NewsMLを活用したニュース検索Webサービスの構築

2002年6月13日

XMLコンソーシアム
応用技術部会 WebサービスWG

アジェンダ

- Webサービス WGの活動状況
- NewsMLを活用した
  ニュース検索Webサービスの構築
- Demonstration
- 所感(接続実験参加者)
- 今後の課題

応用技術部会 Webサービス WG

目的

- XML適用システムの開発を通し、XML技術の向上および普及に努める。
  - プロトタイプ開発を通した技術習得
  - XML利用上の課題の解決技術確立
  - XML製品の利用技術の習得

活動内容

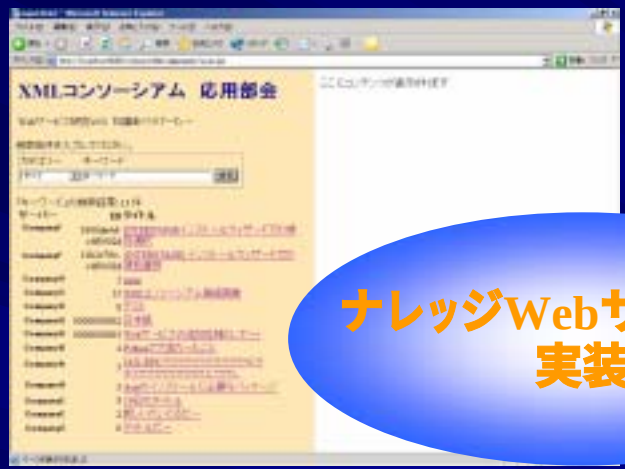
- XML基盤技術の評価を目的とした実証実験、プロトタイピングシステム開発
- 実用(アプリケーション)システムへの適用性評価を目的としたプロトタイピングシステム開発
- ベンダー各社が提供するXML関連プロダクトの評価

WebサービスWGでは
Webサービスを利用したプロトタイプ開発

1年間の活動状況(1)

活動内容

第1回 XMLコンソーシアムDay(11月22日)
第2回 XMLコンソーシアムDay(2月6日)



ナレッジWebサービスの実装

四則演算Webサービスの実装

本格的なWebサービスを実装

Webサービスを体感したい

WebサービスWG発足

2001/06　2002/01　2002/06　活動時期

ナレッジWebサービス実装の評価

- 評価
  - Webサービスの実装作業自体は有意義
    - アプリケーションサーバの使い方
    - インターフェースの定義/実装の習得
    - 本格的なWebサービスを実装した達成感
  - 実装されたナレッジWebサービスに面白みに欠けた
    - 提供サービス(コンテンツ)が少なかった
    - ビジネス要素が少なく、リアリティに欠ける
- ステップアップに向けた方針
  - 実ビジネス要素を加味した、リアルなWebサービスの実装
  - 提供サービスの充実化(異種サービスの実装)
  - 実装方法の更なる拡大(アプリケーションサーバ,開発言語)

