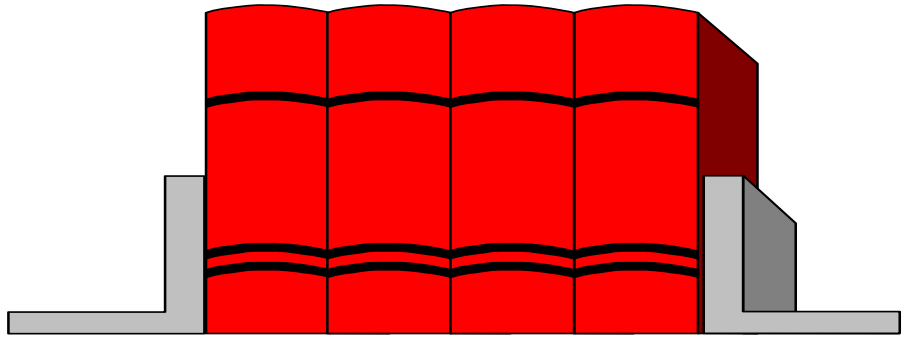
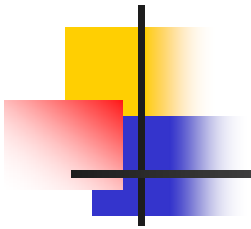


Webサービス事例　辞書検索

2001.11.5　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

- **辞書のXML化**　(DicX)
- **検索エンジン**　BTONIC
- **辞書検索サイトの構築**
- Web**サービスの事例**

DicX

- **辞書・事典構造をXMLで定義**
- http://www.dicx.org
  - Free & Open ＝ JepaX
- **データ構造・記述形式を統一し、辞書間の連携も可能**
- **平易・簡潔なタグの採用**
  →編集作業の効率UP
- **現在　第0.4版　⇒　年内に1.0版へ**

対応コンテンツ

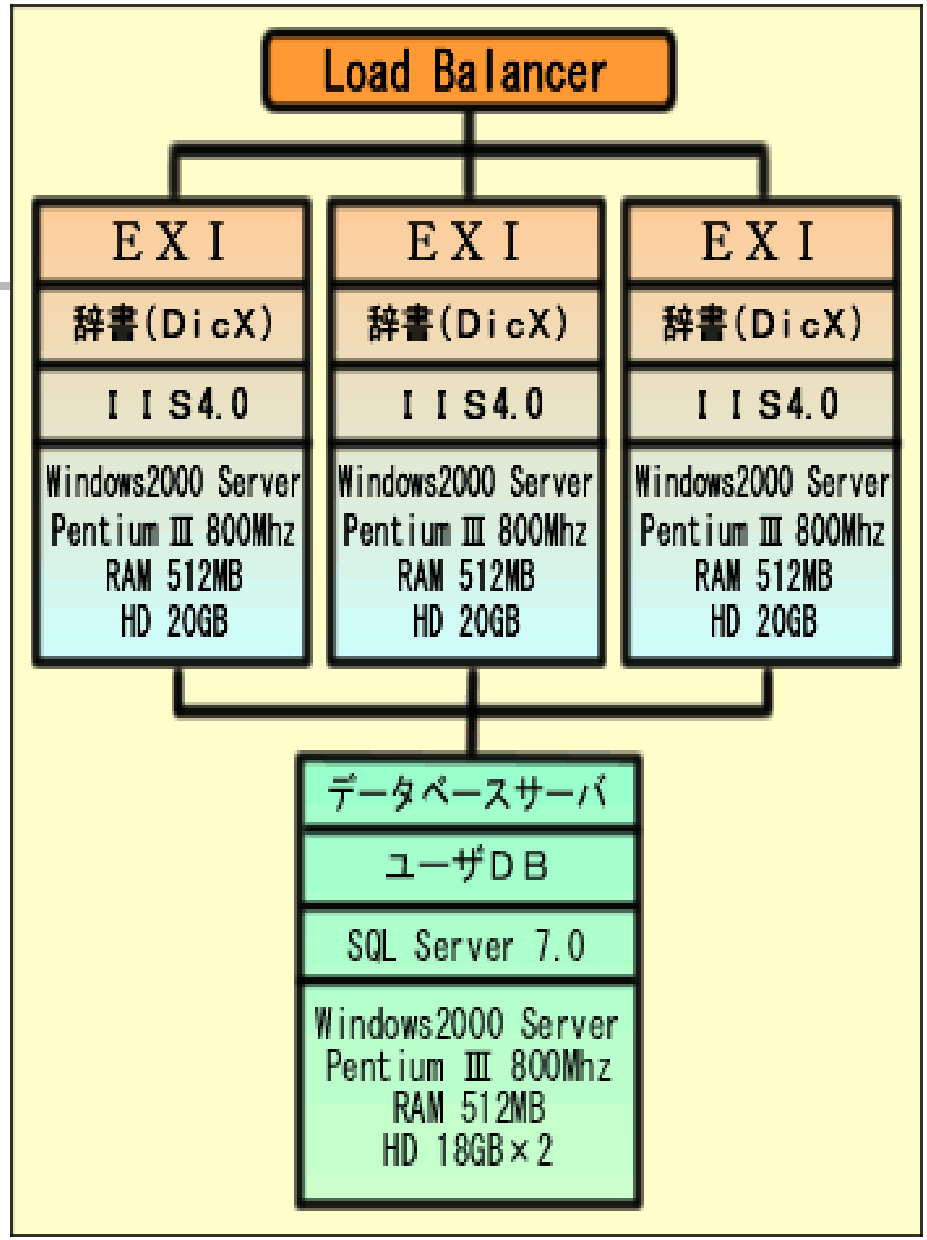
- 小学館　大辞泉、プログレッシブ英和・和英
- 講談社　EOJ、音楽大辞典、NAVIX
- 三省堂　大辞林、エクシード英和・和英、クラウン独和・仏和、デイリーコンサイス英和・和英・国語、新明解国語、センチュリークラウン英和・和英、日本・世界地名など
- 約60タイトルをDicX化しWebサービス
- 大量XMLドキュメントなら、何でもOK

