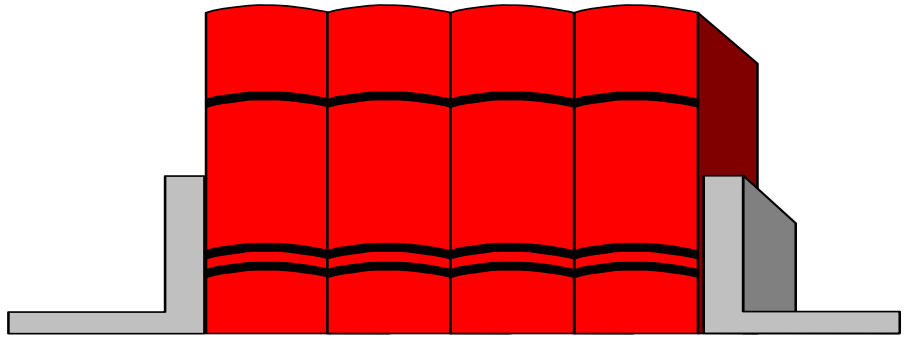
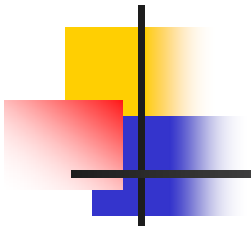


XML全文検索エンジンBTONIC

2004.04.17　XMLコンソーシアム

イースト株式会社　下川　和男

shimokawa@est.co.jp



目次

- BTONICとは
- **開発事例**

BTONICとは

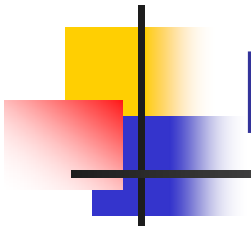
- XML大量ドキュメントの検索システム
  - 12万頁の官報をどう検索するか？
- 三つのインデックスを生成
  - タグ・インデックス：論理構造
  - キーワード・インデックス：検索項目
  - 全文検索インデックス：フルテキスト検索
    ↑生成ツールLaBamba
  - XMLデータ群

BTONICとは(2)

- EXI(EAST XML Index)による高速検索
- データ(.EXIファイル)をシームレスに利用
  - 一回のオーサリングで、WebでもPalmでも
- ×即時更新系データの検索には不向き
- 検索専用なら、高価なデータベース･ソフトは不要
- SQL DBでは表現できない複雑な構造に対応

全文検索　LaBamba

- 全文検索用のインデックス生成ツール
  - 文字パターン方式
  - 形態素解析方式ではないので、検索漏れが起こらない
- インデックスのサイズが小さい
  - 本文(タグなし)の90～120%
  - 大辞林：テキスト30MB、タグ付き76MB、インデックス27MB
    全体で、125MB(XML部分非圧縮)
- 生成時間(大辞林の場合、Pentium III 800クラスで)
  - LaBamba全文インデックス生成　約5分
  - .EXIファイル生成(主にパース)　120~150分

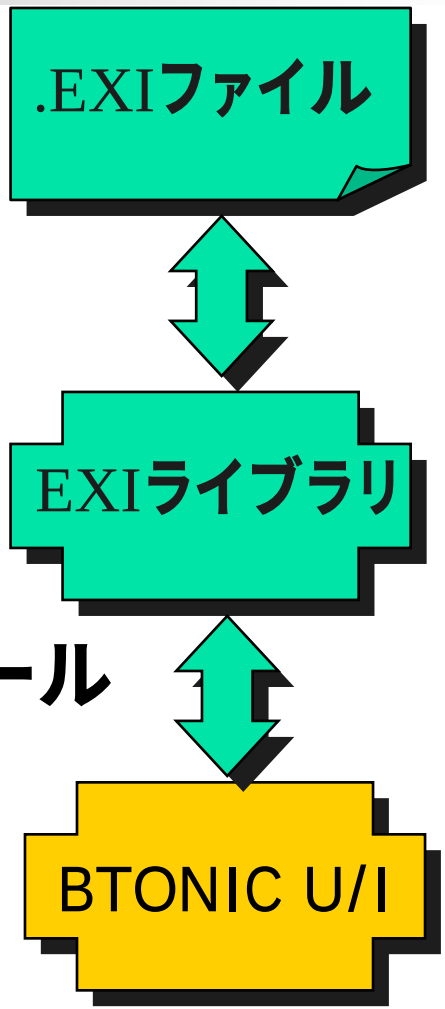


BTONICで扱えるXMLデータ

- 複数の元になるXMLファイルで構成
- インデックスはデータ群に対して付ける
- 一本の大きなXMLファイル
  - 辞書、年鑑、事典、名簿、書誌など
- 複数のXMLファイル
  - 雑誌、新聞、議事録、論文、官報など