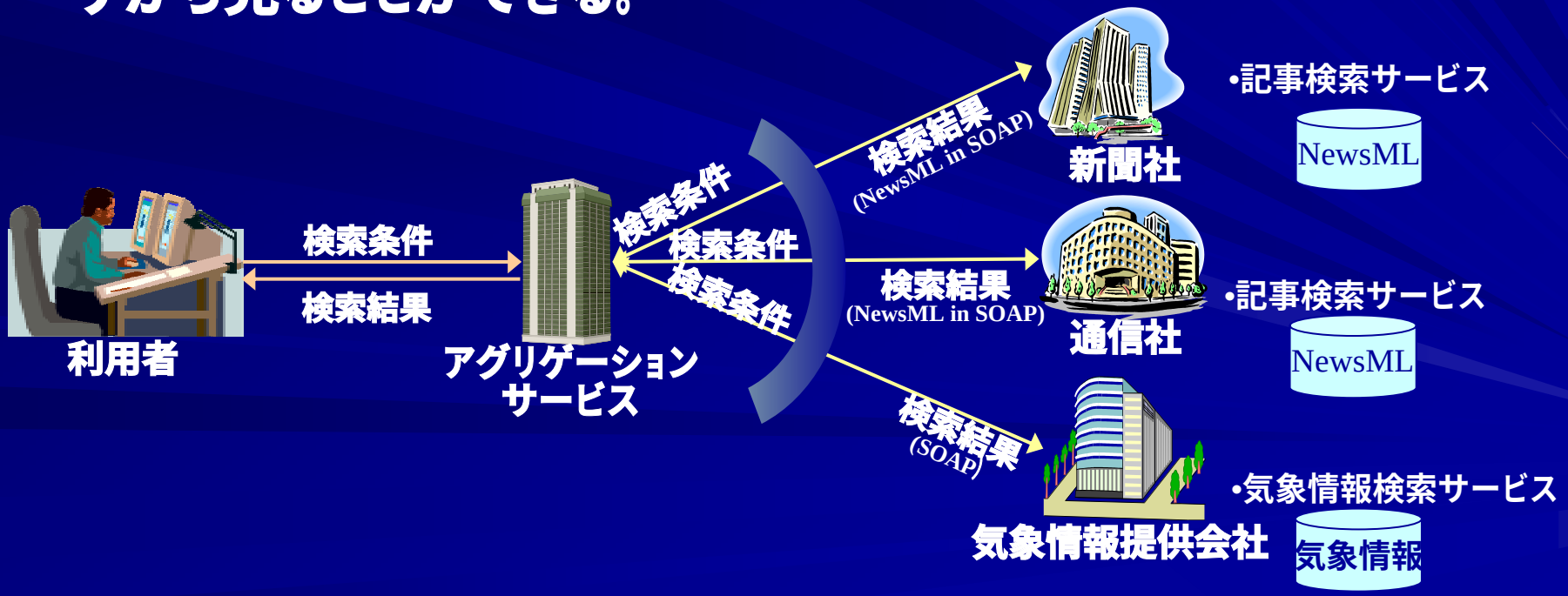
期待
- 充実したサービスの実装
- 実際に使われているコンテンツ

NewsMLを使ったWebサービスの実装

NewsMLを活用したニュース検索Webサービス

提供するWebサービス

- 複数の(仮想)新聞社/通信社/気象情報提供会社が提供する情報/コンテンツを１箇所のポータルサイトから検索
- 検索サービス利用者は、情報の所在(URL)を意識せずに単一のインタフェースで情報を取得することができる。
- 新聞記事は、NewsML形式で取り出せるだけではなく、スタイルシートにより、画像データを含んだHTML形式でWebブラウザから見ることができる。