三省堂 e辞林

- 16点の辞書をXML化(年内に18点)
- 大量ドキュメント検索システム
  - BTONIC(ビートニック)
- 決済システム
  - CyberCash(VISA,Master,JCB)
- 運営
  - 新宿KDDIにハウジング

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

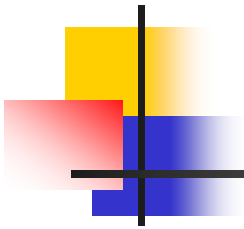


BTONICとは

- 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

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

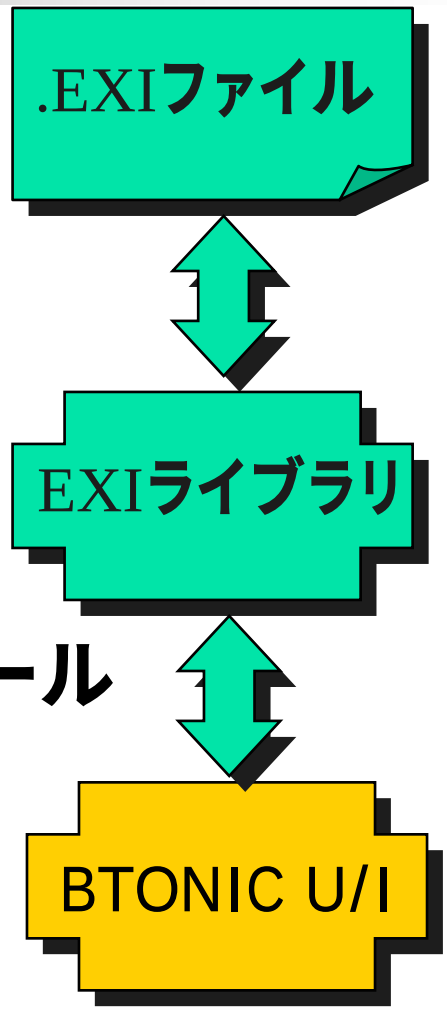
- 複数の元になるXMLファイルで構成
- インデックスはデータ群に対して付ける
- 一本の大きなXMLファイル
  - 辞書、年鑑、事典、名簿、書誌など
- 複数のXMLファイル
  - 雑誌、新聞、議事録、論文、官報など

BTONICとは(2)

