MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

SGML-based Construction and Automatic Organization of Comprehensive Medical Textbook on the Internet

Kengo Miyo, R.N., P.H.N., M.H.S., Kazuhiko Ohe, M.D., Ph.D.

Hospital Computer Center, University of Tokyo Hospital, Tokyo, Japan

Abstract

The amount of knowledge required in practical medicine is large and ever increasing. Medical staff must select and use appropriate pieces of the knowledge from this flood of medical information. Recent Internet technology may be solving these problems because it makes information open to the public immediately after it is created and enables many people to share it. Medical resources on the Internet are however currently not always well organized, because these are often voluntarily provided by the experts of a particular field. We therefore decided to create a comprehensive medical database on the Internet, which is well organized, and of a high quality for practical medical use.

In order to make full use of the benefits provided by electronic media, we created a new structured data set of information. We then commissioned authors to write manuscripts from which we created Standard General Mark-up Language (SGML) documents. We then wrote a translation program that took the SGML and automatically created a fully inter-linked HyperText Mark-up Language (HTML) document.

The translation program generated 4,814 HTML files created from 1,373 number of SGML documents. The total data size including pictures was about 640MB. 205,775 related links were created. We then published our electronic medical textbook described in HTML publicly on the Internet.

Using SGML-based structured data, we constructed a complex electronic medical textbook created organically from simple SGML instances. Our electronic medical textbook is systematic and comprehensive, and has a homogeneous structure. We believe that this is the first comprehensive medical textbook available on the Internet. Furthermore, it was found that our approach to the electronic medical textbook has two major advantages. One is automatic generation of inter-links among documents, and another is easy to maintain documents. In addition, once we construct the electronic textbase in SGML format, the data can be utilized to various application programs on different platforms. Making use of this feature, we are now planning to develop a new style of electronic textbook, which is closely integrated with a Hospital Information System (HIS). The plan would provide medical staff with on-demand access to the electronic medical textbook while using HIS terminals.

Keywords

Electronic Medical Textbook; Internet; SGML; Automatic Organization

Introduction

The amount of knowledge required in practical medicine is large and ever increasing. Medical staff must select and use appropriate pieces of the knowledge from this flood of medical information [1, 2]. At least two difficulties have been recognized. The first is for the user: how may they acquire the medical knowledge. The second is for the provider: how should they provide it. Recent Internet technology may be solving these problems because it makes information open to the public immediately after it is created and enables many people to share it. In fact, the variety of medical resources are increasingly being placed on the Internet [3 - 6].

However, we often face two major problems in using the Internet resources. The first is that they are not well organized. Secondly, resources don't cover comprehensive medical fields, often because these are voluntarily provided by the experts of a particular field. As these resources increase, they become increasingly disordered. If this situation is allowed to continue we will face increasing problems of acquiring appropriate information in a timely manner. Currently a comprehensive medical database on the Internet, which is well organized, and of a high quality is in great demand for practical medical use.

To solve these problems, we should provide comprehensive and well-organized medical resources, or electronic textbook, onto the Internet environment. Therefore, it is important to find out the appropriate method by which construction of large-scale resources would be archived with less manpower. So far most medical textbooks on the Internet have been described in HyperText Mark-up Language (HTML). It has several advantages, e.g., easy to write, easy to include multimedia data. However HTML can not represent the semantic structure of the document, although it could display the visible layout and the Internet links. Furthermore, the HTML standard is frequently updated. If we described original data using HTML, we might

have to rewrite to keep pace with new standards.

For the reasons given above, the Standard Generalized Mark-up Language (SGML) is proposed for the describing the original data of our comprehensive medical textbook. SGML is platform-independent language, which can represent semantic structure of a document, and is adopted in several medical fields [7 - 11]. We, at first, defined the structure of the electronic textbook for internal medicine using the Document Type Definition (DTD), then commissioned authors to write manuscripts from which we created SGML documents. After collecting all the manuscripts in the SGML format, a translation program was developed to generate fully inter-linked HTML documents. We then published the HTML documents generated by this program making them publicly available on the Internet.

Methods

Construction of the SGML-based Medical Textbook

In the past comprehensive and well-organized medical databases were provided on paper media. Many electronic medical textbooks were made that drew on data from these textbooks. However, by its very nature paper media prioritizes the visual while being ill suited to representing complexly structured documents. These limitations make it difficult simply to create an electronic medical textbook that makes the most of the advantages of electronic media. We decided therefore to create fresh data for our electronic medical textbook.