記事検索方法

検索キーワードの指定

検索キーワード
- 分類
- 記事タイトル
- 記事本文
- 日時
- 画像データの有無

キーワードに該当する記事の一覧表示

- (仮想)新聞社/通信社からニュースコンテンツをWebサービスを使って集約

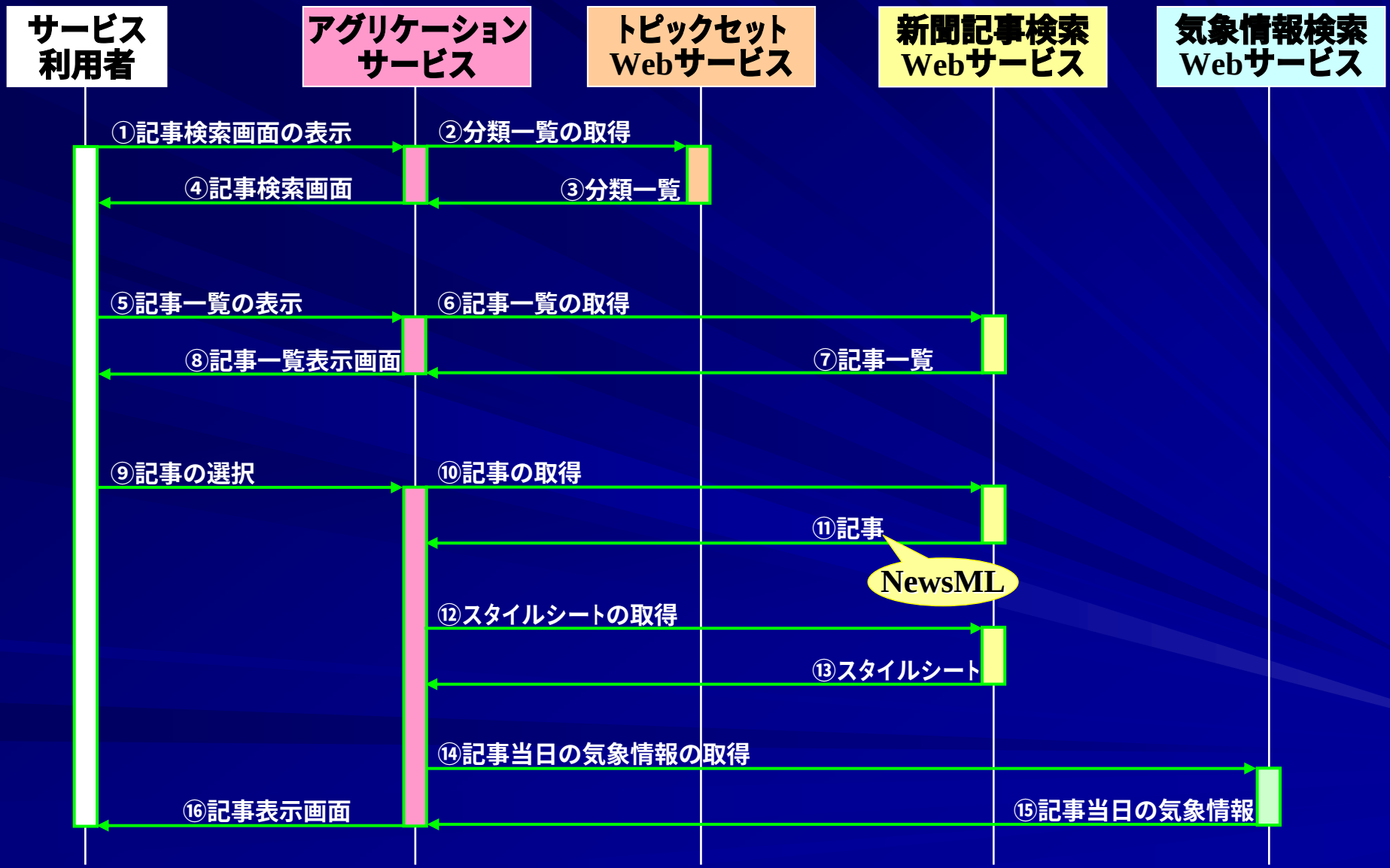
表示する記事の選択

- 選択された記事を(仮想)新聞社/通信社から取り出し
- 記事当日の気象情報を気象情報提供会社から取り出し

記事の表示

- NewsMLと気象情報をスタイルシートを使ってHTMLに変換

処理シーケンス



サービス/メッセージ一覧

トピックセット Webサービス　新聞記事検索 Webサービス　気象情報検索 Webサービス

サービス(service)	操作(operation)	メッセージ(message)		パラメタ(part)	
				型(type)	名前(name)
TopicSetService	getSubjectCodes	②	getSubjectCodes	ー	ー
		③	getSubjectCodesResponse	xsd:string	return
NewsMLService	search	⑥	search	xsd:string	subjectCode
				xsd:string	titleKeyword
				xsd:string	bodyKeyword
				xsd:boolean	hasImage
				xsd:string	dateFrom
				xsd:string	dateTo
		⑦	searchResponse	xsd:string	return
	getArticle	⑩	getArticle	xsd:string	publicIdentifier
		⑪	getArticleResponse	xsd:string	return
	getStylesheet	⑫	getStylesheet	xsd:string	publicIdentifier
		⑬	getStylesheetResponse	xsd:string	return
WeatherService	getWeather	⑭	getWeather	xsd:string	date
		⑮	getWeatherResponse	xsd:string	return

TopicSetサービス：TopicSetService

NewsMLで使用されるTopicSetに関するサービスを提供

- **SubjectCode：記事分類(政治、経済、スポーツ、芸能など)**
  - **大分類(TopicType@FormalName=“Subject”、17個)**
    - ◆FormaName要素の内容が”15000000”は“スポーツ”
  - **中分類(TopicType@FormalName=“SubjectMatter”、286個)**
    - ◆FormaName要素の内容が””15054000”は“サッカー”
  - **小分類(TopicType@FormalName=“SubjectDetail”、616個)**
    - ◆FormaName要素の内容が、“15073018”は”ワールドカップ”
  - **コードによる分類によって、言語に依存しない、分類検索が可能となる**
- **Status：記事の状態**
  - **Usable :公開可能**
  - **Embargoed :公開待機**
  - **Withheld :公開未定**
  - **Canceled :公開取消**
- **Country：国**
  - **JP：日本**
  - **KP：韓国**
- **：**

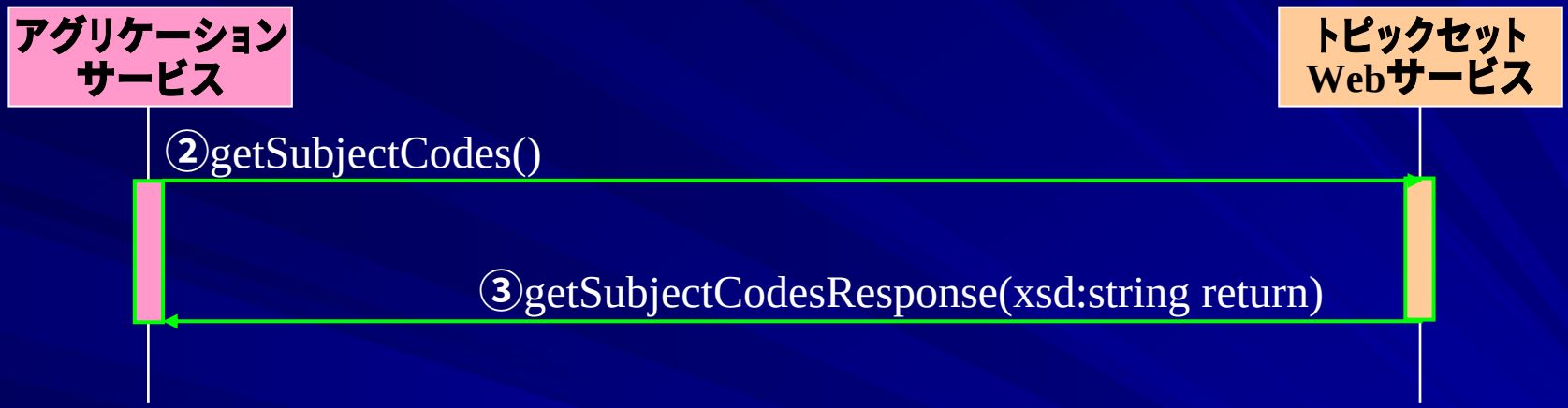
✓NewsMLを分類で検索するためには、SubjectCodeを使用する必要がある。

TopicSetに関するサービスが必要になる。

✓NewsMLの世界で共通なため、新聞社/通信社が提供する検索サービスとは、別サービス化

記事分類一覧の取得:getSubjectCodes

新聞記事の分類名と、分類名に対応するSubjectCodeの一覧を提供する操作



getSubjectCodesResponseで返すデータは、日本新聞協会NewsMLで公開されている用語セット(TopicSet)のSubjectCode(topicset.iptc-subjectcode-ja)から、Subject(大分類)のコードを返す。DTDは、NewsMLv1.0.dtdを利用(下記のDTD参照)。

