The MIHMA Demonstrator Application: bmt line

C. Kindermann* T. Hoppe* D. Marwinski* H. Oertel*
O.K. Paulus* E. Buu† S. Heimann† A. Schmiedel†
R. Tolksdorf* P. Volle†

†Non Standard Logics S.A.
57-59, rue Lhomond
F–75005, Paris

*Technische Universität Berlin
Fachbereich 13, Informatik
FR 6-10, Franklinstr. 28/29
D–10587 Berlin

April 1996

1 Introduction

1.1 Accessing Medical Information On-Line

Three years ago, Beverly Hanscom was diagnosed with chronic myeloid leukemia. After two cycles of chemotherapies her doctors now have recommended a bone marrow transplant (BMT) as the next therapy step. Ms. Hanscom has been lucky: a donor with matching bone marrow has been found through international cooperation of donor registries. The marrow harvest and the transplant are now being prepared to take place within the next weeks.

Early after the initial diagnosis, Ms. Hanscom started to consult a series of background sources to understand the consequences of her having leukemia and the treatment options that existed. Recently she began to use on-line media for this purpose. She is now consulting the World Wide Web for on-line literature three times a week, reading as much as she can about the forthcoming treatment, ranging from basic descriptions to advanced material on newest research results. Especially important for her are experience reports of former BMT patients and on-line discussion groups where she often finds valuable advice and simple cheering-up.

Her using of on-line information was encouraged by the medical team responsible for her treatment. Several of the team members themselves use the Web in addition to printed magazines and conferences for keeping their knowledge fresh and discussing clinical results with colleagues from all over the world using dedicated electronic mailinglists.

While the case of Beverly Hanscom is fictitious and seems to be oversimplified, it nevertheless draws a realistic picture of how access to information will change in a society that naturally exploits the information superhighways – not the version that administrations, politicians, and media industry CEOs have forecast for the next ten years, but today's de-facto Infobahn, the Internet-based World Wide Web.

1.2 The Internet and the World Wide Web

The Internet is the world's largest computer network, or to be more precise, a global network of networks.[1] The Internet allows for the transfer of computer files from a local computer to a remote computer, or vice versa, independent of the computers' brands and the operating systems running on them. It enables users to exchange electronic mail (e-mail). It allows for worldwide discussion of almost every subject in electronic bulletin boards

[1] In July 1995, The Economist estimated in its "Survey of the Internet" (WWW: `http://www.economist.com/internet.htm`) 20 million users at roughly 5 million Internet computers (hosts); in January 1995, already 9 million Internet hosts were determined in the Internet Domain Survey of Network Wizards (WWW: `http://www.nw.com/`).

CGI: Common Gateway Interface to connect a Web server to external application programs or databases

HTML: Hypertext Markup Language used to describe the structure of pages provided by a Web server

HTTP: Hypertext Transfer Protocol used for the communication and data exchange between a Web browser and a Web server

URL: Uniform Resource Locator, the address with which a document or data file can be located on the Internet

Web Browser: Program to display at the user side the contents of the Web

Web Server: Program that provides content to the Web

Table 1: Important Protocols and Interfaces of the World Wide Web Platform

called newsgroups (or mailinglists, if interaction occurs via e-mail). Users can have live on-line conversations typing into their computer (Internet Relay Chat) or through simple video conferencing tools like CU-SeeMe. Plus many other things, and more still to come.

Connecting to the Internet requires standard hardware and software only: a PC or Mac, a modem, and a telephone line to dial up a local Internet provider. The software is available for all important computer platforms, and to a large extend it is free or low-cost.

The "killer application" of the Internet – the one that triggered the Internet's current commercialization and the broad public interest in the Internet – is the World Wide Web. The Web (WWW, W3) is the user-friendly, multimedia interface that incorporates most of the existing Internet functionality in an easier-to-use-manner. The World Wide Web comprises a mixture of tools and methods to integrate heterogeneous multimedia information (hypertext, graphics, video, sound) available on the Internet into a distributed hypermedia information base. It relies on a common addressing scheme for information entities, a set of common communication protocols, and means for the negotiation of data formats (see Table 1).

The term World Wide Web is also used to refer to the overall information space which is spanned by the many types of protocols and services, and the documents and data accessible through them. For the scenario sketched above, the Web provides a bulk of information of different types (as is summarized in Table 2), which indeed serves as an important information source for both medical professionals and patients.[2]

[2]See, e.g., questions and discussions in the mailinglists `bmt-talk` for patients and `BMT-DR` for physicians.

- Periodicals: BMT Newsletter, Mailing lists `bmt-talk` and `BMT-DR`
- BMT specialist literature for physicians and patients
- Books: BMT (patients); Leukemia, Atlas of hematological pictures (physicians)
- Databases: MEDLINE, Cell Line Database, ...
- Descriptions of BMT and leukemia research projects
- Patient experience reports
- Requests and guidelines for bone marrow donation
- Documents about bone marrow/leukemia related medical topics: diseases (breast cancer, AIDS), drugs affecting bone marrow, and genetical aspects of leukemia
- Contact addresses of physicians, researchers, institutions
- References to books, journals, CD-ROMs
- Newspaper articles
- Job announcements
- Other BMT-related issues, e.g., insurance problems

Table 2: BMT and Leukemia Related Material on the Internet

1.3 Finding Information on the Web

Despite the general availability of information on the Internet, today Ms. Hanscom or the experts from the medical team preparing her BMT would find it difficult to locate a particular piece of information. To locate information requires users to visit many different servers, and to browse their respective contents, which is time consuming, confusing, and inefficient. To locate useful information in the Internet's information space by navigation only is so laborious that it severely limits the Internet's usefulness as a source of information.

The limitations of navigation led to the development of value added services which help users to satisfy their information needs more efficiently. Such services have been developed for a variety of server and protocol types, creating systems of varying complexity. The effort required of both the provider and the user of these systems ranges from compiling and consulting lists of interesting sites to maintaining complex distributed applications with powerful search interfaces.

The fundamental value provided by these systems is the reduction of the search space, i.e., to guide users to those parts of the information space where information relevant to their information needs can be found. In simple cases, e.g., the various index-like systems, the search space is reduced to a list of information sources whose contents might satisfy the user's information need as expressed in her query. Due to the large number of information sources available, this is usually realized by maintaining only references to the original on-line information.

Systems of this kind are useful, but they assume that users already have a certain knowledge of a given domain and are able to express their information need adequately in the query language. The user also has to be able to evaluate and relate different pieces of information based on their contents. Finally, because of the sheer volume of indexed information and the resulting time needed to maintain the entire reference basis, references to documents often point to documents not existing any longer or moved to another location.

To facilitate the exploration of a domain, some information services organize the information sources to which they refer into hierarchical structures. The most straightforward approach is to classify all information sources into some directory structure, where each directory is named with a label indicating the contents of the information sources it holds. Such a directory structure can be regarded as a very basic model of the domain covered by the information service; the model makes explicit the relations that hold between different pieces of information. More complex models can be built, and the complexity of the model determines the added value the service provides.

In the remainder of this report, the development of bmt line, a Web-based medical information service is described that directly addresses the scenario presented above. bmt line provides a reference database for Web and Internet resources in the domain of bone marrow transplant and leukemia. The added value of bmt line is based on the provider controlled content of the information it references, the high degree of semantic integration of referenced material which is based on a complex semantic model, the support of different standard user classes, and the support of individual customization of the service through the application of single users' profiles.

Section 2 describes bmt line from a user perspective, the goals that led to its development, its scope, and its different areas with the associated functionalities. *Section 3* briefly describes the MIHMA project in which bmt line has been developed, as well as the technical platform used for service development and operation. The following two sections concentrate on the necessary steps to develop and operate the service. In *Section 4* the development of the semantic model underlying bmt line, and the design of the user interface are discussed. *Section 5* describes the entire cycle of activities required to operate bmt line. *Section 6* summarizes, and *Appendix A* evaluates the expenditure of adding new document references to bmt line.

2 bmt line – A Medical Web Information Service

2.1 Goals for bmt line

The goal for the development of bmt line was to create a Web site that would serve as the main reference point to relevant Internet-based material on the topic of bone marrow transplant and related issues. Rather than

providing new and original material, bmt line would scan the Internet for all kinds of information resources dealing with BMT, and would organize its findings in an intelligent way so that users could find the needed information more easily and with greater precision than by using indexing sites such as Lycos, Alta Vista, or InfoSeek. A user of bmt line should get the impression that after having consulted the service she knows of all or the most relevant information sources for the given topic.

In bmt line, an intelligent organization of information goes beyond flat listings of material found on the Web. Instead, bmt line

- organizes document references around a hierarchy of supported topics;
- provides short descriptions of referenced documents that allow users to assess the value of a document prior to downloading it;
- provides cross-references between related topics thus allowing users to explore the domain of interest by exploiting semantic relationships between domain concepts;
- provides information on authors of papers, people being responsible for Web sites, researchers, physicians, and organizations active in the area, etc., and links this information together through a number of explicitly established relationships;
- offers different views on its information base for different user groups, namely patients and their relatives, bone marrow donors, general practitioners, and domain experts;
- allows customized information presentation for subscribed users based on dynamic generation of HTML pages.

Information organization in bmt line exemplifies what we've called *semantic integration* of information: through explicitly established relationships applied in a systematic and consistent manner, isolated chunks of information are connected to a more meaningful whole. The result is an *integrative information service* which accesses data from existing services with heterogeneous structure and differing but potentially overlapping content; this data is integrated into a new and value added information offer targeted to meet the information needs of users with different user profiles. Fig. 1 shows the basic concept of an integrative information service, and the actions necessary to operate the service and to gain semantic integration (see Sec. 5).

2.2 Scope of bmt line

The scope of bmt line is bone marrow transplant and its particular role as a treatment for different forms of leukemia and other kinds of cancer of the blood forming organs. In Dorland's Illustrated Medical Dictionary,

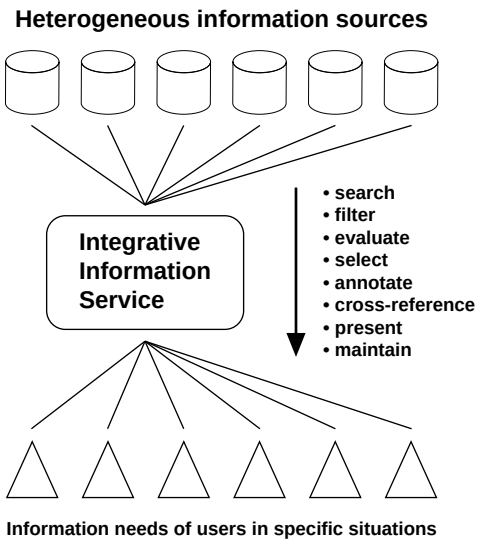

Figure 1: Model of an Integrative Information Service

Leukemia is described as a "progressive, malignant disease of the blood-forming organs, characterized by distorted proliferation and development of leukocytes and their precursors in the blood and bone marrow". Damage to the bone marrow can also be caused by radiation therapy or chemotherapy used to fight different forms of cancer. In both cases, bone marrow transplantation from donors to patients is used to restore the blood-forming capability of the patient's own marrow.