BTONICの実体

- .EXIファイル
  EAST XML Index
- EXIアクセス・ライブラリ
  (プログラム)
- ユーザインタフェース・モジュール
  ブラウザー上で動く
  WEB/LAN/パッケージ



EXIライブラリの機能

- **全文検索**
  - **前方/後方/完全一致、ワイルドカード**
- **キーワード検索**
  - **前方/後方/完全一致、ワイルドカード**
- **項目別検索**
  - **項目間**and/or、from:to
  - **前方/後方/完全一致、ワイルドカード**

様々な用途

- **インターネット**(WEB)　Windows NT/2000
- **イントラネット**(LAN)　Windows NT/2000
- **パッケージ**(CD-ROM、**ダウンロード**)
  Windows PC、Pocket PC、Palm、Zaurus
  Mac、Xbox　(**年内に提供予定**)
- **インターネット版は、**iモード、PDA**にも対応**
- XML Web**サービスにも対応**

商品とサービス

- BTONIC
  - **サーバ**:Web**版**、LAN**版**
  - **単体**:　PC、CE、Palm、Zaurus、Mac
- BTONIC**オーサリング・ツール**
- BTONIC**電子辞書取次サービス**
  - DicX(http://www.dicx.org)
- BTONIC**記事検索サービス**(NewsBOX)
  - NewsML

BTONICの性能・機能評価

- http://www.asahi.com　へ
- 左上の「辞書」を選択
- 「大辞林」と「全文検索」をチェック
- 入力窓に「青森　温泉」、「ドイツ　犬」など and検索