*参考：Subject(大分類):17個,SubjectMatter(中分類):286個, SubjectDetail(小分類):616個

```
<!ELEMENT  NewsML      (Catalog? , TopicSet* ,  (NewsEnvelope , NewsItem+ ) )>
<!ELEMENT  NewsItem    (Comment* , Catalog? , Identification , NewsManagement ,  (NewsComponent | Update+ | TopicSet )? )>
<!ELEMENT  TopicSet    (Comment* , Catalog? , TopicSetRef* , Topic* )>
<!ATTLIST  TopicSet    Duid           CDATA  #IMPLIED
                       FormalName CDATA  #IMPLIED>
<!ELEMENT  Topic       (Comment* , Catalog? , TopicType+ , FormalName* , Description* , Property* )>
<!ATTLIST  Topic       Duid           CDATA  #IMPLIED
                       Details        CDATA  #IMPLIED >
```

getSubjectCodesResponseの戻り値（抜粋）

```
<NewsML>
    :
  <NewsItem>
      :
    <TopicSet Duid="IptcSubjectCodes.subject" Scheme="IptcTopicType" FormalName="SubjectCode">
      <Topic Duid="sr01000000">
        <TopicType Scheme="IptcTopicType" FormalName="Subject"/>
        <FormalName Scheme="IptcSubjectCodes">01000000</FormalName>
        <Description xml:lang="en" Variant="Name">Arts, Culture &amp;amp; Entertainment</Description>
        <Description xml:lang="en" Variant="Explanation">Matters pertaining to the advancement and refinement of the human mind, of
interests, skills, tastes and emotions</Description>
        <Description xml:lang="ja" Variant="Name">芸術、文化、娯楽</Description>
        <Description xml:lang="ja" Variant="Explanation">人間の精神や興味、技能、嗜好、感情の進歩や洗練に関係する事柄。
</Description>
      </Topic>
        :
      <Topic Duid="sr13000000">
        <TopicType Scheme="IptcTopicType" FormalName="Subject"/>
        <FormalName Scheme="IptcSubjectCodes">13000000</FormalName>
        <Description xml:lang="en" Variant="Name">Science &amp;amp; Technology</Description>
        <Description xml:lang="en" Variant="Explanation">All aspects pertaining to human understanding of nature and the physical world
and the development and application of this knowledge</Description>
        <Description xml:lang="ja" Variant="Name">科学、テクノロジー</Description>
        <Description xml:lang="ja" Variant="Explanation">人の自然や物理的世界に対する理解のあらゆる面、およびこの知識の発展や
応用</Description>
      </Topic>
        :
      <Topic Duid="sr15000000">
        <TopicType Scheme="IptcTopicType" FormalName="Subject"/>
        <FormalName Scheme="IptcSubjectCodes">15000000</FormalName>
        <Description xml:lang="en" Variant="Name">Sport</Description>
        <Description xml:lang="en" Variant="Explanation">Competitive exercise involving physical effort. Organisations and bodies
involved in these activities.</Description>
        <Description xml:lang="ja" Variant="Name">スポーツ</Description>
        <Description xml:lang="ja" Variant="Explanation">フィジカル努力を含む競争力があるエクササイズ。組織、および団体は、これら
の活動で含みました。</Description>
      </Topic>
        :
        :
```