In the first version of bmt line, other issues which would also have fit into this context have been developed less extensively or have been left out completely, such as immunological aspects of blood typing.

2.3 Areas of bmt line

bmt line is structured into a set of service sections or areas with associated page types. The service sections are grouped into three main classes: *entry pages* which serve as the major access points to the service, *support pages* which constitute the "service area" of bmt line, and *content pages* which supply users with information about the covered domain.

The entry pages correspond to the traditional "homepage" of a Web site. In bmt line, the entry area, however, consists of multiple pages: bmt line offers "guest" homepages for each of the four supported user classes plus a homepage for subscribed users ("members"). A central "entry page" leads

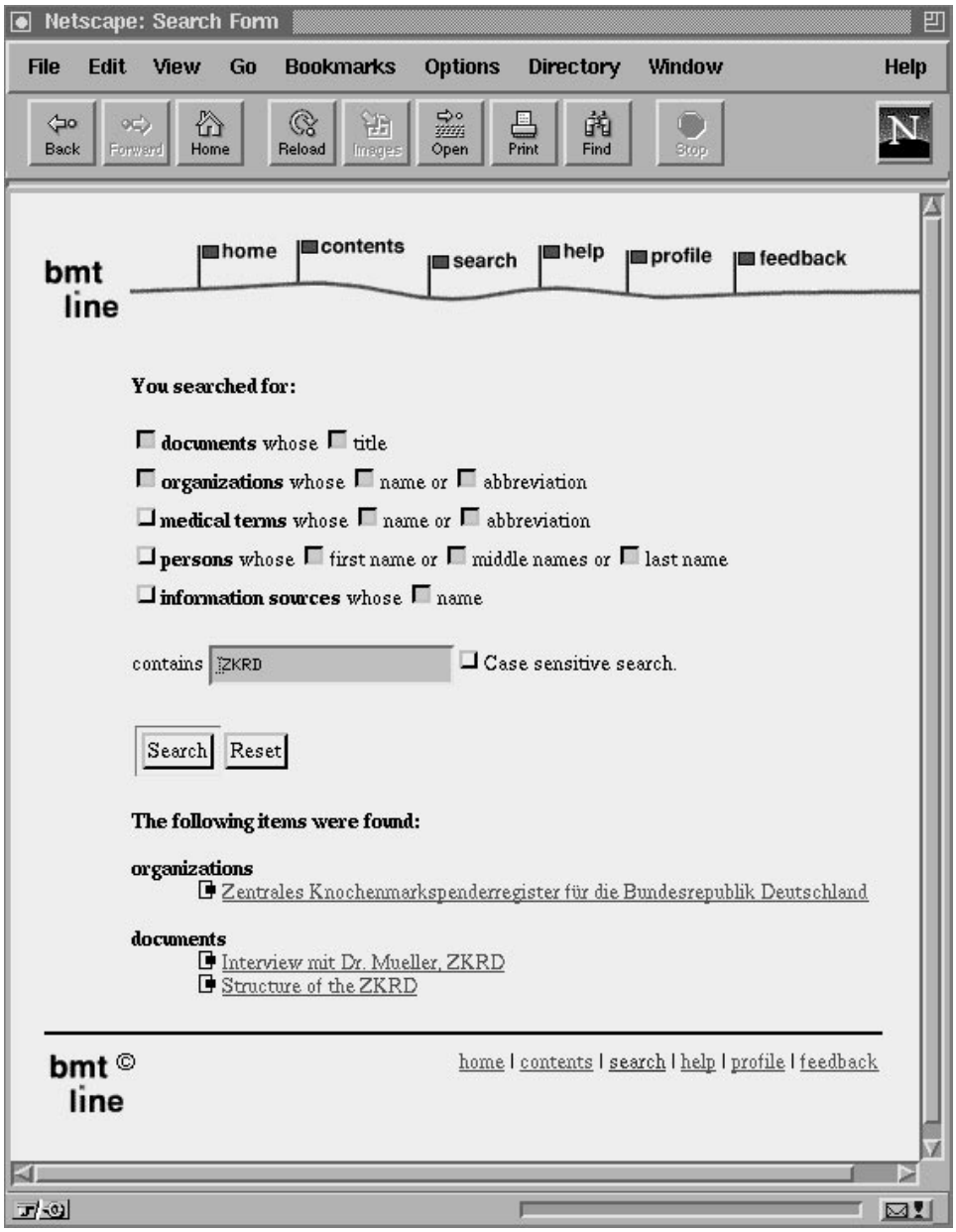

Figure 2: Search Form and Result. The upper part shows the search form with a query for all documents having ZKRD in the title, and all organizations having ZKRD in their name; the lower part shows the structured result list for this query with three matching entries found.

new users to these homepages, lets them subscribe to the service, and offers a kind of sneak access to selected pages of the service.[3]

[3]For a full access to the service it is required that users have explicitly selected one of the possible user classes (which may be changed anytime during a visit at bmt line) – hence the different homepages.

The support part of bmt line comprises a set of pages that provide users with helpful information about the service and assists them in getting efficient access to the real content:

- Overview pages such as a comprehensive catalog and the list of all supported topics help access desired areas of the service, and at the same time make the scope of the service explicit.
- A News Track groups references to information that is more time dependent than other document references within the service. It lists relevant newsgroups and mailinglists, offers direct access to them or their archives, and provides a mechanism that facilitates subscription to mailinglists. The News Track also lists all references recently added to bmt line.
- A Feedback Area asks for active contribution of users. It allows for general entries into a guestbook where users can leave their general impression about the service. Furthermore, users can submit problems encountered during service usage and references to documents missing in the service.
- The different navigation paths through the service are supplemented by a search form that supports query-based access to the service content. Users may chose among different categories for which they want to perform the search, e.g., topics, documents, persons, etc. Selected categories may be combined within a single search; in this case, the result set is structured along these categories (see Fig. 2).
- For users accessing bmt line as anonymous guests, the service offers a subscription form for free "membership" in bmt line. Members of the service can inspect and alter their personal profile at any time, thereby adjusting several service parameters to their personal needs (user profiles are further discussed in Sec. 2.5).

This is rounded off by standard pages such as a help page or an imprint.

The content part of bmt line – consisting of topic pages, the glossary, catalog pages, and internal documents – is described in the following subsection.

2.4 Topic-Oriented Content Organization

In bmt line, information is centered around medical topics. Topics are organized in a hierarchy that reflects the conceptual hierarchy of the medical terms. Each topic is explained through a glossary entry, and thereby linked to other, related topics. For each topic, bmt line knows about relevant information (documents, pictures, general information sources, etc.). The set of topics supported by the service – shown to users on a content overview page – defines the scope of the service in a natural way.

Topic Pages

Because of topic's prevalent role in bmt line, *topic pages*, pages that each correspond to a single medical topic, are made the central information units of the content area. They are linked to other topic pages, to glossary definitions, and to relevant information resources or their catalog pages, respectively.

Figure 3 shows the screenshot of the upper part of a bmt line topic page. The typical topic page starts with the topic's name plus a list of alternative names and spellings. A glossary definition explains the topic and links to a glossary extract in which further relevant terms are explained (see below). The textual explanation is accompanied by a set of relationships to other topics, in this case to more general topics covered by the service.

The general introduction to the topic is followed by a listing of references to documents and information sources covering the page's topic.[4] The association of such information resources to topics is based on an explicit annotation of resources by the service provider, and not on the application of simple keyword matching or other, more advanced Information Retrieval techniques. Each reference contains the name of the information resource, and a short description of its content. It provides links to the resource itself and to the catalog entry within bmt line (see below).

The listing of references is grouped into several subsections. For each topic, a *recommended first reading* is offered which points to a document that gives a quick and competent introduction to a topic. This is followed by different content or document type oriented subsections such as experience reports from other patients, documents on possible treatments, pictures, or research reports. The listing is concluded by a set of pointers to general information sites containing related material.

The content of topic pages depends on the user class a user has selected when entering bmt line. Pages differ in the set of information resource references provided for the particular topic, and on the subsectioning used. For example, on the topic page of the Graft Versus Host Disease (GVHD) a patient will get a different recommended first reading than a general practitioner, an expert, or a potential donor. If they like, users can easily change the selected user class near the bottom of each topic page.

Glossary

The bmt line glossary provides short explanations of all relevant domain terms, and also offers an overview of the service scope. The glossary entries have been extracted and adapted from the Meta-Thesaurus of the UMLS (Unified Medical Language System) [13]. Glossary entries also link to a corresponding bmt line topic page if it exists.

[4] An *extended view* of a topic can be selected which lists also documents covering topics more general and more special than the current one.

