XML 電子辞典システムの開発

00KI000 電大 太郎[†]　　指導教員 絹川 博之

Development of Dictionary System by Using XML

Taro Dendai[†] and Hiroshi Kunukawa

1. はじめに

XML は電子書籍表現に関して業界標準になりつつある．そして今 XML 電子書籍を効率よく利用するためのシステムが求められている．XML 電子書籍システムの対象となる書籍の種類としてまず以下の 3 つを想定する．

- 辞典・辞書
- 学術書
- 新書

今回はその中で辞書・辞典を扱う XML 電子辞典システムを開発する．

辞典は，ある専門分野の知識を得るために利用される．ある用語の意味，つまりその概念を理解するということは，その概念の内容を知るということだけでは不十分であり，その概念がその専門分野の概念体系の中でどのような位置を占め，他の関連諸概念とどのような関係を持っているかということも知ることが必要である．これらの体系，概念相互間の関係を本という媒体で表現するには限界がある．そこで辞典を XML 電子辞典として扱い，利用者が効率よく知識を得るためのシステムが XML 電子辞典システムである．

2. XML 電子辞典システム

本システムは，対象となる辞典の XML 電子辞典データを作成する行程とその作成されたデータを対象に検索・表示する行程からなる（図 1）．XML 電子辞典システムでは，対象となる辞典を XML 表現形式に変換することが必要となる．今回は以下の辞典の記述内容を対象とする．

- 岩波情報科学辞典[1)]
  （1） 見出語とその説明
  （2） 用語の木
- 岩波ジュニア事典シリーズ
  （1） 見出語とその説明

これらの記述内容を本システムで利用可能な共通形式の辞典 XML データにするための変換プログラムが必要である．

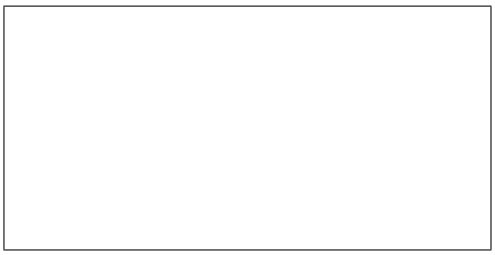

図 1　XML 電子辞典システムの概要

3. XML 電子辞典フォーマット

辞書・辞典を XML 電子書籍として表すためにの形式として DicX[2)] というフォーマットが存在する．DicX は見出しに関する情報の表現に限定されているので「用語の木」の表現を追加した ex-DicX (Extended-DicX) を提案する．

ex-DicX は次の 2 つから構成されている．

（1） ex-DicX-term
ex-DicX-term は見出し語とその説明の表現用であり DicX と同じである．その構造は図 2 のとおりである．

（2） ex-DicX-thesaurus
ex-DicX-thesaurus は用語の木の情報の表現形式であり，各ノードの子ノードが下位語となるようにして木を作り上げている．その構造は図 3 のとおりである

4. XML 電子辞典データの作成

対象の辞典データから本システムが取り扱う ex-DicX 形式のデータを作成する．

- ex-DicX-term
  – 岩波情報科学辞典（1 冊分）
  – 岩波ジュニア辞典シリーズ（6 冊分）
- ex-DicX-thesaurus

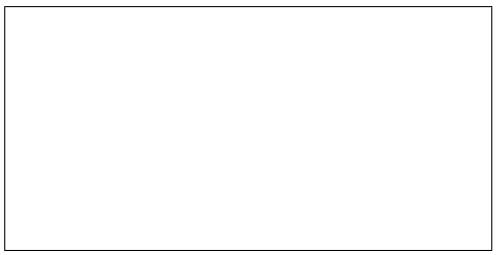

図 2　ex-DicX-term

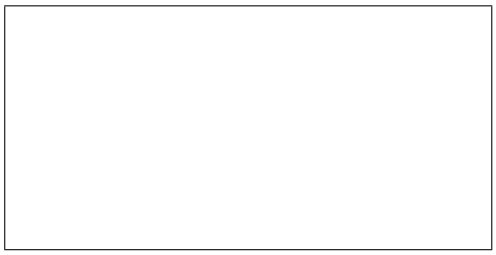

図 3　ex-DicX-thesaurus

† 計算言語学研究室
Computational Linguistics Laboratory

図 4 システムアーキテクチャ

– 岩波情報科学辞典（1 冊分）

4.1 岩波情報科学辞典

4.1.1 見出語とその説明 ex-DicX-term

「岩波情報科学辞典」の見出語とその説明に関する電子データとしてはテキストデータと XML 形式データの 2 形式が存在する．

(1) XML 形式データ
このデータは岩波情報科学辞典独自の XML 形式フォーマットであり ex-DicX 形式とは異なる．見出しに関するデータなのだが，各見出しには木番号に関する情報が含まれていない．