NewsML検索時に使用

検索画面の選択候補に使用

記事検索サービス：NewsMLService

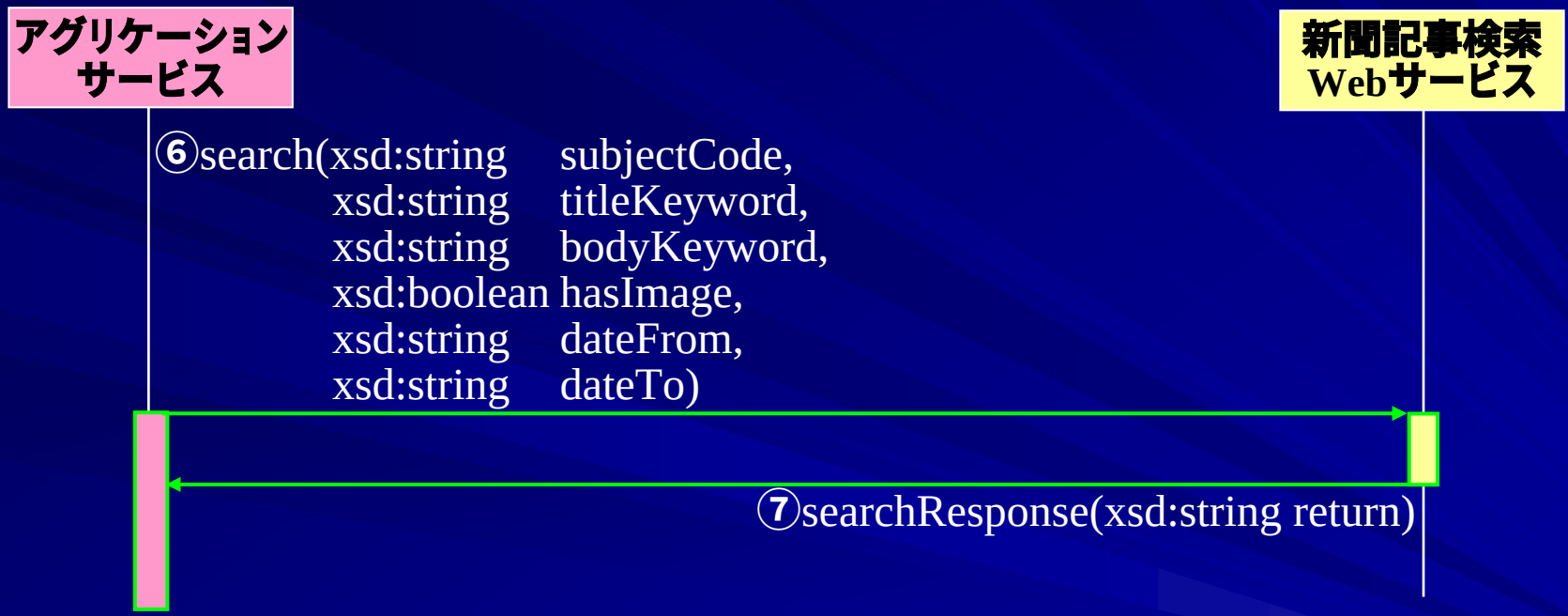
NewsMLの検索とその表示に関するサービスを提供

- **search**：検索条件に従って記事の一覧情報を検索するサービス
  - 検索キーワード
    - 分類　　：指定されたSubjectCodeで絞込み
    - タイトル　　：文字列の部分一致
    - 本文　　：文字列の部分一致
    - 画像データの有無　：画像データが存在するものだけを対象にする場合
    - 記事の年月日　　：開始年月日～終了年月日の範囲指定
- **getArticle**：指定されたPublicIdentifierのNewsMLを取り出すサービス
- **getStylesheet**：PublicIdentifierで指定されたNewsMLをHTML形式で表示するために必要なスタイルシート(XSLT)を取得するサービス

PublicIdentifier：NewsML個々に付与された世界でユニークなID

記事一覧検索：search

利用者から指定された検索条件に従って、該当する記事を検索する操作



記事一覧検索要求メッセージ：search

パラメタ名	パラメタ型	パラメタ説明	検索対象(NewsMLの要素)
subjectCode	xsd:string	subjectCodeで検索する	/NewsML/NewsItem/NewsComponent/DescriptiveMetadata/SubjectCode/Subject/@FormalName 文字列 完全一致
titleKeyword	xsd:string	タイトルで検索する	/NewsML//NewsComponent/NewsLines/HeadLine 文字列 部分一致のみ
bodyKeyword	xsd:string	本文で検索する	//DataContent の中 文字列 部分一致のみ
hasImage	xsd:boolean	イメージがあるものだけを検索する場合にtrue、イメージが無いものも検索対象にする場合はfalse	//ContentItem/MediaType を見て画像データが存在するか
dateFrom	xsd:string	記事の作成日時で検索する	/NewsML//NewsItem/NewsManagement/ThisRevisionCreated ・検索対象は年月日のみ ・開始年月日，終了年月日は「その日を含んで」
dateTo	xsd:string	記事の作成日時で検索する	

example

記事一覧検索応答メッセージ：searchResponse

パラメタ名	パラメタ型	内容
rerun	xsd:string	指定された検索条件に合致した記事の一覧を下記のデータ構造で返す

データ構造

```
<!ELEMENT articleInfo       (article*)>
<!ELEMENT article           (publicIdentifier, title, date, copyrightHolder, hasImage)>
<!ELEMENT publicIdentifier  (#PCDATA)>
<!ELEMENT title             (#PCDATA)>
<!ELEMENT date              (#PCDATA)>
<!ELEMENT copyrightHolder   (#PCDATA)>
<!ELEMENT hasImage          (#PCDATA)>
```

要素名	説明	内容(NewsMLの要素)
publicIdentifier	NewsMLの一意のID	/NewsML/NewsItem/Identification/NewsIdentifier/PublicIdentifier
title	記事のタイトル	/NewsML//NewsComponent/NewsLines/HeadLine
date	記事の作成日	/NewsML/NewsItem/NewsManagement/ThisRevisionCreated
copyrightHolder	著作権保持者	//NewsComponent/RightsMetadata/Copyright/CopyrightHolder
hasImage	画像データの有無 •true：画像データが有る場合 •false：画像データが無い場合	//ContentItem/MediaType を見て画像データが存在するか

記事取り出し：getArticle

指定されたPublicIdentifierのNewsMLを取り出す操作

記事取り出し要求メッセージ：getArticle

パラメタ名	パラメタ型	パラメタ説明
publicIdentifier	xsd:string	利用者が記事一覧で選択した記事のpublicIdentifier

記事取り出し応答メッセージ：getArticleResponse

パラメタ名	パラメタ型	内容	NewsMLの要素
rerun	xsd:string	指定されたpublicIdentifierのNewsML	/NewsML/NewsItem/Identification/NewsIdentifier/PublicIdentifier