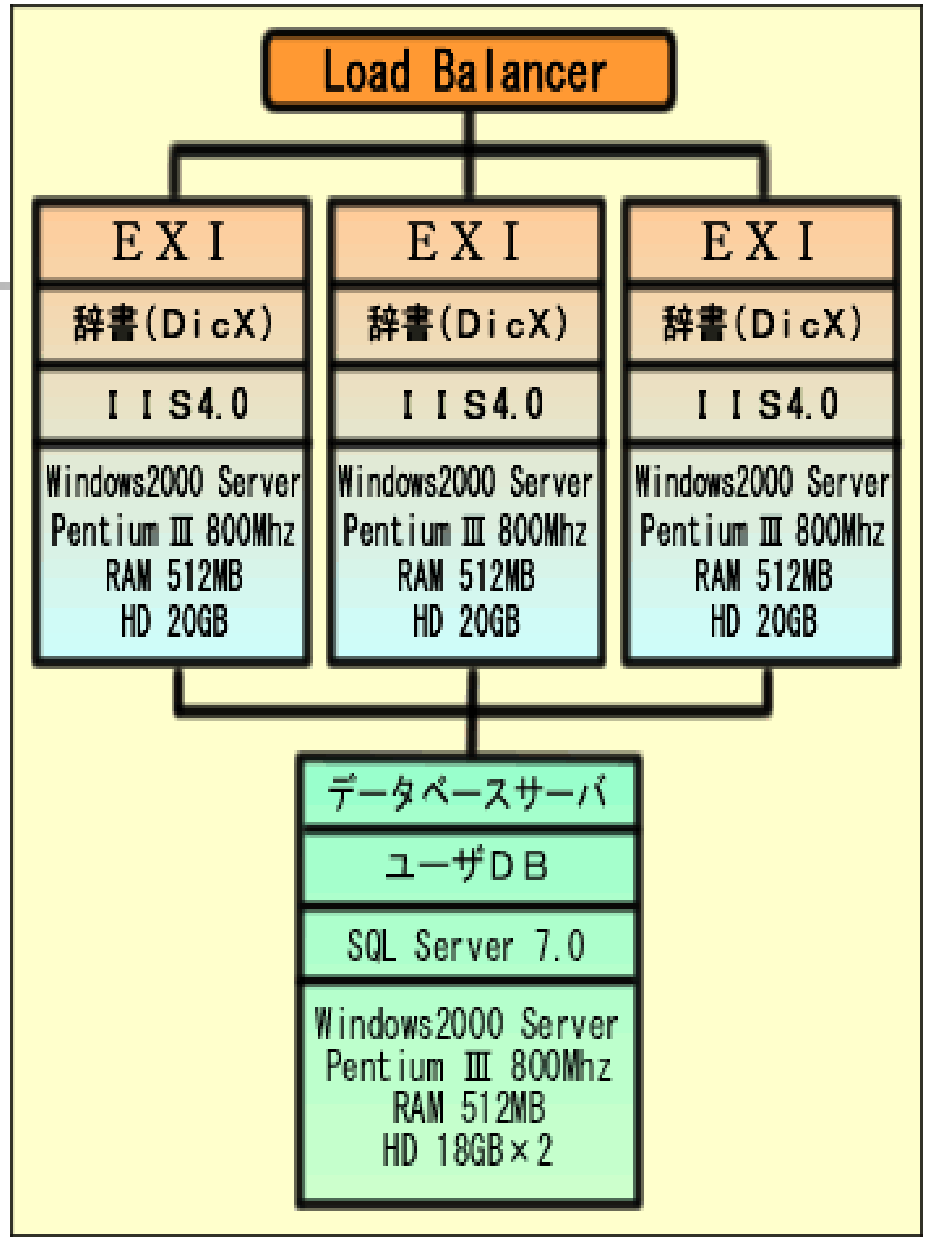
  Googleと同じ心地よさ

事例1: 三省堂　e辞林

- http://www.sanseido.net
- 16点、140万語の辞書をXML化(DicX)
- BTONICでの串刺し検索
  - 複数のXMLファイルを検索
- 決済システム
  - CyberCash(VISA,Master,JCB)

Sanseido.netサーバ

- EXI搭載サーバ　3台
- SQLサーバ　1台
  - 会員管理、ログ管理
  - カード決済
- ロードバランサー　1台
  - URLを3台に振り分け
  - 最大8台まで接続可能
- 回線　4MB(KDDI新宿)
- Asahi.comなどと提携



事例2: ICD病名検索

- http://www.dmsi.co.jp
  - 有料サイト、ビジター・トライアルは可能
- 日本の病名⇒世界標準の病名番号
  - ローカルな病名を登録可能
  - 青ぞこひ⇒緑内障⇒H40(ICD)
- 階層のあるコードブックのような体系なので、XMLに合う。

事例3: 有斐閣　判例CD

- EXI＋IEアプリケーション(HTA)
  - パソコン上で単独に稼動
- 三種類の辞書をHD上で串刺し検索
  - 判例六法、判例百選、判例小事典
- 書籍の項目間ジャンプ

事例4: 施策資料検索

- BTONIC＋アプリの受注作業
- 施策ドキュメントを一般から検索
- アクセス数が多い　⇒　BTONIC
- 施策本文は検索ではなくリンク
  - 見出し、概要、キーワード、カテゴリーのみ
- カテゴリー検索(大分類、小分類)
- 全文検索、キーワード検索

事例5: 小学館JapanKnowledge

- http://www.japanknowledge.com
- 二点の辞書をWebサービス方式で配信
  - 自由国民社：現代用語の基礎知識
  - 日経BP社： デジタル大事典
- SOAP仕様と体験サイトを公開
  - http://btonic.est.co.jp/NetDic/NetDicv05.asmx
  - http://btonic.est.co.jp/NetDicTest/TestV05.aspx
- 事例5:以降はBTONIC＋.NETフレームワーク

事例6: 日本書籍出版協会Books

- http://www.books.or.jp
- 4/16　**一般公開**　1000**万**PV/**月**
- **凸版印刷＋イースト　共同運営**
- **項目間**AND**検索**
  - **書名、著者名、出版社名**
- From-To**検索**　　**発行年**
- **複数ファイルリンク**　　**書誌**XML、**出版社**XML、**著者**
- **iモード、**PDA**にも対応** Mobile Internet Toolkit

事例7: 国語研究所　JiBOOKS

- http://www.kokken.go.jp/public/jibooks.html
- Booksの海外向けサイト　英語環境で稼動
- BooksをWebサービスとして使用
- 三種類のWebサービスの集合体
  - Books　書籍情報検索
  - 文字鏡フォントサーバ　ビットマップフォント配信
  - かな変換　ローマ字→ひらがな変換

事例8: 有斐閣　心理学辞典

- A女子大向け、辞書配信システム
- 学生が大学にログイン
  - アクセス権のある学生の場合はリンク
- Webサービス方式ではない
  - 大学側のシステム開発の問題で