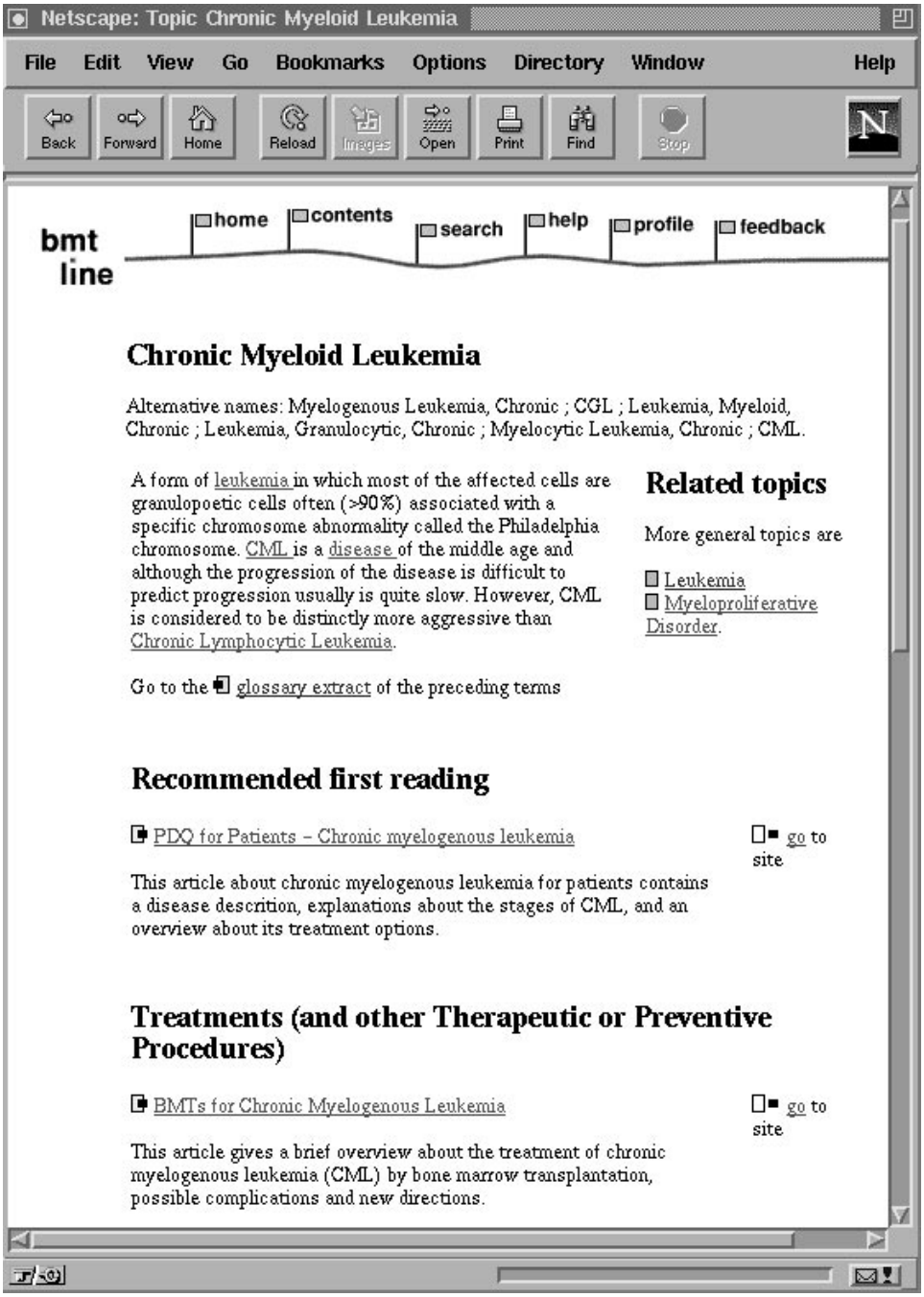

Figure 3: Sample bmt line Topic Page

In addition to the comprehensive glossary, bmt line provides *glossary extracts* which just explain those terms occurring in a small text such as the textual description on a topic page, or a full document provided originally by bmt line.

Catalog Pages

For all references to documents and other information resources important within the context of the service, bmt line provides a *catalog page*. A catalog page contains a hyperlink to the referenced resource, and offers enough information that allows users to assess the value of a document before downloading it: A short description, content format, content language, document size, the URL of the primary Web site and, if existent, URLs of mirror versions, and the like. Maybe even more important, bmt line provides links to more information within bmt line, by listing, e.g., the topics covered, the institution that provides the document, or its authors. Fig. 4 shows a sample document catalog page.

Catalog pages are also used to represent other entities of interest, e.g., medical institutions, domain experts, etc. Each entity's catalog page again provides semantic links to other information within bmt line. For example, the catalog page of a person shows the person's affiliation, its domain of expertise, papers authored, or other information known to the service.

The usage of explicit catalog pages for referenced information is common to a number of well-established reference services such as McKinley's Magellan Site [7]. What's special to bmt line is the way in which cross-references are provided among different pieces of information and the high degree of semantic integration resulting from it.

Internal Documents

Topic pages, catalog pages, and the glossary are supplemented by *internal documents*. These are either documents originally provided by bmt line, or documents that originally were published by other institutions. In the latter case, value is added to these external documents by applying *semantic markup* to it, i.e., by linking important terms within the document to the bmt line glossary, for example. This semantic markup integrates an external document deeper with bmt line's content; such internalizing of documents is used, e.g., for documents proposed to users as a recommended first reading.

2.5 User Profiles

Traditionally, the Internet and the World Wide Web are based on a gift culture: content and functionality are given away for free. The most successful business model for the Web at the moment therefore relies on financing through advertising. The Web allows for much better tracking of access to

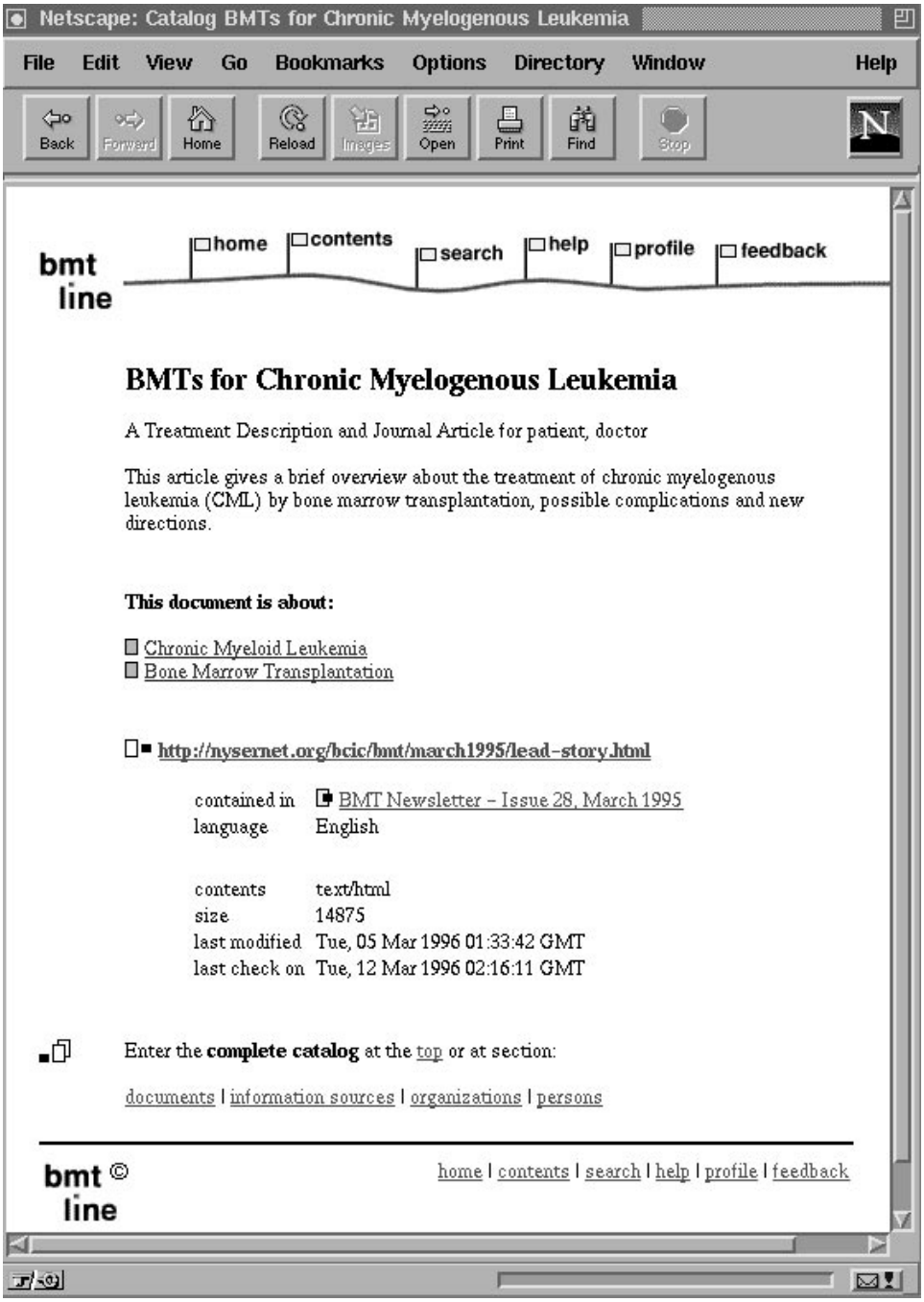

Figure 4: Sample bmt line Catalog Page

information pages than any other traditional marketing medium. Ultimate precision is only gained, however, if a service provider can trace back access of pages to individual users: In this case, advertisers and sponsors of certain Web pages or service areas can be provided with detailed demographic information about Web users accessing their pages or parts of a service worth to place ads on.

For this reason, service providers often require that users subscribe to the service; each time a user starts a session with a service, she has to provide her user identification and password – or similar information – to get access to the service's pages. To make this authentification process acceptable for users, a Web site has to provide some added value for its users, something that makes it worth providing the service with personal information and memorizing the personal ID and password (cf. [12]).

For bmt line it was anticipated that a business model for running the service would rely on advertising and sponsoring. bmt line therefore asks frequent users to subscribe and become "members" of the service. Subscription is free and supplements an anonymous "guest" access mode. bmt line's guest access allows for access to all information provided by the service. Subscription to the service is motivated by a number of additional features which, in summary, enable individual customization of the service.

In the first version of the bmt line, the most important feature with respect to service customization is the ability to select an arbitrary subset of the topics supported by the service as personal topics of interest. Immediate access to these topics is offered on an individual user's homepage; in the News Track, new material added for any of these topics is presented more prominently. Further parameters that subscribed users can preset are the user class they belong to, and general service parameters as, e.g., the time interval to be applied when determining what should go into the What's New section.

Personal topics of interest and individual settings for different service parameters are stored in a user's individual *user profile*. The user profile is initially created when a user subscribes to bmt line. Users may inspect their profiles at any time when interacting with the service, and may edit the profile's content. Information contained in the profile is kept private between service provider and subscribed user. It is not disclosed to any other parties; this includes also the names and (e-mail) addresses of users.

2.6 Extensions of bmt line

Above we have described the functionality of bmt line as it was implemented in the first version of the service until April 1996. During the design of bmt line a number of extensions were discussed which couldn't be realized in the first version but nevertheless should be presented here briefly.

The potential added value for subscribed users was not fully exploited in the first version of the service. For example, it would have been easy to implement a mechanism that every two weeks sent e-mail notifications to service members and informed them about new references concerning their personal topics of interest added to bmt line.

Initially it was planned that users could adjust the service's behavior through a larger number of service parameters, e.g, the interaction language of the service, or the preferred language of sought documents. More of these parameters could be introduced over time. Again, subscribed users could define values for these parameters persistently in their user profiles.

