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

CONSTRUCTING AN ONTOLOGY FOR WWW SUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT)

Brigitte Endres-Niggemeyer Brigitte Endres-Niggemeyer[1], Bernd Hertenstein[2], Claudia Villiger[1], and Carsten Ziegert[2]
[1] Fachhochschule Hannover/University of Applied Sciences
Dept. of Information and Communication, Ricklinger Stadtweg 120
D-30459 Hanover, Germany
{Brigitte.Endres-Niggemeyer, Claudia.Villiger}@ik.fh-hannover.de
[2] Medizinische Hochschule Hannover/Hanover Medical School
Dept. of Hematology and Oncology
D-30623 Hanover, Germany
Hertenstein.Bernd@mh-hannover.de
Carsten.Ziegert@ik.fh-hannover.de

http://summit-bmt.fh-hannover.de/

ABSTRACT

We describe an ontology for WWW summarization in Bone Marrow Transplantation that is currently under construction. It is text-based and qualifies as a grounded ontology. In addition, it is user-centered. For stating medical knowledge, we use first-order logic extended with contexts. The ontology is prepared to serve query scenario formulation, text passage retrieval and summarization proper in a summarization system. It will be stored and managed by an XML database server. Currently, our ontology is developed to the point that its features can be demonstrated.

1. INTRODUCTION AND BACKGROUND

In this paper, we describe an ontology under construction. We focus on content engineering problems. The ontology's aim is to support summarization from the WWW for Bone Marrow Transplantation (BMT), a specialized and life-critical area of hematology. The ontology will be used by physicians, by summarizing agents and by other system participants, such as a text retrieval component described below. It has to comply with the demands of the BMT domain, the physicians using the ontology, and the task of summarizing, i.e. of text processing in a knowledge-based style.

Before we describe our work, we prepare some common ground with our readers. We refer to the literature, we explain the system environment of the ontology as far as needed here, and we give reasons for our own design and presentation decisions.

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

1.1 Role of the ontology in the summarization system under construction

The target ontology is a dense representation of domain knowledge in Bone Marrow Transplantation (BMT). The summarization process it supports follows strategies of human expert summarizers (Endres-Niggemeyer, 1998). Human summarizers rely heavily on knowledge, and so does our target system. Most of the factual and linguistic knowledge is incorporated in the ontology. For every domain concept, the ontology also encompasses statements of relevant knowledge about it. These axioms are formalized so that computer programs, e.g. agents, can exploit them. One of the main differences between thesauri and ontologies is that ontologies state propositional knowledge about their concepts, whereas thesauri are restricted to core paradigmatic relations such as generic ones. From an information retrieval perspective, these statements exclude wrong concept combinations. During summarization, a match with propositions helps to establish the relevance of candidate statements retrieved from texts with respect to the current question. For concept identification in running text, the BMT ontology stores lexical equivalents of concepts and the paraphrases by which they are expressed. Like a classic thesaurus, the ontology also conveys information for human users, in particular for query scenario formulation.
In the following it is described in some detail how the ontology supports query formulation for information retrieval, text passage retrieval, and summarization proper.

Query scenario formulation and query expansion

For question-oriented summarization, we need reasonably precise queries. The better the question, the better chances the system has to come up with a helpful response. In order to have users state their queries in a well-structured fashion that the system can interpret, we provide scenarios, query forms specific for types of questions and situations. Into these forms, users fill concepts of the ontology, formulating their knowledge about current setting and questions about the knowledge they need. The scenarios accept statements and questions. Users can browse the concept, synonyms, hypernyms, hyponyms, and description fields of ontology records (see section 2). We provide definitions and descriptions of concepts assembled from many WWW sources and a few others.