スタイルシート取り出し：getStylesheet

記事(NewsML)の表示方法(スタイルシート)を取り出す操作

- NewsMLの記事本文(//NewsComponent/ContentItem/DataContent)のスキーマ構造は、NewsML毎に異なる。
- NewsComponentは、入れ子構造を持つことができる
- NewsMLをどのように表示するかは、著作権者の意志

- 記事(NewsML)を表示するためのスタイルシートを取り出すための手段が必要
- 今回のスタイルシートは、NewsMLの作成社が提供

スタイルシート取り出し要求メッセージ：getStylesheet

パラメタ名	パラメタ型	パラメタ説明
publicIdentifier	xsd:string	利用者が記事一覧で選択した記事のpublicIdentifier

スタイルシート取り出し応答メッセージ：getStylesheetResponse

パラメタ名	パラメタ型	内容
rerun	xsd:string	指定されたpublicIdentifierのNewsMLを表示(HTML)するためのスタイルシートのURL

- 指定されたPublicIdentifierのNewsMLを検索
- NewsMLの中の/NewsML/NewsEnvelope/NewsService/@FormalNameの値に従って、対応するNewsML用のスタイルシートのURLを返す

気象情報検索サービス：WeaherService

気象情報に関するサービスを提供

気象情報取り出し要求メッセージ：getWeather

パラメタ名	パラメタ型	パラメタ説明
date	xsd:string	気象情報を取り出す日付をyyyymmddの形式で指定

example

気象情報取り出し応答メッセージ：getWeatherResponse

パラメタ名	パラメタ型	内容
rerun	xsd:string	指定された日付の気象情報を下記のデータ構造で返す。 ただし、観測点は東京のみ

データ構造

```
<!ELEMENT weatherInfo    (date,weather,min,max)>
<!ELEMENT date           (#PCDATA)>
<!ELEMENT weather        (#PCDATA)>
<!ELEMENT min            (#PCDATA)>
<!ELEMENT max            (#PCDATA)>
```

要素名	説明
date	日付
weather	天気 晴れ、曇り、雨、 晴れのち曇り、曇りのち晴れ、晴れのち雨、雨のち晴れ、雨のち曇り、曇りのち雨、 晴れ一時曇り、曇り一時晴れ、晴れ一時雨、雨一時晴れ、雨一時曇り、曇り一時雨
min	最低気温(℃)
max	最高気温(℃)

example

コンテンツについて

今回の実証実験を行うにあたって、下記のコンテンツを利用

- 気象サービス様 気象情報
- 共同通信社様　NewsML、及びスタイルシート
- 毎日新聞社様　NewsML、及びスタイルシート
- 読売新聞社様　NewsML、電光用NewsML、及びスタイルシート
  (50音順)
- NewsML WG　JavaコンソーシアムNewsML、及びスタイルシート

共同通信社様のメディア用ニュース素材配信サービスNEWSPACKのコンテンツ(SOAP配信可能)

◆コンテンツの利用に際し、ご協力いただきました各社様に厚く御礼申し上げます。

◆NewsMLに関して技術的指導を賜りました応用技術部会

- 日本アイ・ビー・エム(株) 様
- 日本電気(株) 様
- 日本ユニシス(株) 様
- (株)読売新聞社 様
- その他NewsML-WGメンバー 様

ご協力ありがとうございました。

接続実験参加企業/アプリケーションサーバ

サービス		サーバ
アグリゲーション サーバー	RICOH SYSTEM KAIHATSU COMPANY, LTD. リコーシステム開発株式会社	BEA WebLogic Server 6.1J
トピックセット Webサービス	NISSIN 日進ソフトウエア(株)	Apache AXIS Beta1
共同通信社様 新聞記事検索 Webサービス	NISSIN 日進ソフトウエア(株)	Microsoft Visual Studio .NET
	HITACHI Inspire the Next (株)日立製作所	HITACHI Cosminexus Version 5
読売新聞社様 新聞記事検索 Webサービス	FUJITSU 富士通(株)	FUJITSU INTERSTAGE V4.0L20
	TOSHIBA (株)東芝	IBM WebSphere Application Server4.0
	ORACLE 日本オラクル(株)	Oracle9i Application Server Release2
	NRI 野村総合研究所	
	NEC 日本電気(株)	NEC ActiveGlobe WebOTX Ver4.2
	PFU a Fujitsu company PFUアクティブラボ (株)	