In the version of bmt line described above, the predominant method for finding information in the service is navigation. The search interface provided by bmt line is straightforward but not very elaborated. Since all references to documents, information sites, and other entities are stored in the service's central data repository (object base), an enhanced search interface could be devised which would allow for incremental construction of complex queries, exploiting the underlying structure of the object base. In an early phase of the design of bmt line, it was considered to build such an interface based on the query-by-reformulation approach described in [8]. The idea was abandoned at that time because the required communication between Web browsers at the user side and the bmt line Web server would have resulted in interaction too slow to make this search interface useful. The advent of Java [5] and the small "applets" which could run be run by standard Web browsers, puts the development of such an enhanced interface again back on the agenda.

Finally, bmt line misses to exploit one of the most attractive features of the World Wide Web: its interactivity. The entire process of finding new references and entering them into the service relies on the provider of the service. Instead, means could have been provided that allowed users to propose new documents together with a partial annotation, or to describe themselves,[5] the organizations they work for, or the projects they are assigned to. Still all information could be passed to an editorial team at bmt line to ensure consistent quality within the service. Putting a contributor's name in a "This resource was annotated by..." note on catalog pages would encourage him and other users to send in more data, and make them feel that bmt lineis *their* site. It makes them come back (see [12]). And the service would profit because the time to enter the required data would decrease.

[5] In its first version, bmt line already allows members to make them and their homepage and e-mail address publicly known within the service.

3 The MIHMA Project

3.1 Project Background and Goals

The development of bmt line has been funded by DeTeBerkom GmbH, a 100% subsidiary of Deutsche Telekom AG as part of the research and development project MIHMA. The project consisted of two research teams at the Technical University of Berlin (Project 2100 MIHMA) and at the Paris-based software company Non Standard Logics (Project 2101 MIHMA-P).[6]

The acronym MIHMA stands for Model-based Integration of Heterogeneous Multimedia Archives. The main focus within the MIHMA project was the development of enabling technology to create, run and administer certain types of information services to support information discovery and retrieval on the Internet. To achieve this, a set of generic methods and tools have been researched, and integrated into a prototype of the development platform. An overview of the MIHMA project is [9].

In this context, bmt line plays the role of a prototype application which is used to demonstrate the appropriateness of the MIHMA approach and the usefulness of the software platform developed within the project.

3.2 MIHMA Architecture

The software platform developed in the MIHMA project consists of three main components. There is a run time environment that provides the base for the running on-line service, and an administration environment providing tools to maintain and extend the data collection. Both components make use of the third component, an object-centered knowledge representation system. A detailed description of the MIHMA technology is [14].

Run Time Platform

The run-time platform consists in the combination of an HTTP server handling requests received from WWW browsers, and a MIHMA server responsible for managing the contents of the information system (knowledge base, documents) and delivering HTML pages to be sent to the user. The MIHMA server holds an instance of the BACK system (see below) through which it accesses the object base. It also holds application code which mainly consists of rendering scripts producing HTML text. Figure 5 shows the overall architecture of the run-time platform.

Two variants of this platform exist: one uses the CGI mechanism and can be connected with any HTTP server, the other is based on the Apache server

[6] Unless stated otherwise, in this report we use the name MIHMA to refer to both projects at once.

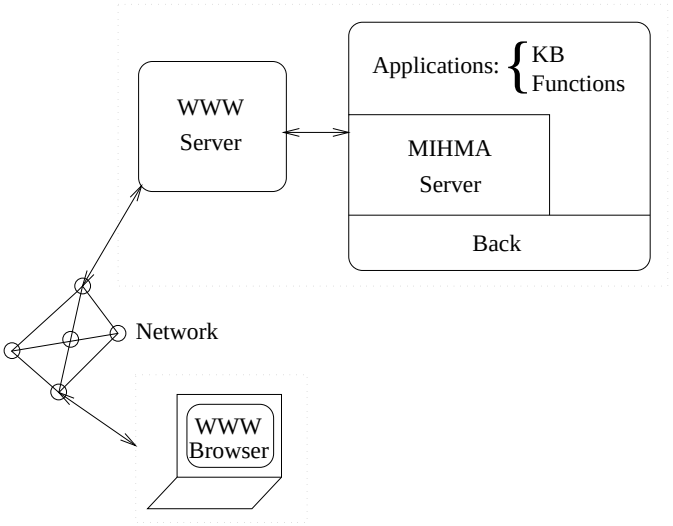

Figure 5: Overall Architecture of the Run-Time Platform

enhanced with a specific extension module interacting with the MIHMA server. The latter solution provides higher performance.

Administrator Workbench

The Administrator Workbench (AWB) is a set of tools to automate some of the administrative processes. In the current prototype system, it supports the location and retrieval of new documents, their analysis and annotation, and the maintenance of document references stored in the object base. The tools are implemented as processes that perpetually run in the background and are triggered by external events. The AWB also provides a simple point-and-click interface for the domain expert to support his annotation task. The administrator controls the results of the AWB processes through the object base editor.

Object Base

All information displayed on Web pages of bmt line is internally stored in a central data repository, the *object base* of the service. In the MIHMA project, the object base is implemented using the BACK knowledge representation system [10, 11].

The BACK system is based on a class of data or representation models called description logics [15]. These models combine the object-centered approach found in object-oriented database management systems with the declarativity of relational databases.

BACK implements a uniform, semantically well-founded, compositional language based on term-forming operators that can be used to build descrip-

tions. Well-formed descriptions according to this language are used for all interactions with the knowledge base (KB): defining new terminology, describing individuals, and queries. Like other DL-based systems, BACK performs a number of actions to organize a KB and to keep them consistent (automatic classification of descriptions, completion of partial descriptions, consistency checking).

Earlier prototypes of BACK have been developed at the Technical University of Berlin [4]. The current version developed by NSL [3] is implemented as a C++ class library, which includes methods for reading and writing KBs from and to files in editable source format. The C++ application programming interface can be used directly by applications.

4 The Making of bmt line

Bringing bmt line into existence required the development of the MIHMA platform described above, and the provision of two further components essential for the service: a semantic model for coding all information in the central object base, and the design of the user interface.

4.1 The Semantic Model of bmt line

In bmt line, a semantic model is used for coding all relevant information within the central data repository. This model provides a thesaurus of the chosen topics, which in turn is used as a basis for navigation through the information space. Information referenced by the service is classified into the semantic model. Furthermore, the model is used to represent and adapt the view on the information space according to the needs of a specific user of the service.

The semantic model is also used to capture necessary data for the search and maintenance components of the AWB, and for the management of subscribed users of bmt line. The administrator directly interacts with the model to annotate new resources and refine descriptions of existing ones.

The semantic model is a combination of a generic model and a domain-specific model usable in the domain of bone marrow transplantation. The generic model provides a core for building services like bmt line but is totally independent of a specific application domain. The generic model contains several submodels focusing on documents, information sources, organizations, persons and other entities relevant for a MIHMA-based integrative information service. A generic topic submodel is part of this model as well.

Reusability has been a major design goal during the development of the generic model. As can be seen in Fig. 6, the ontology of the model has a very modular structure. As was observed in the construction of the domain-specific model used in bmt line this modularity makes extendibility an easy