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

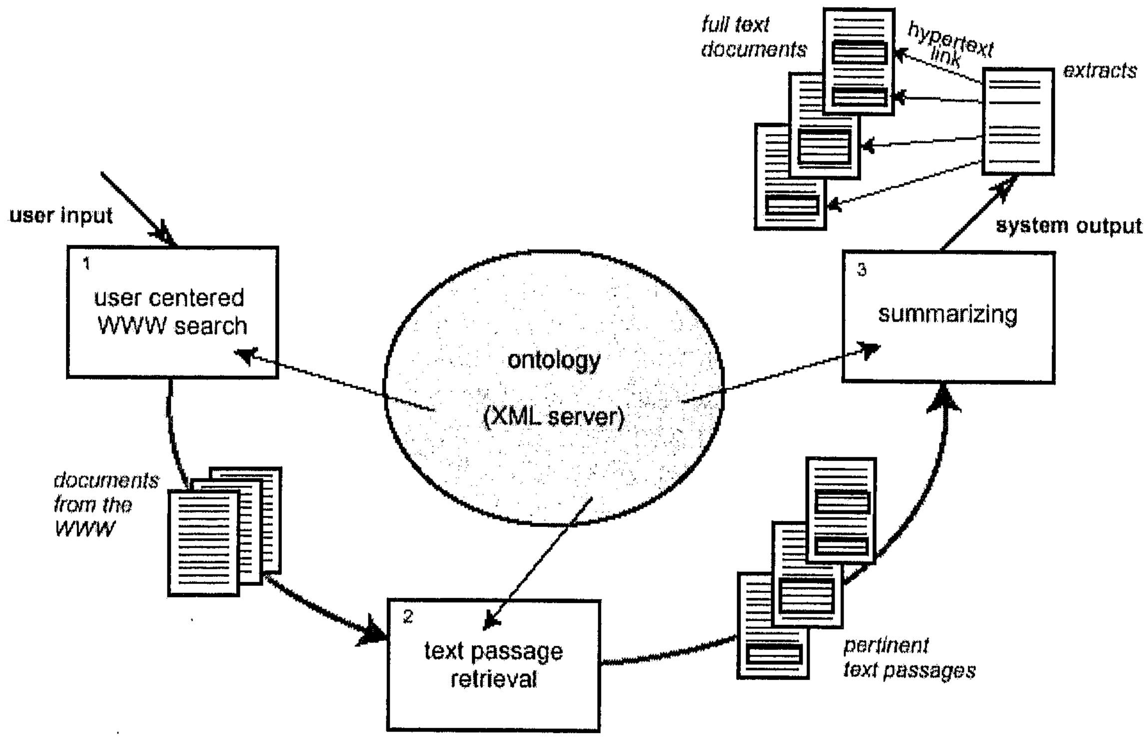

Figure 1: Roles of the ontology in the summarization system: support of query scenario formulation, text passage retrieval and summarizing

After being reformulated in predicate logic form, the query scenario is enriched with related concepts from the ontology, and adapted to the search form of involved search engines (see processing step 1 in figure 1).

Text passage retrieval

As soon as retrieved documents arrive, they are roughly checked for relevance by scanning them for passages that include the query terms. The preliminary relevance check is performed using the concepts of the ontology (see processing step 2 in figure 1).

Summarizing

Summarizing means reducing information to its most important (relevant) points (Endres-Niggemeyer, 1998). Most often, the source information is given as a text, and the resulting summary is a short text. Summarization mostly operates at sentence level, and summaries are composed of relevant statements (often sentences), as opposed to passages including unrelated relevant terms as in text passage retrieval. When the summary is to answer a question, only statements that correspond to the question at a sentence level can be relevant, all others are not. So passages that contain ontology concepts used in the query are checked for question-related relevance at sentence level. For this purpose, the ontology stores propositional knowledge with every concept, practically speaking logical propositions that use the current concept as an argument. In figure 1, summarizing occurs as processing step 3.

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

1.2 Focus on systematic empirical knowledge acquisition

In this paper, we focus on content engineering problems of our ontology. We emphasize the empirical knowledge acquisition, not so much the formal and technical part of ontology engineering. In the methodology literature of ontology engineering (Blazquez et al., 1998; Fernandez et al., 1997; Uschold & Gruninger, 1996), ontology content acquisition is a minor concern. This may be no problem for experimental approaches, but for real-world ontologies, we need a more substantial grounding on evidence. Therefore we propose a systematic empirical research procedure for ontology content, referring to the experience of thesaurus construction (Aitchison & Gilchrist, 1997) and of grounded theory development (Glaser & Strauss, 1980; Strauss & Corbin, 1990). First and foremost, we ground the ontology on domain texts, in line with others (Aussenac-Gilles et al., 2000; Golebiowska et al., 2001) who promote an approach to knowledge modeling based on texts. We argue that papers published in core journals of the domain are safer than other knowledge sources, especially oral information sources such as interviews or expert meeting results, because their contents and wording have been checked by the reviewers of the journal. The idea is to pick up not only concepts, but to formalize text statements comparable to sentences in order to obtain ontology propositions or, technically speaking, first-order logic expressions.

1.3 User-centered/participatory approach

Our target ontology is user-centered, i.e. an ontology that is endorsed by its users because they contributed to its design and content so that they are its responsible co-authors. Since the ontology is central for the future system's practical value and acceptance, it is only consistent to expand the realm of user-centered system design (Norman & Draper, 1986) from the user interface to the ontology, the system's main knowledge representation. During development, our users participate in terms of user-centered or participatory design in the system dedicated to them, especially in the ontology, which is most vulnerable to poor subject knowledge of system developers. Hence, it is better to remember the principles of formative evaluation (Scriven, 1967) and to entrust the medical correctness of the ontology to its users' judgement as soon as ontology records are produced. The less productive alternative would be to wait for a summative evaluation of the final product.

1.4 Target representation format

Since BMT is a hazardous domain, we adhere to logic representation formalisms (Fox & Das, 2000). We state our facts and rules in first-order predicate logic and enable safe and simple conclusions, as opposed to fuzzy approaches, which risk increasing the evasiveness of knowledge presentation often observed in the medical literature.

In medicine many statements are valid only if the limits of their scope are respected. Perhaps more than elsewhere, we have to represent preconditions, explicit and implicit ones. To meet these requirements, we apply the context formalization proposed by (McCarthy & Buvac, 1993). Context expressions are composed of several predicates (propositions). They set up the frame of thinking and state the core information. A context expression as a whole must be true. Last but not least, our technical implementation of the ontology owes much to the

ISSN: 2324-9773

authors of SHOE (Heflin et al., 1999). Our ontology is represented in XML (W3 Consortium, 2000) and managed by an XML database server.