Creating the Structure of an Electronic Medical Textbook

In order to represent the data anew, we first had to create the structure of an electronic medical textbook. We began by investigating the elements required for an electronic medical textbook. After surveying several such textbooks we extracted semantic elements such as etiology, epidemiology, pathology, diagnosis and so on, and expression elements include semantic elements such as headings, text strength and so on. In addition, we considered other potential elements for electronic medical data, such as synonyms, keywords, document version, and so on. We defined these elements for our electronic medical textbook. Our second step was to classify these elements and consider the relationships between them. In this way we created the structure of our electronic medical textbook and described it using DTD [12].

Making SGML Instances

According to the structure we created, we commissioned authors to write manuscripts. Each author was given the list of elements in our structure and filled in each element. When authors returned their manuscripts to us, we added SGML tags defined by DTD. In this way, we made SGML instances which included 1016 diseases, 59 symptoms and 296 laboratory examinations. Figure 1 shows an example of SGML instance about angina pectorisAutomatic Organization

In order to organically combine the SGML instances we wrote a program with an automatic organization mechanism. This program was written in Microsoft Visual Basic version 4.0. It could automatically extract words relevant to other SGML instances from SGML instances. Links between these words and related SGML instances were automatically created. At the same time HTML files for showing pictures were also created and linked to related SGML instances. Finally, in order to make SGML instances open to the public via the Internet, this program translated all SGML tags to HTML tags..

```
<NVD>970329</NVD>
<NDB>
<NG1>循環器疾患</NG1><NG2>虚血性心疾患</NG2>
<NDN>狭心症</NDN>
<NDE>angina pectoris</NDE>
<NSQ>D255</NSQ>
<NDY>angina of effort</NDY><NDY>spontaneous angina</NDY>
<NDY>stable angina</NDY><NDY>unstable angina</NDY>
<NDY>variant form of angina</NDY><NDY>coronary spastic angina</NDY>
<NDY>stable effort angina</NDY>
<NKY>myocardial ischemia</NKY><NKY>coronary arteriosclerosis</NKY>
<NKY>coronary risk factor</NKY><NKY>coronary spasm</NKY>
<NKY>stunned myocardium</NKY><NKY>hibernating myocardium</NKY>
<NKY>intimal thickening</NKY><NKY>atherosclerosis</NKY><NKY>anginal pain</NKY>
<NKY>Holter心電図</NKY><NKY>percutaneous transluminal coronary angioplasty</NKY>
<NKY>coronary artery bypassgraft</NKY><NKY>CABG</NKY>
<NKY>isosorbide dinitrate</NKY><NKY>ISDN</NKY><NKY>Ca拮抗薬</NKY>
<NKY>β遮断薬</NKY><NKY>percutaneous transluminal coronary recanalization</NKY>
<NKY>PTCR</NKY><NKY>intraaortic balloon pumping</NKY><NKY>IABP</NKY>
...
<NAB><NAN>山本　一博</NAN><NAI>大阪大学医学部第1内科</NAI></NAB>
<NFB>
<NFL ID = D255P010><CAP>[ 写真1 ]冠動脈の断面図(1)</CAP><NTE>内膜の肥厚を認めるが，比較的正常冠動脈壁の構築が認められる冠動脈の断面像．</NTE></NFL>
...
</NFB>
<EDF>
狭心症とは，心筋における酸素の需要と供給の均衡が破れ，心筋が一過性に虚血に陥るために生ずる胸部症状を主症状とする臨床症候群である．発現機序，重症度，発作発現状況からいくつかに分類される．<P>
発現機序から労作性狭心症(angina of effort)と安静狭心症(spontaneous angina)に，また症状の推移から安定狭心症(stable angina)と不安定狭心症(unstable angina)に分類される．また，安静狭心症のうち，発作時の心電図でST上昇が認められるものを異型狭心症(variant form of angina)と呼ぶこともある．<P>
...
</EDF>
<EET>
狭心症は心筋の一過性の虚血により発生し，心筋虚血(myocardial ischemia)は冠血流量の絶対的ならびに相対的減少，心筋酸素消費量の増加，動脈血の酸素運搬能の低下などにより生ずる．
<EET1>
<EET1HD>A. 冠血流量の減少</EET1HD>
心筋虚血をきたす原因として，主に冠動脈硬化による狭窄と冠攣縮による狭窄および冠動脈血栓が挙げられる．
...
```

Figure 1 - An Example of SGML Instance

Linking Related SGML Instances Together

This process consisted of three steps. First, the program extracted synonym elements and keyword elements of SGML instances to create a database of related words. Second, using this database words for links, which are included in each SGML instance, were identified. Third, anchors were added to identified words. All anchors gave information on the link target. Through this process, each SGML instance was linked to other SGML instances based on their semantic relationship.

Creating HTML Files for Showing Pictures

Pictures are shown in a different window from document one. This allows users to see both the information that is in the pictures and the documents related to them. This program also created HTML files for showing pictures. It extracted figure description elements from SGML instances. Based on this description HTML files for showing pictures were created. Simultaneously, anchors were added to all related words of SGML instances to link make links between these HTML files and related SGML files.