(2) テキストデータ
見出しに関するデータ．説明文のない同義語・参照用語からなる見出しが抜けている．

両者の不足情報を相互に補うと「岩波情報科学辞典」の本の情報を再現することができる．そのための岩波情報科学辞典 XML 化プログラムを作成し，ex-DicX 形式の XML データを作成した．

4.1.2 用語の木 ex-DicX-thesaurus

「岩波情報科学辞典」のシソーラスを表す「用語の木」に関する電子データは存在しない．そこで今回は「岩波情報科学辞典」の木から手作業によって ex-DicX 形式のデータを作成した．

4.2 岩波ジュニア辞典シリーズ ex-DicX-term

「岩波ジュニア辞典シリーズ」は XML 形式データとして存在するが，6 冊とも本システムが扱う ex-DicX 形式とは異なる岩波ジュニア辞典シリーズ独自のフォーマットであり，4.1.1 節の (1)(2) とも異なる．そこで独自の XML 形式データから ex-DicX 形式の XML データに変換するための岩波ジュニア辞典シリーズ XML 化プログラムを作成し，「岩波ジュニア辞典シリーズ」の ex-DicX 形式の XML データを作成した．

5. 検索・表示プログラムの機能と構成

XML 電子辞書の検索・表示機能は以下のとおりであり，そのプログラムの構成を図 4 に示す．

- 複数辞典 XML データ制御機能
- 参照用語タグ付け機能
- 用語の木と見出し表示機能
- 全文検索機能
- KWIC 検索機能

5.1 制御機能の必要性

4 章でも述べたが，ex-DicX 形式の辞典は 7 種を数える．

表 1 辞書データの容量

辞典名	内容	容量
情報科学辞典	見出語とその説明	4.93MB
情報科学辞典	用語の木	105KB
岩波ジュニア辞典シリーズ	見出語とその説明	2.38MB

今後新たに ex-DicX 形式の辞典が作成される可能性がある．そこで新たに辞典が追加されることを想定した柔軟なシステムにする必要がある．また複数の ex-DicX 形式 XML データを一度に扱うことのできるような制御機能が必要となる．

5.2 高速化の実現方法

XML データはプログラムで扱う際に DOM ツリーという形でメモリに読み込まれる．DOM とは XML 文書のための API である．DOM は一度にすべての文書をメモリに取り込むため，その読み込み時間によって処理を遅くしてしまう．そこで要求があるたびにメモリに読み込むのではなく，システム起動時にメモリに読み込み常駐させ，どの DOM ツリーを使うことで処理の高速化を実現する．

5.3 複数 XML データの制御

複数の辞典 XML データの見出し情報の部分を取り出しマージして 1 つの DOM ツリーにする．こうすることで 1 つの DOM ツリーのみで複数の辞典を同時に扱うことが可能になる．また，マージする際に各見出しがどの辞典の情報なのかということを示すために，取り出したノードの親ノードにそれぞれの辞典名，ISBN を属性として付加する．

6. 考　　察

本研究で作成した辞書データの容量を表 1 に示す．

検索表示プログラムでは起動時に辞書 XML データを DOM ツリーとしてメモリ上に常駐させている．これは処理の高速化を実現すると同時に多くのメモリを消費していることになる．大きな容量の辞書データを利用しようとした場合メモリに乗らない可能性がある．

7. おわりに

本という媒体での辞典とは断片的な情報が五十音順に並んでいるだけで，用語の意味的な情報としては不規則な並びになっている．これでは辞書の持つ本来の特徴を活かすことができない．本システムではそれら断片的な情報をお互いにリンクさせることによって，自分の調べたい用語に関する情報のみを集めることができる．

今回は辞書についてのサブシステムとして開発を行ってきたが，学術書，新書，辞典を連携させ，使い勝手の良い XML 電子書籍システムとして完成させることが今後の目標である．

参 考 文 献

1) 長尾 真 他, 岩波情報科学辞典, 岩波書店, 1995.
2) DicX, `http://www.dicx.org/`
3) 北内 啓, 宇津呂 武仁, 松本 裕治, 誤り駆動型の素性選択による日本語形態素解析の確率モデル学習, 情報処理学会論文誌 Vol.40, No.5, pp.2325–2337, May 1999.