1.5 Time of presentation

We are presenting our BMT ontology when about a fifth of it is realized. The message is that a considerable part of the execution has been done and that the ontology features are stable enough for presentation. Nothing really new can be expected by waiting for an implementation rate of some 80%, but only more implementation of the same. Please consider three arguments:

- There will be no definite state of completion of the ontology, but at best a smooth transition from ontology construction to updating and maintenance. We have to decide about describing it on other grounds than completion.
- Slight adaptations excluded, an ontology design cannot be changed during execution. An ontology is a highly organized object, possibly as complicated as a car. Nobody would start car production without a suitably detailed plan or design. This plan is open to restricted local adaptations, for instance one may exchange the brakes of a certain producer against cheaper or better ones of his competitor. But it would be disastrous to deviate from the design in major issues. Adherence to the design in all major points is a conditio sine qua non for success in production processes relying on the cooperation of many persons and the functional integration of many parts, be it in a car or an ontology. This is why one cannot ignore an ontology design during production. Unless one fails totally or in part, the ontology will conform to its design. As design precedes realization, one can explain the interesting points earlier by referring to the design rather than by waiting for a complete execution some years later. Consequently, we should present our ideas at an early stage.
- A reasonably good design keeps its value for later researchers. Examples for this effect are the SUSY summarization system (Fum et al., 1985), which was never implemented, and HEARSAY II (Erman et al., 1980), a classical speech recognition system that survives because of its pilot use of a blackboard architecture.

1.6 Style of presentation

The presentation style of this paper reflects our empirical approach:

Form follows function in inductive empirical modeling
Empirical modeling is important to system design. In our inductive empirical approach, form follows function and content. The function an ontology has to serve is roughly defined by its domain and task. This is why we talk about the tasks of the ontology in its domain before we describe the ontology itself, whose features are determined by domain and task constraints. This sequence of presentation may contradict the habits of colleagues who are less concerned with an empirical justification of their systems, but there are good reasons for keeping presentation in parallel with the inductive style of empirical argumentation.
We describe the way we set up our ontology for empirical researchers, so that they can understand what happens, criticize it and perhaps try it out or find a better solution. Our

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

description is far from a specification for programmers, but near to the degree of detail provided in standard sources of thesaurus and ontology construction.

Real-world data size

In our ontology and in this paper we have to do with real-world data sizes. Records describing single concepts of our ontology widely differ in length, but many of them reach or exceed the length of this paper. Unfortunately, we must therefore disappoint readers who expect substantial parts of the ontology to be presented here. With the given technical means, we must restrict ourselves to description as if we were reporting on any normal length movie, such as "Casablanca" or "Ladykillers". More of the ontology can be seen integrated in the scenario interface at our website (http://summit-bmt.fh-hannover.de/demo/run), or during a computerized presentation.

2. TASK-ORIENTED STRUCTURE OF THE ONTOLOGY

Our BMT ontology under construction must be about its domain. An idea of the concepts the ontology organizes is transmitted by its top (see figure 3) discussed below in section 4.2. What we record about an individual concept is determined by the tasks of the ontology and by its users. The ontology will serve summarizing and its subtasks involving human participants (users) and agent and non-agent system participants, demanding different types of information. The structure of concept records (see example in figure 2) reflects these requirements. The names of slots/fields reveal concerns of semantic organization and of gathering evidence from texts. Basic ontology records include the following main fields:

- concept
- synonyms
- hypernyms
- hyponyms
- sort
- description
- assertions
- occurrences

Our concepts are meaning items given by their use in a social or professional group, in our case physicians in Bone Marrow Transplantation. We call concepts by their name (the preferred term for them) and dedicate to every domain concept of interest quite an extensive concept record. The concepts of the ontology may have alternative names, called synonyms as usual. When checking whether a name is a synonym, we ask whether it can replace the preferred name of the concept in a given linguistic environment, without a noticeable change of meaning. This includes the identity of the word class. Hypernyms and hyponyms establish the generic hierarchy, the common structural backbone of ontologies as well as classification systems and thesauri.

ISSN: 2324-9773

The purpose of the field sort is to attach every concept to its class in SummIt top, the upper model of the ontology (see figure 4). While the generic hierarchy is still being constructed, all items are thus hooked to the top of the ontology. In the description field, we give basic