- EXI(EAST XML Index)による高速検索
- 様々な用途に対応
  - 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
- データ(.EXIファイル)をシームレスに利用
- ×即時更新系データの検索には不向き

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

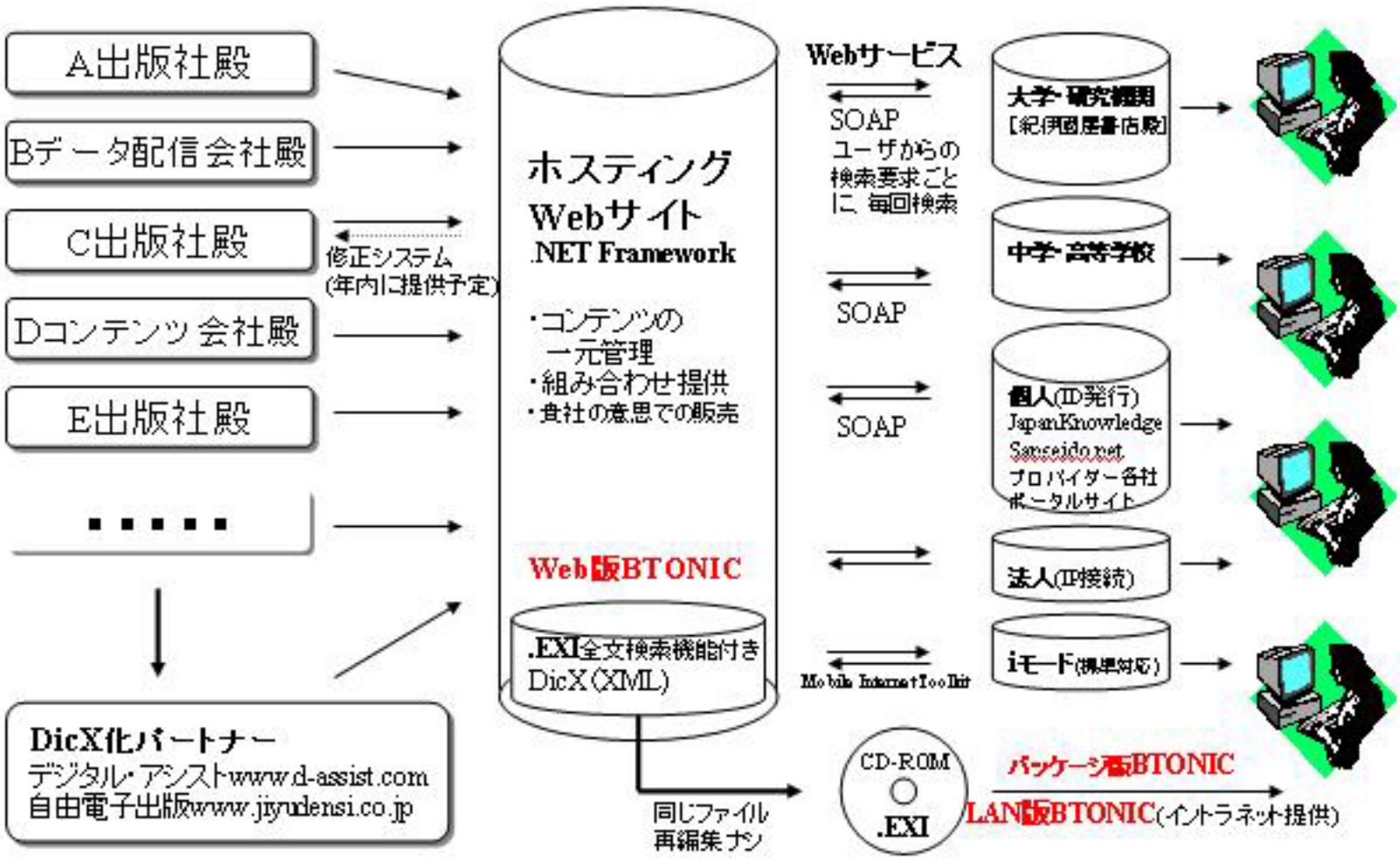
- **全文検索**
  - **前方/後方/完全一致、ワイルドカード**
- **キーワード検索**
  - **前方/後方/完全一致、ワイルドカード**
- **項目別検索**
  - **項目間**and/or、from:to
  - **前方/後方/完全一致、ワイルドカード**