Translation of SGML Tags to HTML Tags

Finally, the program translated all SGML tags included in the documents to HTML tags. In order to perform the translation, it refers to a translation table, which described the correspondence

of SGML tags to HTML tags. This automatic process creates necessary links among related HTML files and, as the result, generates a complete electronic medical textbook in HTML format.

Making Search-Functions

For searching documents, we wrote a Common Gateway Interface (CGI) program. This program was written in the C language. When users input any keywords, it uses the database created above to search and present documents.

Results

Generated Data

Data size of HTML files generated from SGML instances was about 32MBs. Data size of pictures used in our medical textbook was about 610MBs. 4,814 HTML files generated from 1,373 SGML instances. 205,775 related links were created. We used a Windows NT 4.0 PC with a Pentium 166 MHz processor for the data generation. It took us 5 hours and 27 minutes to generate all data.

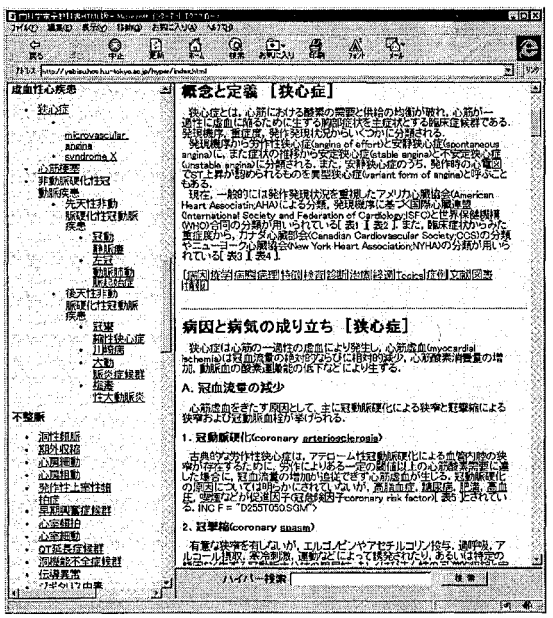

Figure 2 - User Interface of Our Electronic Medical Textbook

User Interface and Functions

Figure 2 indicates our electronic medical textbook. It can be displayed using Netscape Navigator version 3.0 or later, or Microsoft Internet Explore version 3.0 or later. The left area of Figure 2 displays a table of contents. A table of contents is constructed using a tree structure based on the levels of documents. The right area displays the documents.

This area of Figure 2 displays the HTML document of angina pectoris. Documents are divided into sections based on semantic elements such as etiology, diagnosis, treatment and so on. A list of sections included in each document is shown at the end of each section. If users click on this list, the section selected is displayed.

Figure 3 - Sample of Picture Windowr

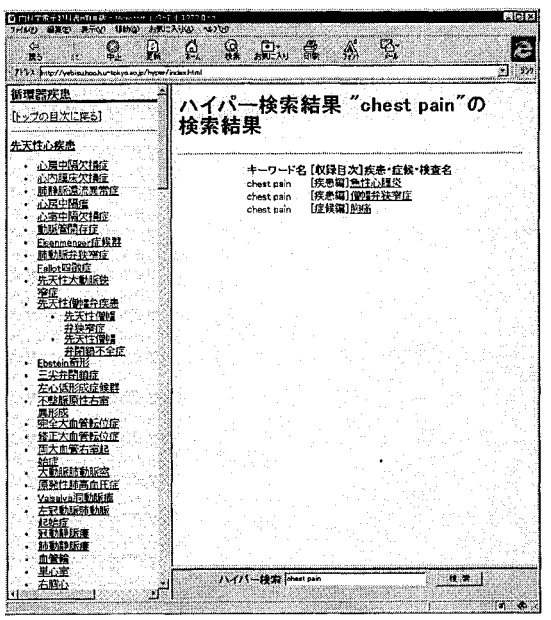

Figure 4 - The Results of Searching the Documents related to "Chest Pain"

Words related to other documents are rendered in blue. If a user clicks on such a word, a related document is displayed. Pictures elated to documents are also linked from the documents. If users click the words linked pictures, another window is opened and the picture is shown. Figure 3 indicates a sample of picture window. The picture of angina pectoris taken with coronary angiography is shown in this window. In some pictures, users can apply further functions such as showing high-resolution (1200 pixels X 1600 pixels) image, or selecting some nidi within the image to be shown, etc.

Users can search documents by inputting any text in the text box shown below the right side of Figure 2. Above right of Figure 4 indicates the results of searching. When users select one of these results, the selected document is displayed.