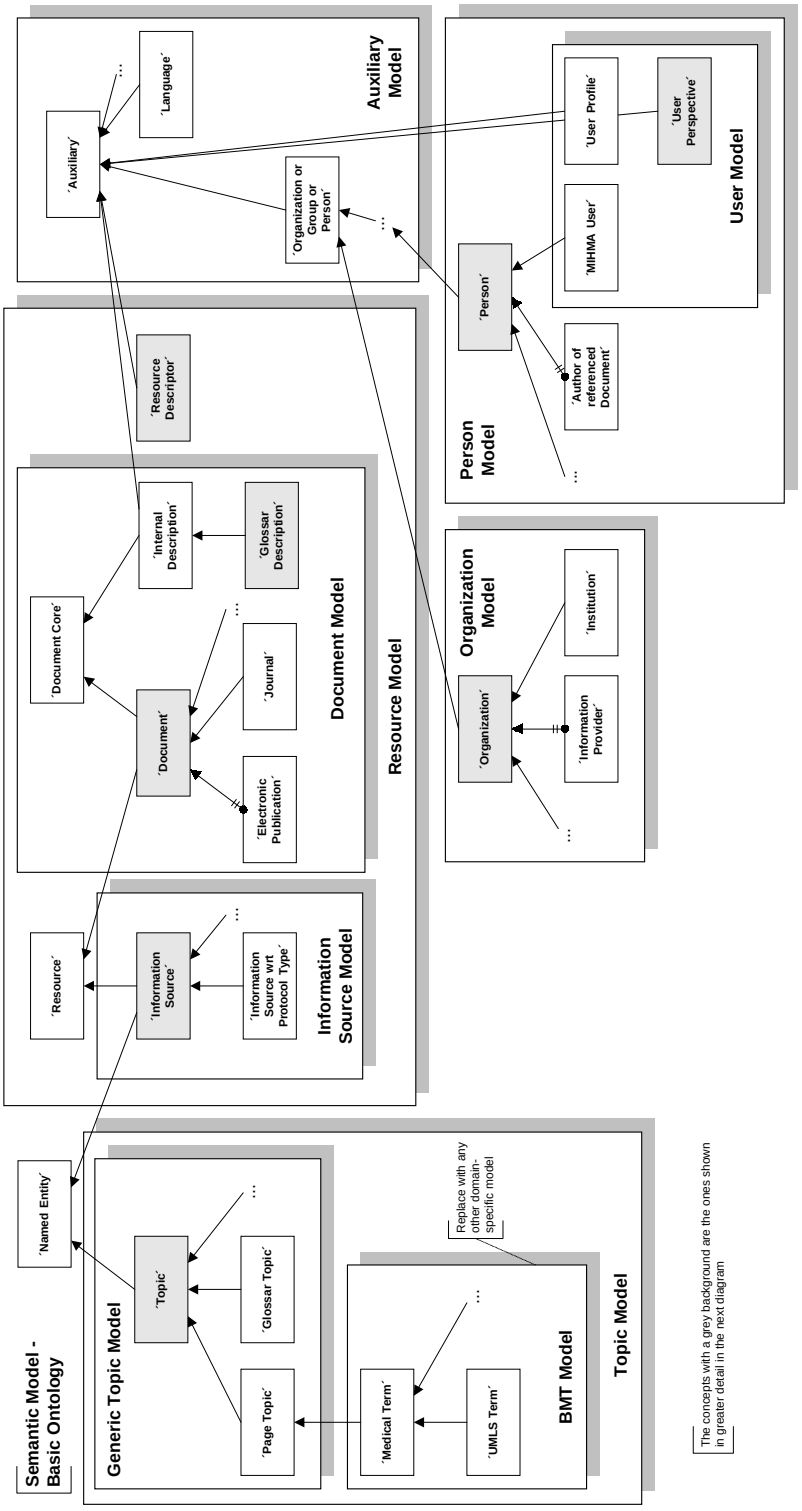

Figure 6: Basic Ontology

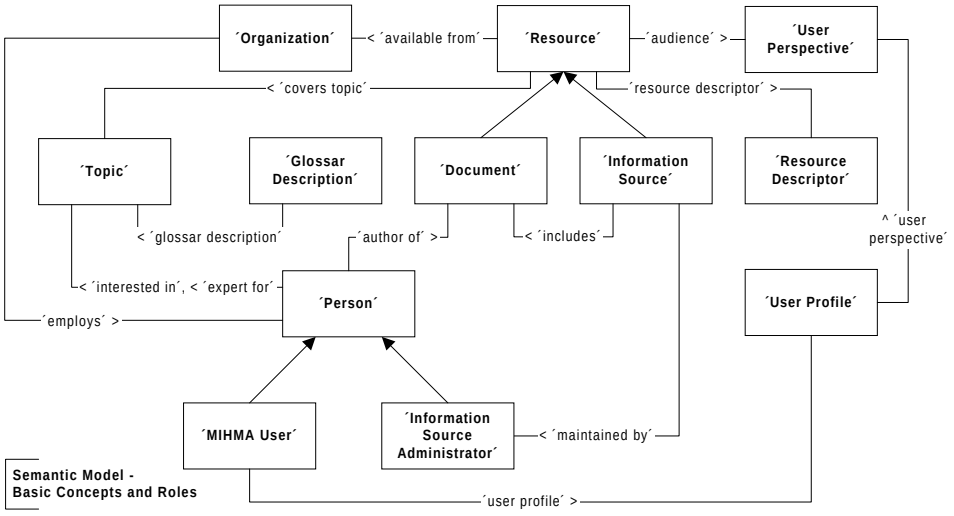

Figure 7: Basic Concepts and Roles

and straightforward process; during this process, the basic concepts and roles as shown in Fig. 7 did not have to be modified at all.

The domain-specific model developed for bmt line extends the submodels of the generic model to be suitable in a medical domain. Especially the document and organization submodels had to be augmented with additional constructs to make it possible to adequately describe document types and organization forms found in a medical domain. The most important part of the domain-specific model is the medical topic model. Figure 8 shows a substantial part of the topic hierarchy as presented to a user of bmt line.

The semantic model was developed using the Unified methodology for object-oriented analysis and design [1]. It is coded using the BACK knowledge representation system described above. Our experiences in using the BACK system for the development of an integrative information service are documented in [11].

As will be discussed in the next section, the topic structure has been extracted from the standard thesaurus and refined by a medical expert.

4.2 Information Extraction

The application domain specific part of the semantic model comprises three important categories of information: topic terms which are are used for the annotation of resources with medical terms, a hierarchical structure for the topic pages of bmt line, and glossary definitions for medical terms appearing in bmt line.

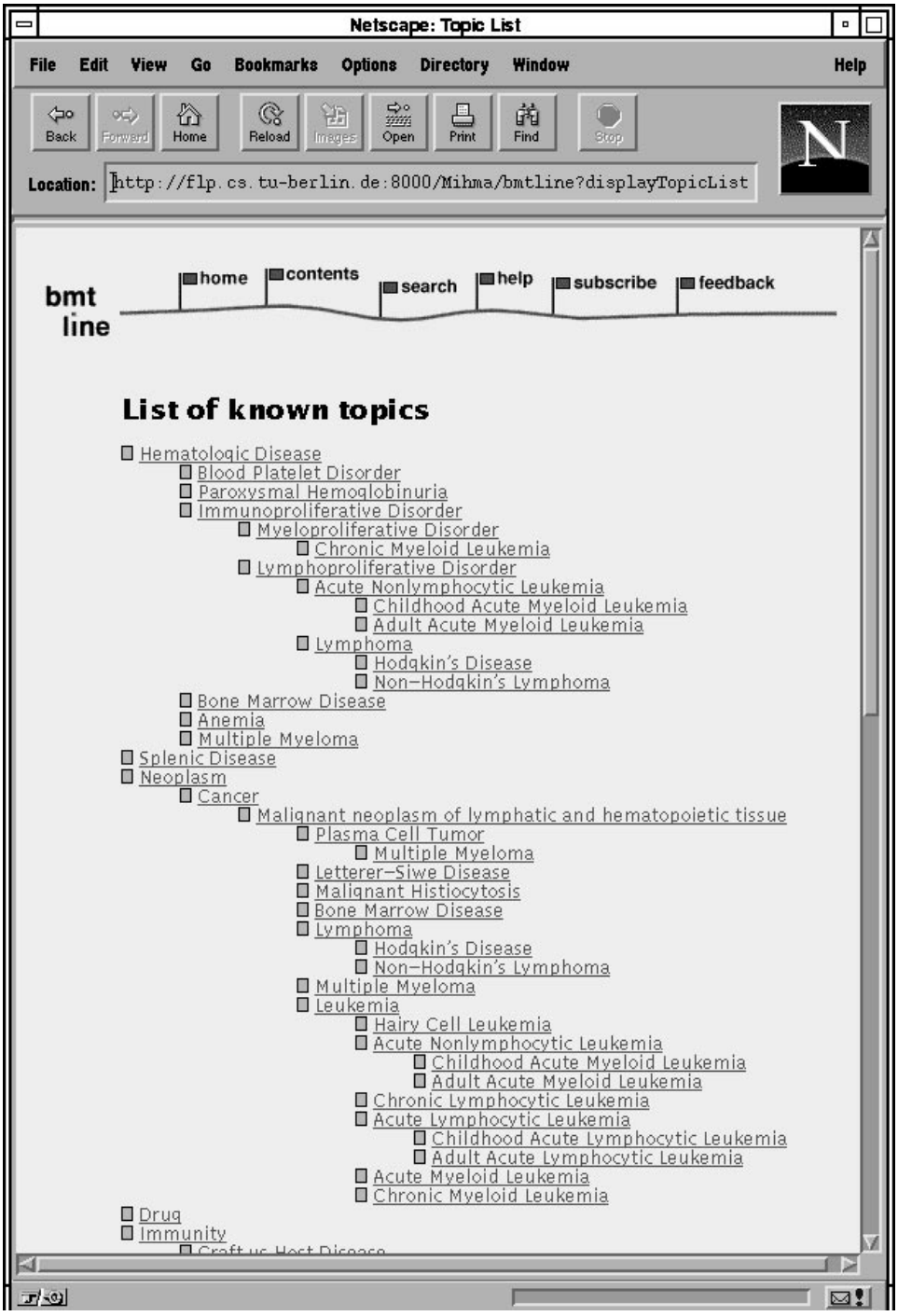

Figure 8: Topic Hierarchy

While for certain domains it might be necessary to handcraft the semantic model, in medical applications one can extract information from existing terminologies and thesauri. Through the Unified Medical Language System (UMLS) project of the U.S. National Library of Medicine the situation is even better: the most important medical terminologies have been integrated into a single thesaurus, the UMLS Meta-Thesaurus [13].

For bmt line, the UMLS Meta-Thesaurus was used to identify and extract the medical topics for the domain specific parts of the semantic model. This activity was done in close cooperation of a medical expert in an incremental manner: a set of core topics was presented to the medical expert, who extended the set, criticized the extracted topics, and proposed changes of the scope of the application motivated by his background knowledge. During the four iterations needed to agree on a stable medical model, a set of basic questions arose:

- Should medical experts and laymen see the same topic structure?
- Even though the hierarchy extracted from the UMLS is correct in medical terms, is the complete hierarchy actually needed?
- Should all topic terms appear in the glossary, and should all glossary entries become topics with an associate topic page?

Eventually, it was decided that experts and laymen will see the same basic topic structure reduced to a set of topics the medical expert marked as relevant, and that the glossary entries are a superset of the set of topics with associated topic pages.

In a final extraction cycle, the topic concepts marked by the medical expert were extracted. The hierarchical structure of the extracted terms was manually adjusted by the medical expert, since the initial structure contained in the UMLS was no longer adequate for the reduced set of topics. A small set of additional topics were introduced to complete the model. Additionally, textual definitions were extracted for all concepts reachable from the core topics through semantic relationships modeled in the UMLS. These definitions where approved by the medical expert, who wrote also a handful of new definitions which were missing in the UMLS. These definitions form the glossary entries of bmt line.

The experience we gained from this process can be summarized as follows:

- Extraction of concepts from an existing thesaurus or knowledge source is actually more a knowledge elicitation task than simple extraction of the information.
- A domain expert is essential in this process in order to direct term extraction from a domain viewpoint, and to restructure and adapt the resulting terminology until it fits the actual application's needs.

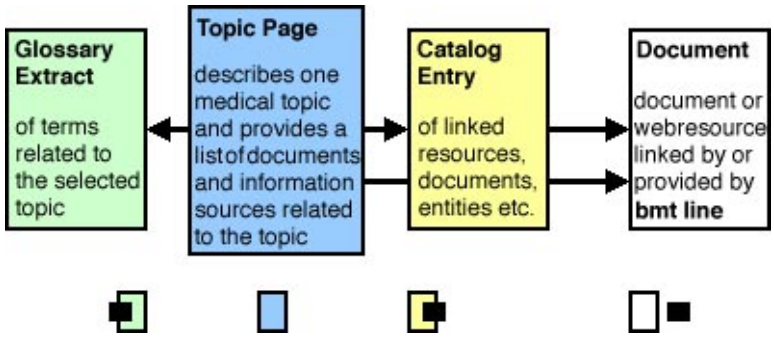

Figure 9: Overview Map for bmt line Content Area

4.3 Designing the User Interface

In the early days of the World Wide Web, the mere presence of a company or institution on the Web was worth lot of attention. It didn't matter that most Web pages literally looked gray and as if put together by the organization's computer department. Things have changed since then. Today, if a Web site should get a fair chance to be accepted by its targeted user group, a well designed appearance and user interface are indispensable.