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と同じ速さ

BTONICホスティング



インターネット第三世代

- **インターネット第一世代** 1980**年代～**
  - DNS、Mail、ftp、・・・
- **インターネット第二世代** 1993**年～**
  - http、Netscape/IE　←HTML
- **インターネット第三世代** 2001**年～**
  - B2B、SOAP　←XML
  - ＝Web**サービス**

Webサービスとは(2)

- **サーバが関数(機能)を持つ**
- **クライント(サーバ)がその機能を呼ぶ**
- **この呼び出し方法が**SOAP
  - Simple Object Access Protocol
- **中身のデータは**XML

.NETフレームワーク

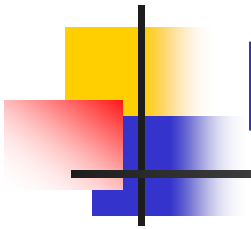
- **マイクロソフト社のサーバ製品名**
- Web**サービス**(SOAP)**対応機能を**OS(Windows2000)**に追加**
- Windows.net Server**からは標準添付**
- **開発環境は**VisualStudio.net
- Mobile Internet Toolkit**で携帯電話や**PDA**にも対応**

.NETフレームワークβ2

- β2は製品への組み込みが可能
- 10/22 PDCで新版を発表
- β1⇒β2改訂ドキュメント170ページ
  - 8年前のOLEよりマシ
- クライアント側はライブラリを提供

辞書サーバのSOAP連携

- 複数の辞書サーバで串刺し検索
- 親サーバでの一元表示
  - 子サーバからのデータ提供
  - 辞書コンテンツの一元管理が可能
- Office XPスマートタグ連携
  - 右クリックでの辞書引き
- http://www.est.co.jp/btonic



NetDictionary

- インターネット出版社アットマークITとデジタルアドバンテージ社の共同事業
  - イースト、三省堂が協力
  - http://www.webservice.jp/netdict/
- アットマークITの記事連携
  - 辞書を題材に、Webサービスの構築方法を解説
  - http://www.atmarkit.co.jp/fdotnet/netdict/index

辞書サーバ間インタフェース

- XMLでのデータ交換
- Unicode対応、外字対応

- **検索要求　⇒　単語、検索条件、検索数**
  - **候補一覧　←　検索結果リスト**
- **詳細要求　⇒**
  - **詳細項目　←　検索結果項目**

Neverending-Dictionary.com

- **世界中の辞書サイトをWebサービスで結合**
- **SOAPでのインタフェースを規定**
- **各国語の辞書引きが行える**
- **各国のサーバを各社が編集、更新**
- **常に成長する、インターネット辞書**

JTB.NET

- JTBの法人営業の業務すべてをWebサービス化したデモサイト
- http://net.est.co.jp/jtb/about/
- 国検索、空港検索、ホテル検索
- 社内稟議・決済システムとの結合

JTB.NETメソッド一覧

分類	メソッド名	機　能
共通	EnumCountries	国情報を列挙します。
	EnumAirports	空港を列挙します。
ホテル	EnumHotels	指定された都市／空港に関連するホテルを列挙します。
	FindHotel	予約可能なホテルを検索します。
	GetHotelInfo	ホテル情報を取得します。
航空便	EnumCarriers	航空会社を列挙します。
	FindFlight	予約可能な航空便を検索します。
	GetFlightInfo	航空便情報を取得します。
予約	OrderTour	ツアーの新規申し込みと予約内容の変更を行います。
	CancelTour	ツアーをキャンセルします。
確認	EnumTours	申し込まれたツアーを列挙します。
	GetTourContents	ツアーの予約状況を取得します。
(他)	NotifyState	クライアント側から予約に関する状況を通知します。

計画中のWebサービス

- 外字配信
  - 文字鏡番号⇒フォント・ビットマップ
  - JISコード⇒文字鏡URLフォント
- ISBN検索
  - ISBN⇒書名、出版社名、著者名
- 書誌情報配信
- NewsMLタグ追加

商品とサービス

- EXI＋生成ツール：製品化
- BTONIC：サービスとして提供

何をどのように売れば良いのか
サッパリ判らない！

お問い合わせは

- shimokawa@est.co.jp