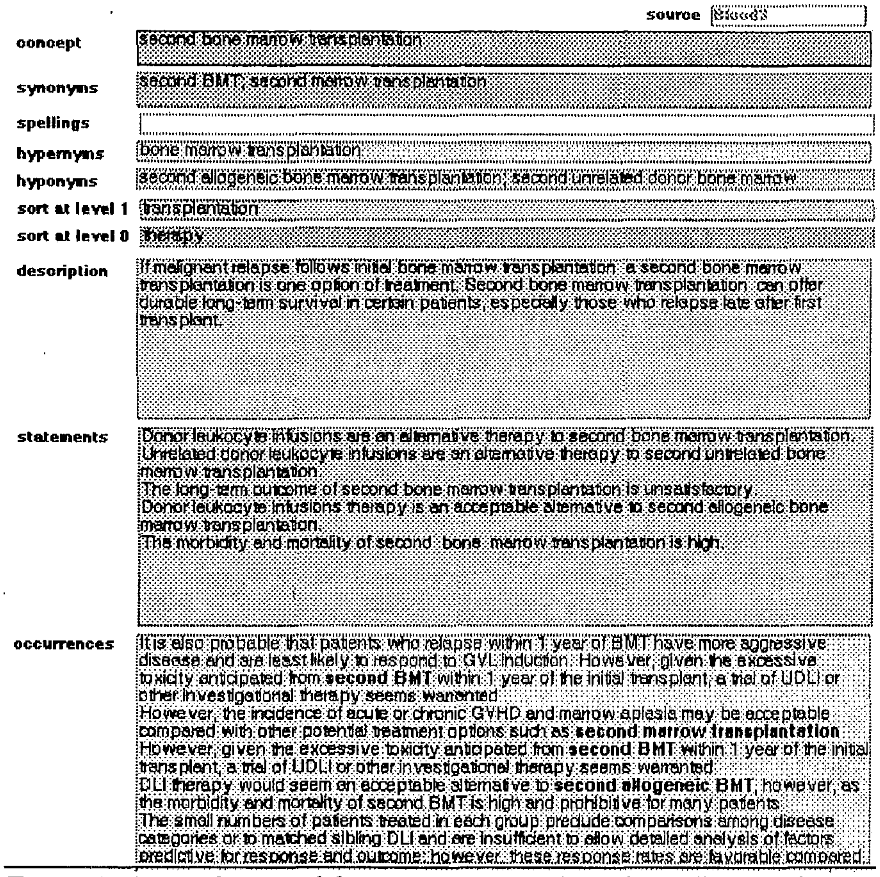

Figure 2. A sample record from a text-oriented ontology: "second bone marrow transplantation" in Blood3

information about the meaning of the concept, at least a lexicon-style definition. The description helps developers and users who are not familiar enough with the current concept to handle it. The occurrences of a concept in context are recorded in the occurrences field. Every entry there should include the concept with enough textual environment (context in the linguistic sense) that the utterance there can be interpreted, and that knowledge drawn from it can be stated, normally without returning to the source paper. In the assertions field we enter good knowledge statements from the concept's occurrences in a source we exploit. The assertions are made in a stand-alone manner. From them, we derive our formalized axioms (see formalized contexts below in section 5).

Depending on the tasks, different fields of an ontology record are used by different agents:
Query scenario formulation. During query scenario formulation, users refer to the ontology for concepts they can fill into the scenario form. They can browse the concept, synonyms, hypernyms, hyponyms and description of ontology records.

Query expansion. After being reformulated in predicate logic form, the query scenario is reworked and enriched with synonyms and other related terms from the ontology, and adapted to the search form of involved search engines (e.g. Medline).

ISSN: 2324-9773

Text passage retrieval. As soon as retrieved documents arrive, they are roughly checked for relevance by scanning them for passages that include the query terms. We process only documents with relevant passages.

Summarizing. Summarization agents need formalized assertions. The ontology supports them and their helpers in different ways. Perhaps the most interesting ones are the recognition of concept occurrences and knowledge processing during text interpretation and relevance assessment. In running text, concepts adopt various appearances. They show up under synonyms, they are expressed by paraphrases, they figure in specific phrasings, they are often wrapped in rhetorical decorations etc. In order to identify occurrences of a concept, we equip the ontology with equivalence information. It consists of rewriting rules specifying for instance that synonymous expressions are replaced by their preferred term (the ontology concept name), or that rhetorical embellishments are reduced. The summarization agents use the generic hierarchy of the ontology and the context expressions that state concept-related knowledge. The generic relation is used for example when a concept that occurs in a text does not directly match a concept of the query, but only its hypernym. Predicate logic expressions help to recognize whether the statement in the text conforms to the expectations set by the query, i.e. the agents check whether concepts are related as requested by the query before accepting the respective statement for a summary.

3. EMPIRICAL STRUCTURE OF THE ONTOLOGY

The empirical structure of our ontology explains how the ontology is derived from empirical evidence (see figure 3). It maintains the connection of ontology concepts to their sources in textual evidence. In our case, evidence mainly takes the form of concepts occurring in BMT texts, in different phrasings, synonyms, etc., and participating in statements of BMT knowledge. Our empirical procedure must discover these concepts with the propositions of knowledge to which they contribute, and organize them in the form of an ontology that can be used for finding the concepts and related propositions in other texts. This implies a switch from natural language to a formal representation of concepts and propositions. In the following we describe the empirical procedure and its results.

3.1 Ontology production procedure and resulting Empirical ontology structure

Our ontology construction follows an empirical procedure adapted from thesaurus building (Vickery, 1997; Endres-Niggemeyer, 1999). We apply the principles of qualitative field research, especially the framework for inductive theory development. According to the grounded theory framework proposed by (Glaser & Strauss, 1980), our ontology is a grounded theory of its domain ("one that is derived inductively from the study of the phenomenon it represents" – Strauss & Corbin, 1990). All concepts are justified by and connected with their evidence, found almost always in text.

Our ontology is built up according to the ideal 13 steps procedure (Endres-Niggemeyer, 1999) presented in table 1, with some adaptations to practical conditions. The first adjustment was stimulated by a domain expert. Instead of exploiting small numbers of papers from each subdomain, he proposed choosing two sorts of papers for the whole domain: BMT papers from Blood (2000), a core journal of the domain where current issues are discussed, and BMT educational papers from the Association of Hematology (2000) which present fundamental

ISSN: 2324-9773

topics in a tutorial style. The second major adjustment is due to the interdisciplinary composition of the group. The team members have backgrounds in linguistics, AI, computer science, medicine, biology and chemistry. Like this, a breadth-first approach was imposed by the mere fact that the qualifications for doing the whole job of concept creation and testing are distributed over several team members. With our breadth-first approach we now have arrived at stage 5 of the procedure in table 1, with the restriction that our first round of scenario testing has not yet given rise to a WWW search.

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

Table 1: The step-by-step procedure of empirical ontology construction

Figure 3 gives an overview of the empirical structure of the BMT ontology under construction. Partial ontologies with a light gray background are completed, planned ones have broken lines and clear backgrounds. All others (with normal lines and white backgrounds) are currently under construction.