With so many well designed services available on the Web, users will simply not accept poor presentations but will consider a service with a weak design low-quality, independent of its actual content or the superior technology underlying it. Users have a very strong, though not articulate, understanding of what looks good and what doesn't. The look and feel of a service influences users' attitude towards it, and thereby their willingness to learn how to put the service to good use.

As important as the visual appearance of a service is its structure, i.e., the way the content of individual pages is organized and how pages are linked to each other. Examples of the first point are the maximum length of lists, how to present graph structures, how much information to put on a page, how to separate parts of a page, and so on. The second point is about user's orientation within the service. One must decide what type of links should be on all pages, how they should be labeled, how many links should be on a page, how to use color to indicate one's position, etc.

For bmt line, quite some effort has been spent on these questions. In a very early prototype of the service, a nearly one-to-one mapping of the structure of the domain model to the hypertext structure was used. The result was informative, but not adequate for a real on-line service. A methodological separation of the domain model and the hypertext was needed, i.e., both had to be developed rather independently of each other. The rendering layer of the MIHMA platform is then responsible for the mapping between the two. The hypertext structure and the visual presentation of the service

have finally been developed with the aid of a professional designer, which greatly improved the service's structure and appearance.

The first and most important task of the designer was to develop the different areas of bmt line and the corresponding page types. Simple optical means are applied to distinguish the different areas and page types, including:

- different background colors for the different areas (light gray for entry area and support area, white for content area)
- different layout grids for different page types
- different coloring in the page headers for different page types

In order to keep loading of bmt line pages fast, all images used in the service are kept small in byte size by using only few colors and keeping images small.

Navigation support is given by implicitly distinguishing *structural links* and *semantic links*. Structural links connect the different areas of the service and are mainly provided in the global button bar in the header and footer of every page (see, e.g., Fig. 4). Semantical links connect information units within the content area, e.g., a topic and the corresponding glossary definition, a topic with related topics, or a topic and the relevant documents and information sources. Semantical links are used as textual links within a page body.

Semantical links are also marked up with small icons that provide a simple way of indicating which type of page a semantic link leads to. These "link labels" and their association to page types are shown in Fig. 9. Their appearance is oriented at the arrangement of content area pages shown in the figure and provided to users on the help page of bmt line. Starting with the central role of topic pages, a small black "bit" indicates the direction within the content area: back/left to some background explanations (glossary), forward/right to a more detailed description of an entity (catalog page), or out of the service to the references document or site.

Care was taken to ensure that all navigational support also works when a user connects to bmt line using a textual browser like Lynx, or has switched off image loading in her graphics-oriented browser. In this case, link labels are replaced by textual alternatives, and the text-based navigation bar in the page footer provides the same options as its graphical counterpart in the page head.

Further details of the user interface design are discussed in [6].

5 Operating bmt line

5.1 The Operation Cycle

Once the basic ingredients (semantic model, general layout structure, graphics design, etc.) are in place, operation of the service can begin. Operation

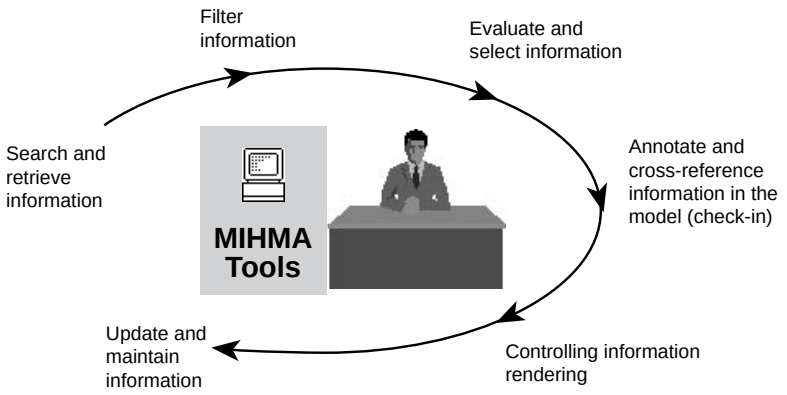

Figure 10: Activities for Operating bmt line

of bmt line requires a series of actions which are performed in an continuous cycle (Fig. 10):

- Searching for and retrieving of new information resources to be included into the service
- Filtering information to exclude obviously irrelevant resources
- Evaluating and selecting information resources
- Adding the selected resources into the object base by annotating and cross-referencing them
- Controlling the rendering of the object base content
- Updating and maintaining the object base

Each of these tasks is described in more detail in the following subsections.

A critical point of the bmt line operation cycle is the question to which degree this process can be automated. Fortunately, the tools developed in the MIHMA project can handle a number of the steps listed above, including the search, retrieval, and filtering of new information resources, parts of the modeling and annotation process, the generation of the bmt line Web pages from the data repository, and the maintenance of references in the data repository.

Figure 10 suggests that a single person is involved in operating bmt line. This, of course, is an abstraction of reality. Usually, there will be different roles for the respective tasks, which will be instantiated by experts from the corresponding fields. For bmt line, we identified three tasks and roles: the editor, the domain expert, and the administrator.

The editor is responsible for the overall, visible quality of the service, its appearance, and the way the information offer is structured. The administrator

is responsible for the technical part of the service, for programming, keeping the system running, and knowing the internals of the various components.

The domain expert is responsible for the quality of the information sources referenced by the service, and for the quality of the domain model. The expert needs not necessarily to know how the system works, or how the model and the reference descriptions are actually produced. He doesn't even have to know how to locate information over the network. His main task is to evaluate information sources wrt. their relevance and quality, to describe them correctly, and to ensure the correctness of the underlying domain model. It is impossible to create a value added information service of any standing without the help of an expert of the application domain. The expert acts as a human filter to separate the noise from the valuable information.

Running a service like bmt line as a commercial service would require further staff. For example, a web assistant would take care of subscribed users, would personally respond via e-mail to user entries in the Feedback Area of the service, and the like. She may also assist the Administrator in the annotation process, or the Editor in copy editing documents originally provided by bmt line.

5.2 Searching, Retrieving, and Filtering Information

The discovery of new documents to be included into bmt line is supported by an automatic search process within the AWB. The goal is to automatically locate documents relevant to the supported domain in order to release the administrator from the task of searching the Internet for new documents.

The search process gets hints to documents and locations to check for documents by a number of feeds. The contents of relevant Usenet newsgroups and mailinglists is scanned for Internet references that are passed to the search component. Queries are periodically issued to large index-services on the Internet, whose results are also passed to the search process. The administrator can manually submit hints to the search process which are processed with higher priority. Finally, users of the service can send hints using the Feedback section of bmt line.

For each reference, the search component retrieves the document, and asks an analysis component for a relevance rating. If the document is considered relevant, it is passed to the annotation component together with an annotation suggestion. In addition, references found in the relevant document are taken as hints to further relevant documents.

5.3 Evaluating and Selecting Information

The search component locates and retrieves documents from the Web, applies a keyword-based filter to preselect potentially relevant documents, and

forwards the resulting set of documents to the domain expert. The domain expert confirms or rejects the relevance judgment after evaluating each document's content, and produces an annotation in terms of the domain model. This annotation is then associated with the reference to the original document. The annotated reference is stored in the object base, and the local copy of the document is deleted.

To support the expert's annotation task, a HTML form has been developed which provides a simple point-and-click interface. Thereby, the expert does not need to know the formal modeling language and can concentrate on the document content.

Although the expert has to actually read most of the documents retrieved by the automated search process, this part of the process was quite effective. Obviously, a domain expert is able to very quickly assess the relevance of a document and determine the topics covered by it.

The expert's annotations cover only the medical aspects of a document. Additional information which is more or less independent of the particular domain, like the document size, or its creation and expiry date, are extracted from the communication protocols by the search component, or have to added by the administrator.

5.4 Annotating and Cross-Referencing Information

Each resource selected by the expert has to be described as an object that will be introduced into the central data repository (object base) underlying the service. When the administrator starts to describe a resource he already receives partial descriptions that result from the expert's medical annotations and the data determined by the search component.

Basically, the final introduction of a resource's description into the object base consists of the following steps:

- Verification of the domain expert's annotation
- Inspection of the resource to be described via a Web browser
- Identification and manual search for related information
- Identification of objects in the data repository already representing this related information
- Introduction of new objects into the data repository if such object do not already exist
- Establishment of relations in terms of the semantic model between those objects (cross-referencing)

For each new *document* to be entered into the object base, the administrator needs to identify the information source which includes the document, the

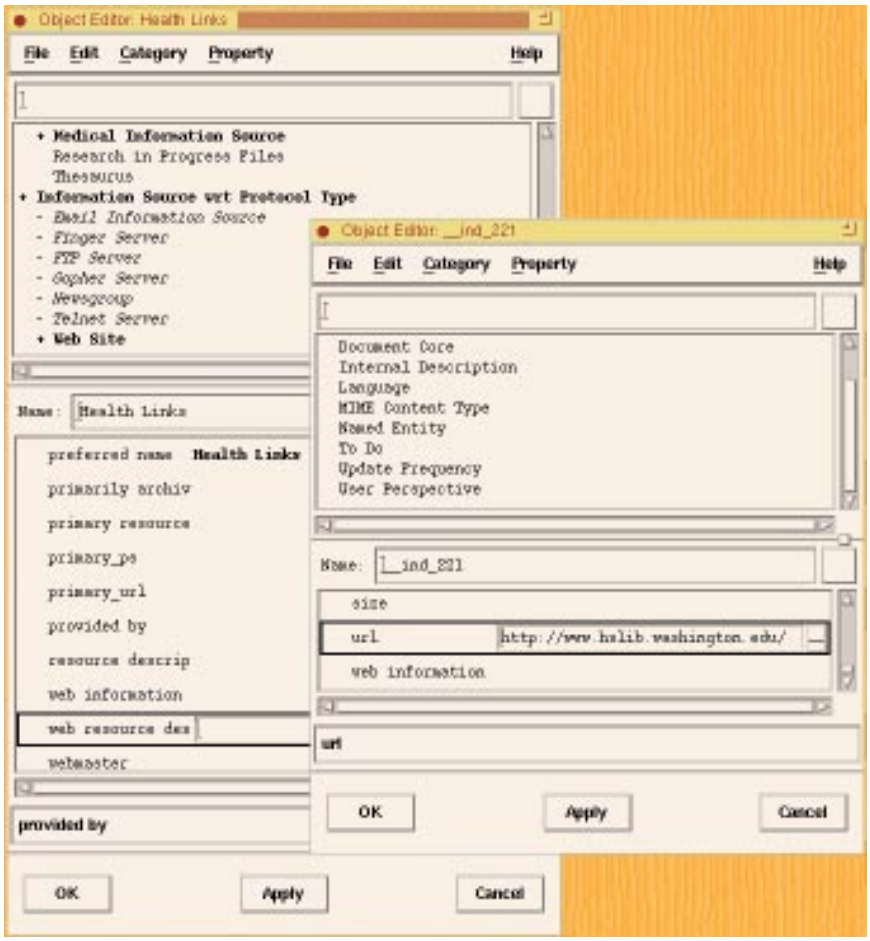

Figure 11: Object Base Editor of the AWB

organization which publishes the document, authors which have written the document, addresses of these organizations and persons, and so on. For each new *information site*, the administrator has to identify organizations which provide the information source or make it available, persons which administer it, and relationships to other general information sources, e.g., whether one site mirrors the other.