接続実験のシステム構成

利用者
- Webブラウザ

アグゲーションサーバー
- WebLogic Server 6.1J
- リコーシステム開発

トピックセット Webサービス
- Apache AXIS Beta1
- 日進ソフトウエア

共同通信社様新聞記事検索 Webサービス
- Visual Studio .NET
- 日進ソフトウエア

毎日新聞社様新聞記事検索 Webサービス
- Cosminexus Version 5
- 日立製作所

読売新聞社様新聞記事検索 Webサービス
- INTERSTAGE V4.0L20
- 富士通

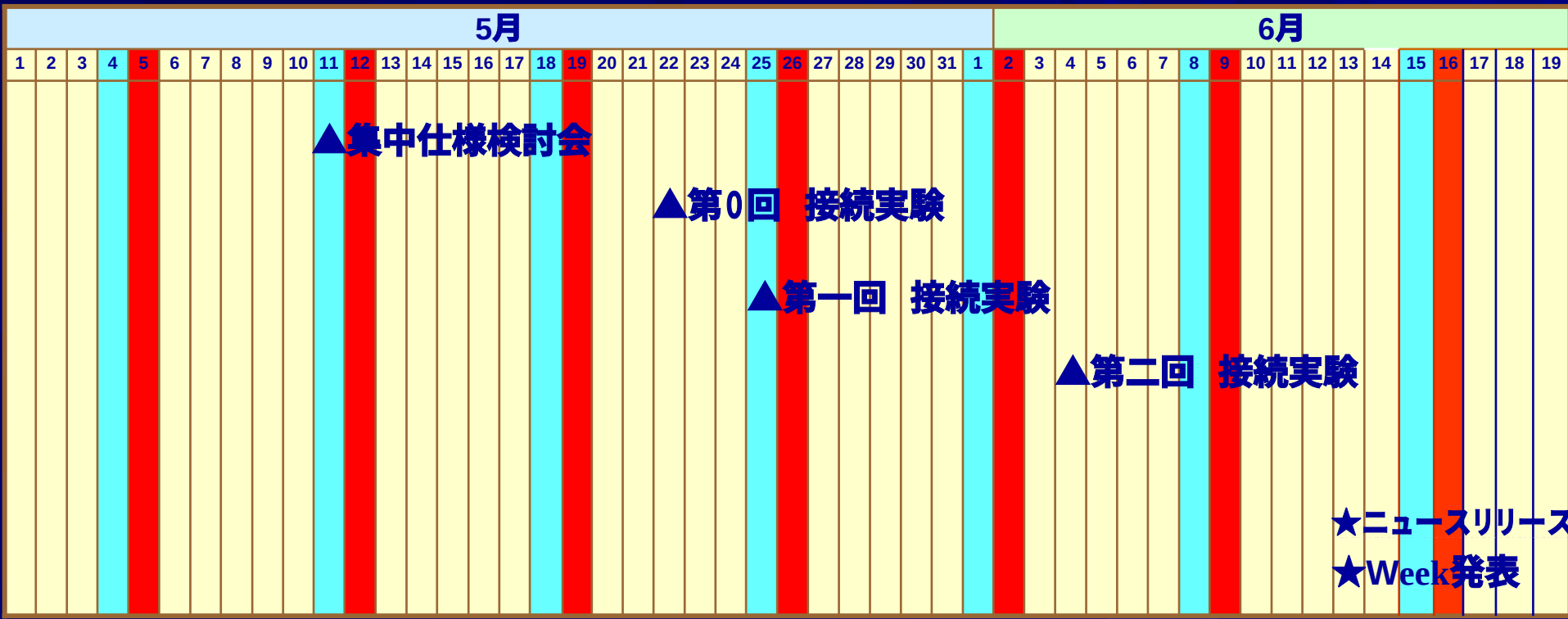
読売新聞社様電光記事検索 Webサービス
- WebSphere Application Server4.0
- 東芝

読売新聞社様電光記事検索 Webサービス
- Oracle9i Application Server Release2
- 日本オラクル

Javaコンソーシアム記事事検索 Webサービス
- Orbix E2A XMLBus Edition 5.1
- 野村総合研究所

気象情報検索 Webサービス
- ActiveGlobe WebOTX
- 日本電気

開発/接続実験スケジュール



実装方法に依存しないWebサービス

- アプリケーションサーバ
  - 9種類のアプリケーションサーバを利用
- 開発言語
  - Java
  - C#
- XMLデータの格納方法
  - RDB(カラムマッピング、バイナリ格納)
  - XML専用データベース
  - フラットファイル
- XMLデータの検索方法
  - XPath
  - 文書管理システム
  - 全文検索エンジン
  - DOMで独自処理