First of all, we mine our concepts from papers of the domain. But thesaurus builders know from experience that concepts asked for in user queries may systematically differ from concepts offered in the pertinent literature. So we also exploit a corpus of user scenarios. When we glean concepts from user query scenarios, the empirical evidence is somewhat

ISSN: 2324-9773

weaker than elsewhere. Often, the physicians' scenario descriptions are very short, context information is poor or almost missing. To make user-defined scenarios fit for WWW

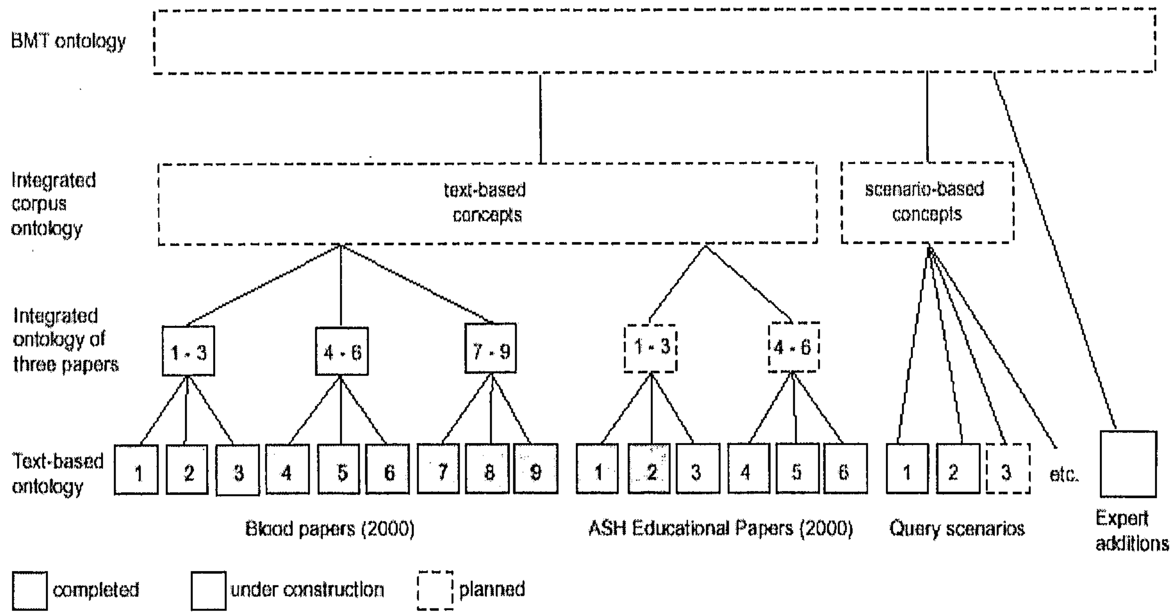

Figure 3: The empirical structure of the BMT ontology

retrieval, we have to modularize them and to fill them up with concepts from external sources. All concepts needed for stating usable scenarios are entered into the respective scenario ontology.

Besides core papers and user scenarios, our third knowledge source are concept records included by domain experts. Domain experts are needed for clearing up what is not or not directly said in the exploited papers, and they choose the acceptable views from the many standpoints offered by their colleagues.

From a technical point of view, ontology construction is supported by general-purpose tools: by Conc, a freeware concordancer, and by Filemaker applications. Source documents are PDF files. For WWW searching, we use the metasearch engine of Mac OS and other search engines.

However, some technical support does not change the fact that our ontology content is the result of intellectual work. As long as we are setting up the kernel of the ontology, we see no other chance. Later, small-scale updates can come from users. Automatic learning techniques may be applied.

First, we produce text-based ontologies, one for every exploited paper (or set of user scenario descriptions). The text-based ontologies record all concept occurrences in context. If these contexts convey valid knowledge, more concise statements are entered into the assertions field of the record. They are transformed into predicate logic expressions later on. As soon as we integrate text-based ontologies, the bulky occurrence data are left behind, but they can be

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BOŃE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

recovered from higher levels of ontology integration. For the sake of ease and safety, integration is done by bundles of three text-based ontologies at a time.

During integration, conceptualizations are checked and adapted if necessary, double records are removed, and statements found in the source databases are entered into the respective record in the integrated ontology. If necessary and possible, the concept descriptions are improved. A return to the source paper is sometimes necessary for checking and changing conceptualizations. As a result we have the modular overall ontology, constructed according to the common modularity concept of computer science ("divide et impera").

3.2 SummIt top – the upper model of the ontology

SummIt top, the upper model of the ontology (its uppermost level is shown in figure 3), has been conceived deductively by drawing from several sources. Adopting the MeSH classification proved to be awkward:

- The MeSH taxonomy is not strictly generic, which is a prerequisite for our ontology with its strict isa-relations for inheritance management.
- Parts of the MeSH taxonomy are oriented towards special scientific disciplines (e.g. genetics). In our task-oriented environment, however,
- Physicians ask for the consequences of a genetic aberration, and not how it is seen in genetic theories.
- The MeSH taxonomy is broad and its granularity does not suffice for our purposes.