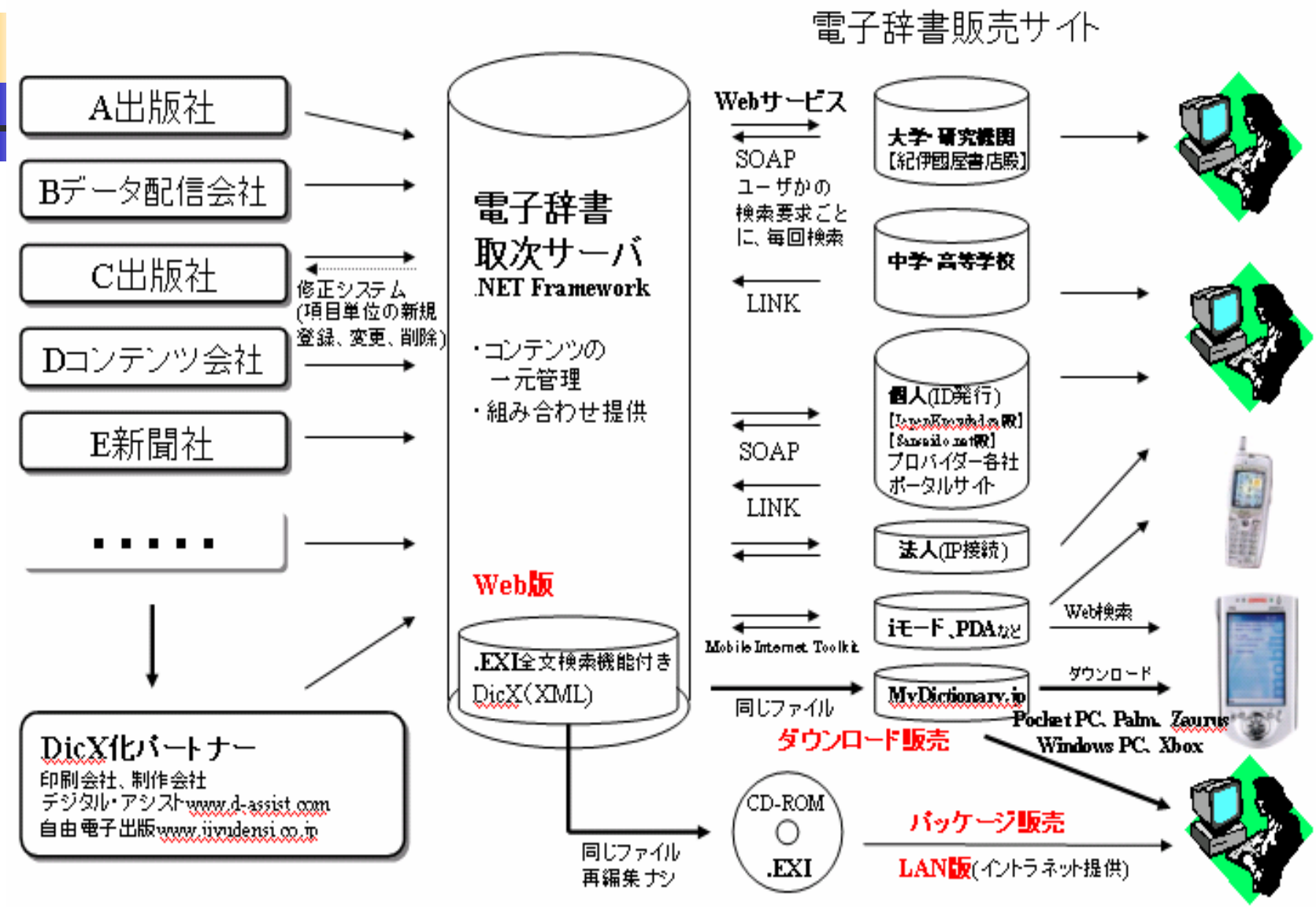
事例9: Grove世界音楽事典

- 5月販売開始　小学館/講談社/イースト
- 世界最大のクラシック音楽事典
  - 日本語版　全20巻　80MB
- バッハ、日本など項目内に目次あり
  - 三階層の目次を項目内で表現
  - 単行本1冊ほどの情報
- 英語版とのローマ字串刺し検索
  - Khachaturian　ではなく　ハチャトリアン

事例10: 新聞検索

- 8月公開予定
- 業界新聞の記事検索(10年分)
- 記事入稿･編集システムも開発
  - こちらは、SQL Server
- SQL ⇒ NewsML ⇒ EXI　自動生成
  - 月水金に自動更新

電子辞書取次 (BTONICホスティング・サービス)



BTONIC体験キャンペーン

- **大量XMLドキュメントを貴社限定で配信　【無料】**

- XML、CSV、**タグ付きテキストなどをご提供ください。**
- **1ヶ月以内に試作検索サイトのURLをご連絡します。**

- **百聞は一見にしかず**

お問い合わせは

- http://www.btonic.com
- http://www.est.co.jp

- **営業担当**： kumano@est.co.jp