The process of annotating objects is supported by the object base editor of the AWB. Based on basic categories selected by the administrator, this tool presents further semantic relationships that can or must be filled. For example, if a document is described the tool will ask for its author(s), and when a general information source is annotated it will propose to enter the responsible administrator. Fig. 11 shows the two windows of the object base editor. In the left window, a general information site is being edited, for which, in the right window, its resource descriptor (URL plus additional information) is modeled.

During the process of adding new resources to the object base it is unavoidable that the semantic model needs to be adjusted, modified, or augmented. One consequence is that the descriptions of previously annotated objects have to be reconsidered and possibly have to modified too. A quality assurance process is required to ensure consistency of the descriptions after such modifications.

Also contributing to the overall quality of the service's information base is a coherent description style. In order to achieve this, editing guidelines are inevitable, especially if several persons should enter object descriptions. Such guidelines could also define an effective editing strategy.

The annotation and cross-referencing process can quickly become the bottleneck of the overall process. While the evaluation of a document can be handled by the expert quite efficiently, the time to determine the additional information can become a time consuming task. Especially in the early phases of the construction of the information service, each document may trigger the description of a wealth of additional entities, like a document's authors, the original information source, the affiliated organization, and so forth. Since all of this information has to be determined manually this task consumes most of the time required for the annotation task. Fortunately, the administrative overhead decreases over time, and can be expected to be neglectable in the long run. A statistical evaluation of the time spend for annotating documents for bmt line is presented in Appendix A.

5.5 Controlling Information Rendering

In the MIHMA model, an application contains a rendering layer of software that is responsible for accessing the object base, retrieving the data and composing HTML pages that make sense to a certain user in a particular situation. It is this set of functions – together with the contents of the object

base – that actually determines which documents will be available by the server.

Whenever a user fetches information from the bmt line server, top-level functions are called in response to such requests and produce whole pages. The top-level functions directly correspond to the different page types. Auxiliary functions are called by top-level functions to compose the content of the page. Some are specialized to produce layout elements (like header and footer), whereas others produce content sections (such as sorted lists of document descriptions).

Standard rendering methods are applied to all objects in the object base, depending on their description in terms of the semantic model. For example, when a new book object is introduced into bmt line it is ensured that a standard rendering is applied which will differ from the rendering of, e.g., an object representing an image. Sometimes it is desirable, however, to customize the rendering of a particular object. For example, the editor asks the administrator to present a particular book of exceptional value in a very specific way, e.g., by listing a number of selected chapters on the book's bmt line catalog page. The administrator then will devise a specialized rendering script for this individual case.

The separation of object base content and rendering is quite convenient when entering new data into the object base because the administrator can concentrate on the content, neglecting appearance in the service. Especially in the case of customized rendering for individual objects, however, careful examination of the appearance of the affected object is required. Unfortunately, the MIHMA platform as used for the construction of bmt line didn't support such testing very well. Immediate feedback on how newly introduced objects would appear on the different pages would have made quality assurance a much easier task.

5.6 Updating and Maintaining the Object Base

The Internet is a dynamic medium, where documents are added, deleted, or moved to new locations, where the content of documents changes, and where links between documents are established and removed. Servers in the Internet may be temporarily off-line, or restrict the access to documents. Internet connections are sometimes slow causing timeouts. For a reference service like bmt line, the critical issue is that most of these changes are outside the control of bmt line's provider, and that no central instance exists that informs bmt line about these changes.

In order to keep the content of bmt line up-to-date, the validity of URLs referenced by the service therefore needs to be checked on a regular basis. This process is usually called 'link verification'. Furthermore, it must be ensured that document annotations maintained in the object base still correspond to the contents of the referenced documents in the cases where documents have been modified by their owners.

In order to perform URL maintenance a tool was developed, called "maintainer", which ensures that certain problems are detected and corrected:

- Temporarily unaccessible URLs are marked as unaccessible for proper rendering.
- Unaccessible URLs are marked for manual removal by the administrator.
- The URLs of moved documents are corrected automatically in the object base.
- Resources whose content has changed are marked for manual inspection and correction.
- Redundant URLs are detected and result in a notification of the administrator.

Basically the maintainer is a wrapper for the publically available link verifier MOMspider [2], which we extended to cope – besides the Web's HTTP protocol – with FTP and Gopher URLs. It's the maintainer's responsibility to determine URLs from the object base which need to be verified, to schedule the next maintenance date for those URLs, to trigger the link verifier on those URLs, to analyze the result of the link verification, and to invoke appropriate follow-up actions.

The experiences gained with the maintainer running on a regular basis can be summarized as:

- Timeouts during the link verification were the most frequent problem. In most cases the maintainer dealt with this situation automatically through further link verification attempts with increased timeout values.
- Frequently, the link verifier detected changed documents. On the one hand this is due to frequently updated documents, on the other hand this is due to the applied simple approach of counting and comparing word frequencies of text documents in order to identify documents with changed contents. Here, further research for more advanced content extraction and comparison methods would be necessary.
- Only rarely the link verifier detected unaccessible or moved resources.
- While it is desirable to use a fully automated and autonomous maintenance procedure, the current state of the used Internet protocols requires manual intervention by the administrator from time to time. For example, on two occasions the link verifier encountered problems which led to incorrect behavior and could not be resolved by the maintainer; these problems were due to servers which seem to comply only partially to the HTTP protocol.

6 Summary and Outlook

On the Internet and the World Wide Web, a wealth of information is available for almost every subject one could think of. On the other hand, finding a particular piece of information suited to satisfy a personal information need is often time consuming, confusing, and inefficient.

Using a scenario from medicine, a patient looking for special information on bone marrow transplant (BMT), we described this problem in detail. With bmt line, we presented an integrative information service that helps users to find information about leukemia and more efficiently. We presented the service in detail, describing its different areas and functionality. bmt line provides a reference database for Web and Internet resources in the mentioned domain. The added value of bmt line is based on:

- the provider controlled content of the information it references
- the high degree of semantic integration of referenced material which is based on a complex semantic model
- the support of different standard user classes
- the support of individual customization of the service through the application of single users' profiles

The development of bmt line was carried out in close cooperation with medical experts from ZKRD, the German Central Registry for Bone Marrow Donors in Ulm, and was based on enabling technology and a methodology developed in the MIHMA project. The different steps necessary for the making and operation of bmt line were discussed in detail.

The future of bmt line depends on a financing of its continuous operation. Different options are currently considered. At the moment it seems most promising to find an external sponsor that helps the ZKRD operate the service until a definite business model has been developed. This business model should also include final development of the service which still is in a prototypical status. Experiences from the current operation of the service may help determining the directions for further development.

Acknowledgments

The development of bmt line has been funded by DeTeBerkom GmbH as part of the R&D projects 2100 (MIHMA) and 2101 (MIHMA-P). Graphics design and product identity for bmt line by Magdalene Kübler. Medical advice was provided by Dr. Carlheinz Müller and Thorsten Heitzmann of ZKRD (Zentrales Knochenmarkspender Register für die Bundesrepublik Deutschland). Dirk Lutzebäck implemented the final version of the search component and

the interface between bmt line and the domain expert. Nurhan Yildirim supervised the final production of bmt line.

Bibliography

[1] Grady Booch and James Rumbaugh. Unified method - overview version 0.8, 1995. This is a temporary name for this work; it may be renamed in a future version of these documents.

[2] Roy T. Fielding. Maintaining distributed hypertext infostructures: Welcome to momspider's web. In *Proceedings of the First International Conference on the World-Wide Web*, May 1994.

[3] P.-A. Hasenfratz, Ph. Volle, and M. De Cecco. *BACK++ User's Manual.* Non Standard Logics S.A., Paris, France.

[4] Thomas Hoppe, Carsten Kindermann, J. Joachim Quantz, Albrecht Schmiedel, and Martin Fischer. BACK V5 Tutorial and Manual. KIT Report 100, Department of Computer Science, Technische Universität Berlin, Berlin, Germany, March 1993.

[5] Java: Programming for the Internet. `http://www.javasoft.com/`.

[6] Magdalene Kübler. Entwurf des Erscheinungsbildes und Konzeption der Nutzerführung in einem medizinischen Online-Dienst. Master's thesis, Hochschule der Künste Berlin, 1996.

[7] Magellan – the McKinley Internet Directory. URL: `http://www.mckinley.com/`.

[8] Peter F. Patel-Schneider, Ronald J. Brachman, and Hector J. Levesque. ARGON: Knowledge representation meets information retrieval. In *Proceedings of the 1st Conference on Artificial Intelligence Applications*, pages 280–286, Denver, Col., 1984.

[9] Oli Kai Paulus et al. The MIHMA Project Overview Report. Technical report, Department of Computer Science, Technische Universität Berlin, Berlin, Germany, April 1996.

[10] Christof Peltason. The BACK System - An Overview. *SIGART Newsletter*, 2(3):114–119, 1991.

[11] Albrecht Schmiedel et al. Using the BACK System in MIHMA. Technical report, Non Standard Logics, Paris, April 1996.

[12] Jim Sterne. *World Wide Web Marketing.* Wiley, New York, 1995.

[13] U.S. Department of Health and Human Services, National Institutes of Health, National Library of Medicine. *Unified Medical Language System*, 7^{th} experimental edition, January 1996.

[14] Philippe Volle et al. The MIHMA Project Technology Report. Technical report, Non Standard Logics, Paris, April 1996.

[15] William A. Woods and James G. Schmolze. The KL-ONE family. *Computer and Mathematics with Applications, Special Issue on Semantic Networks in Artificial Intelligence*, 23(2-9):133–178, 1992.

A Expenditure of Annotating New Instances

The approach of an integrative information service underlying bmt line relies on manual evaluation, annotation, and cross-referencing of references to documents and information sites through domain experts and staff of the service provider. This part of the process could easily become the bottleneck of the entire operation, or could become too cost intensive.