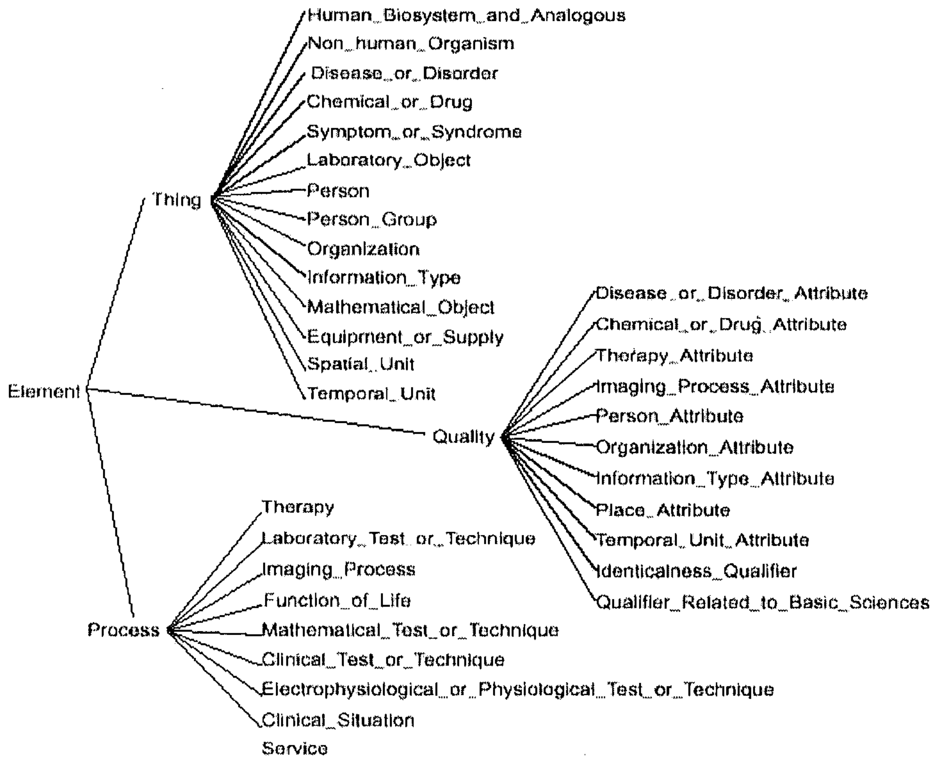

Figure 3: The ontology top in its current state

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

So we developed our own taxonomy. First we selected those MeSH classes, which fit the requests of the task. They were reformulated so that the expression of generic relations was facilitated. We enhanced the classes by assigning 455 concepts to them and inserting classes (e.g. clinical situation) if concepts could not be mapped to the existing taxonomy. If necessary, class designations and the internal structuring of classes were changed. The resulting classification was compared with other ontologies comprised in the metathesaurus of UMLS (Unified Medical Language System) and the Generalized Penman Upper Model (Bateman et al., 1995). Again, we reformulated class names and reorganized of the internal structure of classes. For instance, the internal structure of the class `Disease or Disorder' was focused on diseases occurring in the domain of BMT; further diseases were subsumed under the class `Other Disease'. The provisional top structure is useful because it sets an anchor to the generic hierarchies, but it is also often enough disputed by our empirical data. We adapt it inductively as needed.

3.3 Some bookkeeping

After illustrating the way the ontology is set up, we give a quantitative account of the concepts acquired so far. Table 2 shows the size of the ontologies derived from Blood (2000) papers. The first column gives the technical names of the Blood papers we exploited and the number of concept records in the resulting ontology data bases. From Blood1, 235 concept records were created, from Blood2 173, from Blood3 166, and so on. After integration, the combined ontology Blood1-3 contains 479 concept records. Of the 574 records of the source ontologies, 95 records have disappeared. In fact, new concept records have been also created, so that the turnaround was higher than displayed by the final figures.

Text-based Ontologies	Records	Integrated Ontologies	Records
Blood1	235		
Blood2	173		
Blood3	166	Blood1-3	479
Blood4	160		
Blood5	415		
Blood6	241	Blood4-6	781
Blood7	134		
Blood8	227		
Blood9	133		

Table 2: Overview of ontology records

The scenario ontology of the first survey turn contains 211 concepts resulting from 28 scenarios. We have recently started lexical equivalences, the same is true for formalized

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

contexts. Here stocks are still modest at the time of writing: Some 250 context expressions are ready for use and approximately the same number of predicates are equipped with equivalences.

4. FROM TEXTUAL CONTEXT TO FIRST-ORDER CONTEXT EXPRESSIONS

Context expressions (McCarthy & Buvac, 1993) are the main propositional knowledge representation format of the BMT ontology. Context expressions aim at formalizing in first-order logic what linguistic context achieves in natural language text: What is said there is understood in the light of the conceptual environment set up by earlier and – less frequently – later utterances in the same text. Predicate logic contexts are obviously useful structures for stating medical knowledge because they limit the scope of an assertion. Context expressions assert that the proposition p is true in the context c:

$$ist(c, p)$$

Both context and core propositions are predicate logic expressions. Following the example of Prolog, we accept only conjunctions and implications between context and core propositions. Disjunctions are expressed by alternative facts or rules. All predicates use ontology concepts as arguments. Context expressions must be true in the eyes of a domain expert.