Demonstration

by リコーシステム開発) 八尋

接続実験の様子

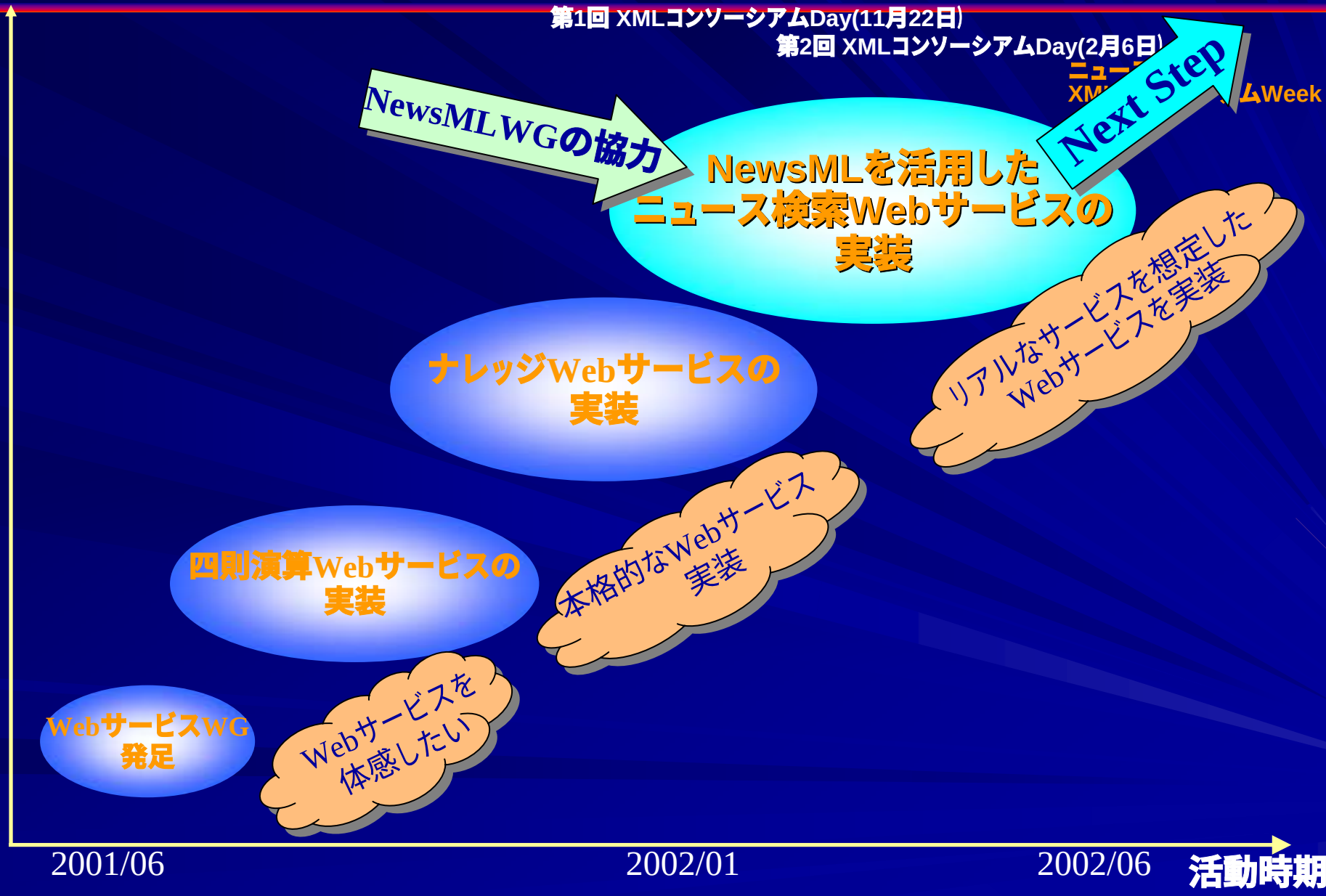
2002年6月4日 日進ソフトウエア様社内に於いて

所感

- リコーシステム開発) 八尋
  - アグリゲーションサービス
- 日進ソフトウエア) 定村
  - TopicSet　Webサービス
- 日進ソフトウエア) 荒本
  - 共同通信社様　新聞記事検索Webサービス
- 富士通) 山本
  - 読売新聞社様　新聞記事検索Webサービス
- 日立製作所) 葛坂
  - 毎日新聞社様　新聞記事検索Webサービス
- 東芝) 山田
  - 読売新聞社様　電光記事検索Webサービス
- 日本オラクル) 鈴木
  - 読売新聞社様　電光記事検索Webサービス
- 野村総合研究所) 坂田
  - Javaコンソーシアム記事検索サービス
- NEC)高橋
  - 気象情報検索Webサービス
- PFUアクティブラボ)松山
  - 取り纏め

(敬称略)

1年間の活動状況(2)



活動評価

もう一度振り返って

- 目的
  - プロトタイプ開発を通した技術習得 — 達成
    - 算術演算Webサービス
      - Webサービスを体験
    - ナレッジWebサービス
      - 複数のWebサービスを連携
    - ニュース検索Webサービス
      - 実世界で使われているコンテンツを利用

StepUp

  - XML利用上の課題の解決技術確立 — ほぼ達成
    - Webサービスを使ったシステム開発手順の確立
      - ユースケース
      - 外部インターフェース(WSDLなど)
      - サービス実装
      - 接続
  - XML製品の利用技術の習得 — 達成
    - 計11種類のアプリケーションサーバを使用
    - XMLストレージ
      - RDB(カラムマッピング、バイナリ化格納)
      - XMLデータベース

今後の課題

- 技術的には
  - アグリゲーション型(Pull型)以外のWebサービス
    - 配信型(Push型)
    - 非同期型
    - コンテンツからサービスへ
  - トランザクション処理
    - コミットとロールバック (課金処理など)
    - 動的フロー制御(動的なWebサービスの利用)
  - UDDI
    - ローカル(企業内、業界内、地域)
    - グローバル
  - サービスレベル
    - 性能：スピード
    - 品質：信頼性
    - セキュリティ (暗号化、外部攻撃)
- 方法論として
  - XMLコンソーシアムが提供するWebサービス
  - 他WG、部会、Webサービス推進委員会との連携
    - ✓NewsML WGとの連携が成功！

Webサービスの普及の一助となれば幸い

ご清聴ありがとうございました。

END