In preparation of a business case for a continued operation of bmt line, we collected statistical material gathered through the prototypical operation of bmt line during the MIHMA project. The results of an evaluation of these statistics is presented in this appendix.

The figures evaluated below cover the annotation of new documents through the administrator (cf. Sec. 5.4). They do not include the time needed by the domain expert to provide medical annotations of documents.

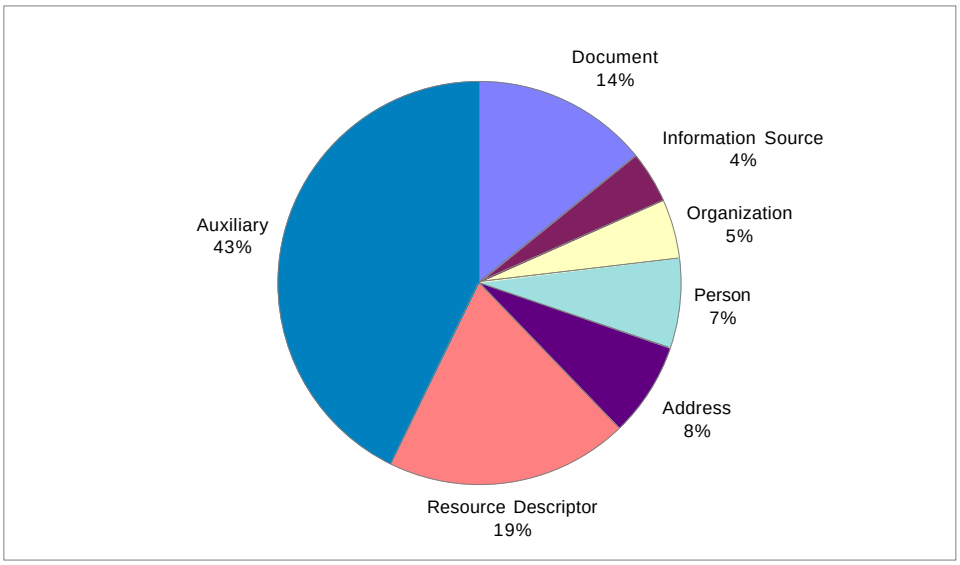

Figure 12: Distribution of Objects Over Main Categories of the Semantic Model (in %)

Until the end of April 1996, 2166 objects have been annotated for bmt line. The distribution of annotated objects over the main categories of the semantic model is shown in Figure 12.

It is interesting to see that only 14 % of all annotated objects are documents, and only 4 % of all annotated objects are general information sources. While documents and general sources are the main entities referenced by bmt line, a large number of further objects have to be modeled to achieve the desired degree of semantical integration. On the one hand, these are entities from the real world such as organizations, persons, or addresses. On the other hand, these are objects basically used for internal operation purposes of

bmt line, including Resource Descriptor objects (which model the location of resources on the Web) and auxiliary objects (among others, the textual descriptions used on catalog pages of the service).

The following figures are based on a selected set of annotated objects. This set contains 109 newly annotated documents; the respective expenditures where measured after an initial training phase of the responsible staff member, and after a stable version of the annotation editor was made available.

For the selection of documents, the chart in Figure 13 shows how long it took to annotate them. Each bar shows the time needed for annotating one document. It is interesting to see that the minimum time required for annotating one document is 3.5 minutes, and the maximum time required is 50.2 minutes. That is a difference of approximately 45 minutes. The average time to annotate one document was 12.5 minutes.

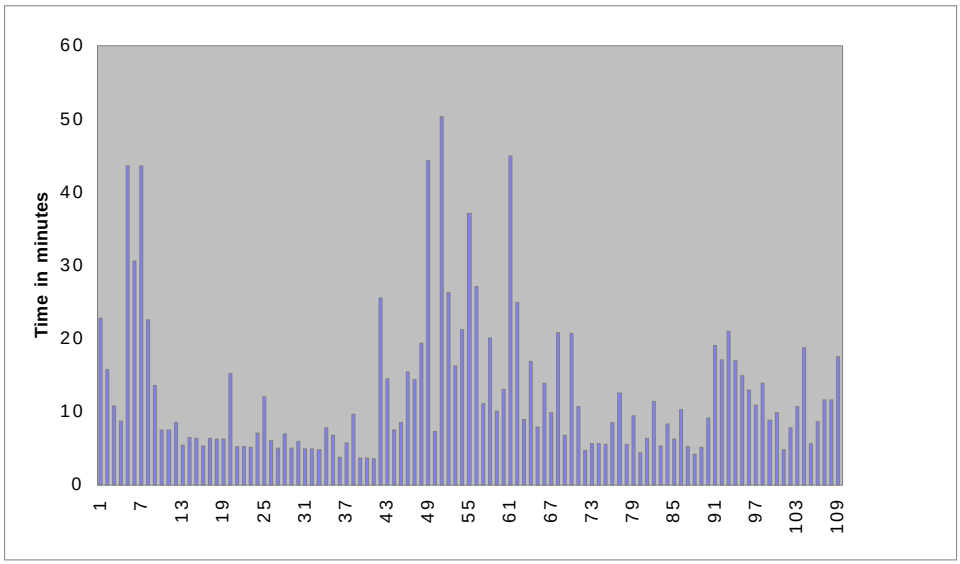

Figure 13: Time Needed for Annotating New Documents

The chart in Figure 14 explains why the periods of time differ. As mentioned above, adding one new document to the object base usually requires several other objects to be added as well, or to change existing objects in order to keep the object base consistent. In Fig. 14 each bar shows how many objects had to be changed or added to describe one new document. The minimum number of objects affected is 2 and the maximum number of objects affected is 22. The average number of affected objects is 5.3.

The differences can have multiple reasons. When only few other objects were affected, this could've been because, for example,

- all other objects had been annotated before (e.g., the document's author and the information source providing the document)

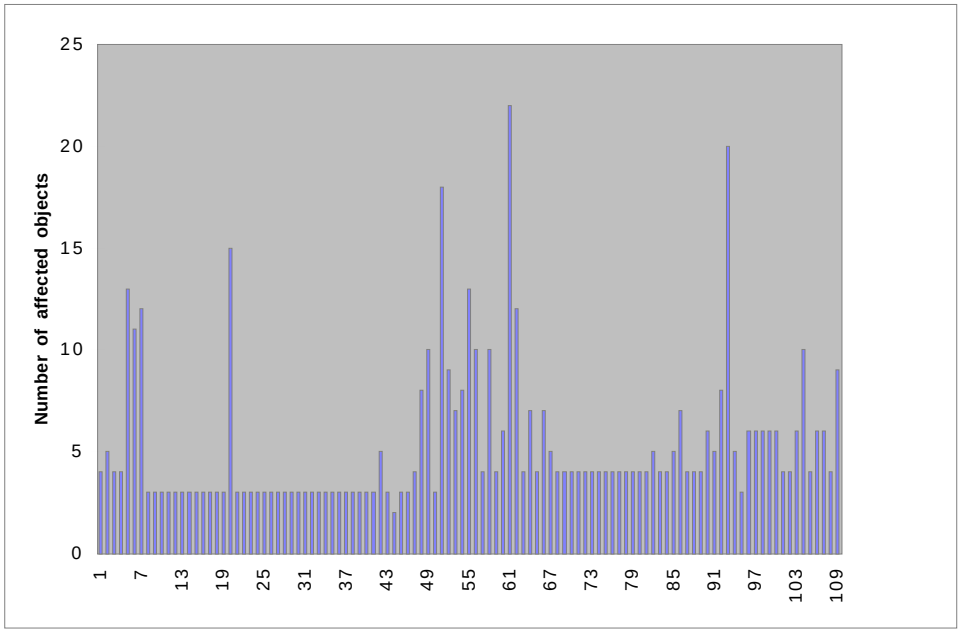

Figure 14: Number of Objects Affected by Adding a New Document to the Object Base

- the respective information (author, organization, address) was still unknown or couldn't be determined.

When many other objects were affected this could've been because, for example,

- the document had multiple authors
- The document was mirrored on another information source (a case explicitly represented in the bmt line object base)
- the document consisted of chapters that were worth being described as separate objects.

For the above two charts, the thin line in the chart of Figure 15 shows the time needed for annotating one document divided by the number of objects affected. The average time to add or change a single object is 2.4 minutes. The thicker straight line is the trend implied by the thin line. It shows that the person who annotated new documents became faster.

The figures discussed above strongly depend on the particular setting in which bmt line was prototypically operated. It is questionable whether these numbers are representative for an operation of bmt line on a larger scale, not to speak of other services for which several parameters may be different, e.g, the attempted degree of integration.

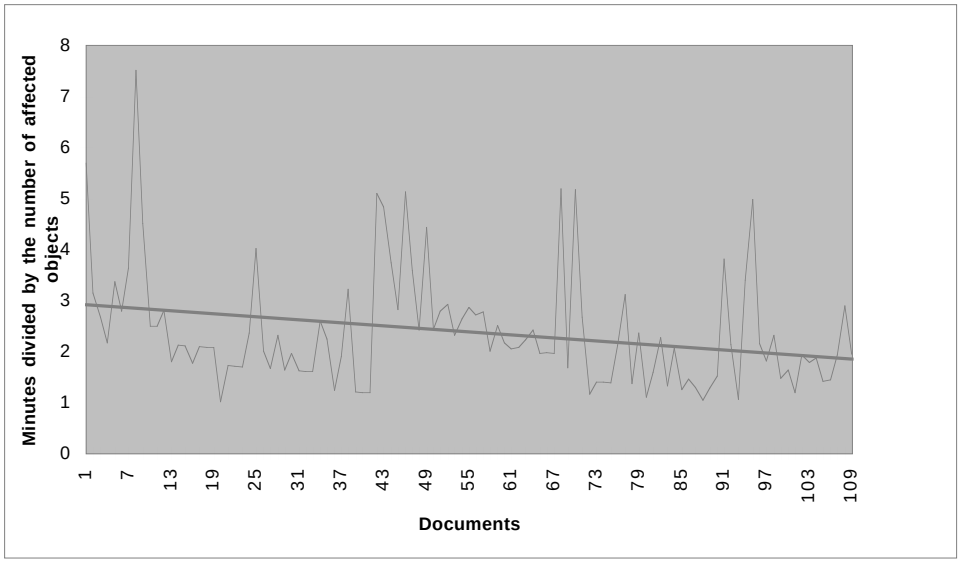

Figure 15: Minutes Divided by the Number of Affected Documents