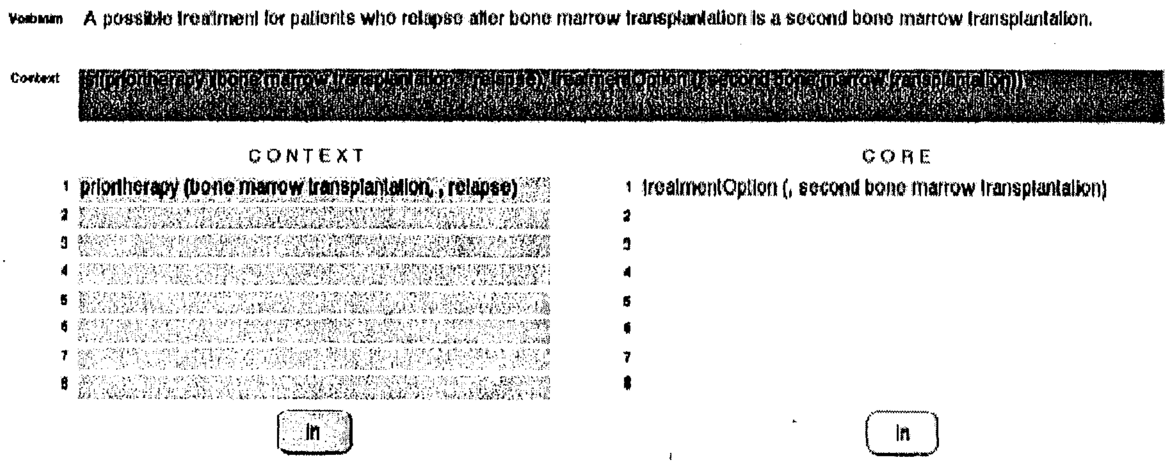

Figure 5: A simple context expression is generated from context propositions on the left and core propositions on the right hand

Figure 5 shows a context expression that is ready to enter the context database. It is accompanied by its formulation in natural language (top line). The context base accepts predicate expressions from the predicate base that are transferred either into the context area or into the core area. From the arriving context and core propositions, the context expression (second line) is generated. In figure 6 we show how first-order context expressions are built from normal first-order statements, including predicates with a list of arguments defined in

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

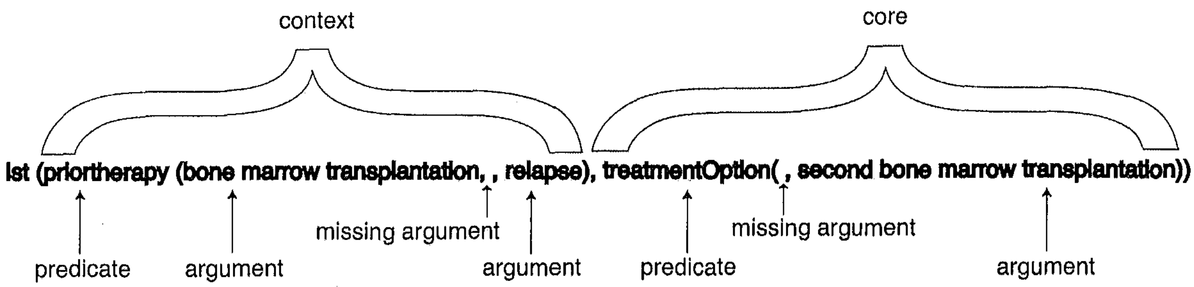

Figure 6: Formal structure of a first-order context expression

the papers of our corpus. We extract contexts, often sentences, but also longer sequences if necessary for stating a context in stand-alone manner. While there are often quite explicit statements of medical knowledge, this is by no means always the case. We present a counterexample. It demonstrates that we can mine more and more relevant knowledge from a text by skilled interpretation. In Blood3 (Porter et al., 2000) we find:

> DLI therapy would seem an acceptable alternative to second allogeneic BMT, however, as the morbidity and mortality of second BMT is high and prohibitive for many patients.
>
> However, the incidence of acute or chronic GVHD and marrow aplasia may be acceptable compared with other potential treatment options such as second marrow transplantation.
>
> However, given the excessive toxicity anticipated from second BMT within 1 year of the initial transplant, a trial of UDLI or other investigational therapy seems warranted.

Among many other things, we learn here that DLI therapy is an alternative to second allogeneic BMT. Doing a second bone marrow transplantation is presupposed three times. Readers can conclude this without bothering about the medical terms our example sentences are teeming with. The formulation in our ontology (cf. top line in figure 5) states the point as follows:

> A possible treatment of patients who relapse after bone marrow transplantation is a second bone marrow transplantation.

Although the context of our interest does not show up literally, it is a knowledge item whose validity is not confined to the paper at hand. It had to be derived by a competent human understander. She or he restates good assertions, elaborating implicit text know-ledge if needed. Textual formulations are stripped of rhetorical accessoires (example replacement rule: incidence of X -> X) and prepared for later formalization.

On their way to the context knowledge base they pass some normalization steps:

- Their concepts are standardized: only ontology concepts, no synonyms are accepted.
- They are reconstructed according to some syntax from the syntax base holding the syntaxes for predicate expressions. Thus they are normalized according to the number and kind of arguments.
- The resulting predicates are complemented with a natural language translation and

ISSN: 2324-9773

stored in the predicate base for later reuse.
- The predicates are entered into the context expression, either into the core or the context position.

The context expressions are not stored in individual concept records, because almost all of them are shared by several of them. Hence they are stored in and allocated from a central database. Since the information for recognizing concepts in running text aims at mapping textual occurrences to predicates, the textual equivalences are stored in the central pool together with the predicates.