Discussion

Until now, trials of systematic provision of medical resources via the Internet have been underway. William R. Hersh et al. have provided linked list of medical resources on the Internet constructed using the MeSH thesaurus [13]. Their work is important but still has the problem that present resources on the Internet have a heterogeneous structure. Our electronic medical textbook is systematic and comprehensive, and has a homogeneous semantic structure. We believe that this is the first comprehensive medical textbook available on the Internet.

It was found that our approach to construct the well-organized medical resources using SGML has two major advantages as follows:

1. It allows automatic generation of inter-links among documents, because each original SGML-based document completely includes the relational information, which could be referred to by other documents.
2. We can manage and maintain the SGML instances independently from each other, because each document has no relational information, which refers to other documents explicitly.

These are very important to maintain a medical textbook, which has a complex body of knowledge and is rapidly changing.

Future Work

A key aspect of this study is adoption of SGML in order to represent the structures of medical documents. SGML doesn't depend on a particular platforms or particular applications. Once we construct the electronic textbase in SGML format, we can utilize the data to various application programs on different platforms. Thanks to this feature, we could develop an original application running on Windows 95, by which medical staff can browse all the content of the SGML-based data as comprehensive medical textbook. Furthermore, making use of this feature, we are now planning to develop a new style of electronic textbook, which is closely integrated with a Hospital Information System (HIS) [14]. The plan would provide medical staff with on-demand access to the electronic medical textbook while using HIS terminals.

Conclusions

We developed an SGML-based comprehensive electronic textbook by using an automatic organization mechanism to translate SGML instances to HTML documents. We suggested the usefulness of the structured data described in SGML and of our automatic organization mechanism.

Acknowledgments

We wish to thank Mr. Masayuki Kajino, Mr. Tetsuo Sato, Ms. Atsuko Yamamoto and Ms. Mariko Nakai of Nakayama-Shoten Co. Ltd, Tokyo, Japan.

References

[1] Evidence Based Medicine Working Group. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*. 1992; 268:2420-2425.

[2] William R and Anna D. Evidence based medicine: An approach to clinical problem-solving. *The British Medical Journal*. 1995; 310:1122-1126.

[3] Kleeberg P. Medical uses of the Internet. *Journal of Medical Systems*. 1993: 17: 363-366.

[4] Cimino JJ, Socratorus SA, and Clayton PD. Internet as clinical information system: application development using the World Wide Web. *Journal of American Medical Informatics Association* 1995: 2: 273-284.

[5] Lowe HJ, Lomax EC, and Polonsky SE. The World Wide Web: A review of emerging Internet-based technology for the distribution of biomedical information. *Journal of American Medical Informatics Association* 1996: 3: 1-14.

[6] Cimino JJ. Beyond the Superhighway: Exploiting the Internet with Medical Informatics. *Journal of American Medical Informatics Association* 1997: 4: 4: 279-284.

[7] Pearson P, Francomano C, Foster P, Bocchini C, Li P, and McKusick V. The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Research*. 1994: 22: 17: 3470-3473.

[8] Lincoln TL, and Essin DJ. A document processing architecture for electronic medical records. *Medinfo*. 1995: 8: 1: 227-230.

[9] Jonassen K and Saboe R. The use of text encoding in the development of a terminology and knowledge system associated with the Norwegian version of the ICD-10. *Medinfo*. 1995: 8: 1: 51-55.

[10] Kahn CE Jr. A generalized language for platform-independent structured reporting. *Methods of Information in Medicine*. 1997: 36: 3: 163-171.

[11] Yoshihara H, Ohe K, Ohashi K, Yamamoto R, Yamazaki S, Hirose Y, Matsui K, Hishiki T, Yamashita Y, Kimura M, Minagawa K, and Oyama H. Standardization of exchange procedures of clinical information, and an experiment of clinical data exchange using Medical Markup Language (MML).http://www.miyazaki-med.ac.jp/medinfo/SGmeeting/document/MML_Sympo97/MML.html

[12] Maler E, and El Andaloussi J. *Developing SGML DTDs, From Text to Model to Markup*. New Jersey: Prentice Hall, 1996.

[13] Hersh WR, Brown KE, Donohoe LC, Campbell EM, and Horacek AE. CliniWeb: Managing clinical information on the World Wide Web. *Journal of American Medical Informatics Association* 1996: 3: 273-279.

[14] Miyo K, and Ohe K. An Interface to Retrieve Electronic Medical Textbook from a Hospital Information System. *Proceedings of the Second China-Japan Joint Symposium on Medical Informatics* 1997: 84-88.

Address for correspondence

Hospital Computer Center
University of Tokyo Hospital
7-3-1, Hongo, Bunkyo-ku
Tokyo
Japan.
mailto: miyo@hcc.h.u-tokyo.ac.jp
http://yebisu.hcc.h.u-tokyo.ac.jp/index.html