5. TECHNICAL IMPLEMENTATION OF THE ONTOLOGY

As stated above, the ontology in the current state of development is stored in a set of Filemaker databases. First, the basic ontologies are developed and stored there. Second, the Filemaker application environment serves the engineering of syntaxes, predicates and context expressions. For operational summarization, however, a common and easy-to-access data structure seems appropriate. We decided to develop an XML structure according to the restrictions made by the authors of SHOE who demand semantic interoperability and therefore generality in the names of the XML tags. Thus, our tags are called <concept>, or <proposition> holding subtags like <sort> or <predicate> instead of using tags like <cancer> and implementing concepts in a manner like <cancer> leukemia </cancer> as dismissed by the SHOE team. We use plain XML with own DTDs for the subtasks of storing and retrieving basic ontology data like synonyms or hyponyms and for storing and retrieving predicate logic expressions, i.e. context expressions according to (McCarthy and Buvac, 1993). For managing XML data we use the XML database server dbXML (dbXML Group, 2001). dbXML allows XPath queries, which provides a simple possibility of searching all synonyms for a given concept, of investigating the set of hyponyms or of retrieving all predicate expression having the given concept as an argument. For integrating queries in an application, dbXML provides a Java API.

6. CONCLUSION

We have reported on an ontology for WWW summarization in Bone Marrow Transplantation. Due to the requirements of the task, the users, and the domain, it comes with some less common features. It is text-based and qualifies as a grounded ontology. In addition, it is user-centered. This counts in a real-world application. For stating medical knowledge, we use first-order logic context expressions. The ontology is multifunctional. It has to serve query scenario formulation, text passage retrieval and summarization proper in the summarization system. It will be stored and managed by an XML database server. Currently, the ontology is still growing, but its innovative features are set. A demo of its use in question scenario formulation can be seen at the project website http://summit-bmt.fh-hannover.de. We proceed on our way to application.

REFERENCES

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

Aussenac-Gilles, N., Biebow, B. & Szulman, S. (2000). Corpus analysis for conceptual

ISSN: 2324-9773

modelling. EKAW 2000 [On-line]. Available: http://www.irit.fr/ACTIVITES/EQfiSMI/GRACQ/WSEKAW2000/accepted.html

Bateman, J. A., Henschel, R. & Rinaldi, F. (1995). Generalized Penman Upper Model [On-line]. Available: http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html

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. Available: http://delicias.dia.fi.upm.es/miembros/ASUN/kaw98.ps.zip.

dbXML Group (2001). The dbXML Group L.L.C. [On-line]. Available: http://www.dbxmlgroup.com/

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

Endres-Niggemeyer, B. (1999). Empirical Methods for Ontology Engineering in Bone Marrow Transplantation. International Work- shop on Ontological Engineering on the Global Information Infrastructure, Dagstuhl Castle, May 25, 1999 [On-line]. Available: http://www.ik.fh-hannover.de/ik/person/ben/OntoBone.pdf

Erman, L., Hayes-Roth, F., Lesser V., & Reddy, R. D. (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty," Computing Surveys, 12(2), 213-253. (also in R. Engelmore & T. Morgan (Eds.), Blackboard Systems, (31-86). Addison-Wesley).

Fox, J. & Das, S. K. (2000). Safe and Sound: Artificial Intelligence in Hazardous Applications. Cambridge MA: MIT Press.

Fum D., Guida G. & Tasso, C. Evaluating Importance: A Step towards Text Summarization. IJCAI 85 – Proceedings of the Ninth International Joint Conference on Artificial Intelligence (pp. 841-844) Los Angeles, CA, August 18-23 1985.

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

Golebiowska, J., Dieng-Kuntz, R., Corby, O. & Mousseau, D. (2001). Building and Exploiting Ontologies for an Automobile Project Memory, IJCAI 01 Workshop on Ontologies for Information Sharing [On-line]. Available: http://www.tzi.de/buster/IJCAIwp/Finals/golebiowska.pdf

Henflin, J., Hendler, J. & Luke, S. (1999). SHOE: A Knowledge Representation Language for Internet Applications. Technical Report CS-TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park [On-line]. Available: http://www.cs.umd.edu/projects/plus/SHOE

McCarthy, J. & Buvafic, S. (1993). Notes on Formalizing Context. In R. Bajcsy (Ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJAI-93) (pp. 555-560). San Francisco, California: Morgan Kaufmann. Available: http://www-formal. stanford.edu/jmc/context3/context3.html

Norman, D.A. & Draper, S.W. (Eds.) (1986). User centered System Design. London: Erlbaum.

Porter, D. L., Collins Jr, R. H., Hardy, C., Kernan, N. A., Drobyski, W. R., Giralt, S., Flowers, M. E. D., Casper, J., Leahey, A. , Parker, P. , Mick, R., Bate-Boyle, B., King, R. & Antin, J. H. (2000). Treatment of Relapsed Leukemia after Unrelated Donor Marrow Transplantation with Unrelated Donor Leukocyte Infusions. Blood: Journal of the American Society of Hematology, 95(4), 1214-1221.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne & M. Scriven, (Eds.), Perspectives of Curriculum evaluation (39-83). Chicago: Rand McNally.

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

Strauss. A. & Corbin, J. (1990). Basics of Qualitative Research. Grounded Theories Procedures and Techniques. Newbury Park California: Sage.

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.

W3 Consortium (W3 C) (2000). Extensible Markup Language (XML) 1.0 (Second Edition) [On-line]. Available: http://www.w3.org/TR/2000/REC-xml-20001006

ISSN: 2324-9773

Endres-Niggemeyer, B. (2001). CONSTRUCTING AN ONTOLOGY FORWWWSUMMARIZATION IN BONE MARROW TRANSPLANTATION (BMT). 12th ASIS SIG/CR Classification Research Workshop, 13-30. doi:10.7152/acro.v12i1.13791

ISSN: 2324-9773