2005.11.5

平成 17 年度情報処理軽井沢セミナー報告書

愛媛大学　総合情報メディアセンター
木村　映善

「OAI-PMH サーバの実装と RSS アグリゲーション及びハーベストの提供の試み」

背景

愛媛大学総合情報メディアセンターは学内の情報基盤の統括、学生の情報リテラシーの教育、高度情報利用技術を活用しての学術研究支援、そして e-Learning による大学の教育システムの整備といった役割を担っている。学術研究支援と e-Learning の取り組みを通して、学術情報、研究情報の効率的な蓄積、管理のノウハウを模索してきた。その中で、OAI-PMH というメタデータをハーベストする手法の存在を知り、その応用と適用の仕方について同研修において学んだ。当初、国立情報学研究所メタデータ・データベース共同構築事業での「OAI-PMH の NII メタデータ・データベースへの適用について」等の先進事例について学び、それに倣って試験的なシステムを構築することを考えていた。

先述したように総合情報メディアセンターは書誌情報以外にも、e-Learning や本メディアセンターが支援している、太陽地球系物理観測データの管理に代表される観測データなど、様々なデータを扱っている。現在認識している問題点として、図書情報以外の幅広いメタデータを配信、管理する手法が確立されていないということから出発している。

そこで、OAI-PMH をサポートしている、DSpace に代表される実装について調べたところ、（１）OAI_DC 以外のメタデータをサポート、（２)大元のデータとの連携（e-Learning のコンテンツや観測情報等へのアクセス）を同時に満たしている実装は見かけなかった。そこで、本研修の前半の座学で得た情報を元に、OAI-PMH を利用した実装のプロトタイプの開発を試みた。

OAI-PMH について

OAI-PMH はメタデータを刈り取る(Harvest)するための標準的手法であり、HTTP/REST アーキテクチャに基づいたシンプルなモデルである。標準で、書誌的情報のメタ情報のスキーマとして OAI_DC を必須サポートとしている。さらに、そのほかの様々な（独自に定義したものを含めて）メタデータのスキーマについても、クエリ要求時に metadataPrefix の指定をすることで、リポジトリがその指定スキーマをサポートしている場合は、クエリ結果を指定スキーマの形で返すことが出来る。このことによって、OAI-PMH は(1)ハーベスタはシンプルな実装でメタデータを刈り取れる、（２)様々な形式のメタデー

タを扱える汎用性を持つ、といった利点を提供している。

OAI_DC の問題と、独自スキーマの開発について

OAI-PMH では最小限にサポートすべきメタデータとして、OAI_DC が規定されている。しかし、OAI_DC は DC 要素のうち任意をサポートすればよいと定義されており、全ての DC 要素を吐き出すとは限らない。この点において、データリポジトリとハーベスタの相互の期待するデータの粒度のミスマッチがおこりうる。

そして、そもそも DC 要素のみではメタデータとしては役不足の感がある。現に、国立情報学研究所の事例としては junii 形式メタデータの OAI-PMH 上の実装を行っている。

そのため、観測データ、教育コンテンツの為に独自のメタデータ形式を作成し、配信することを検討した。

演習の目標

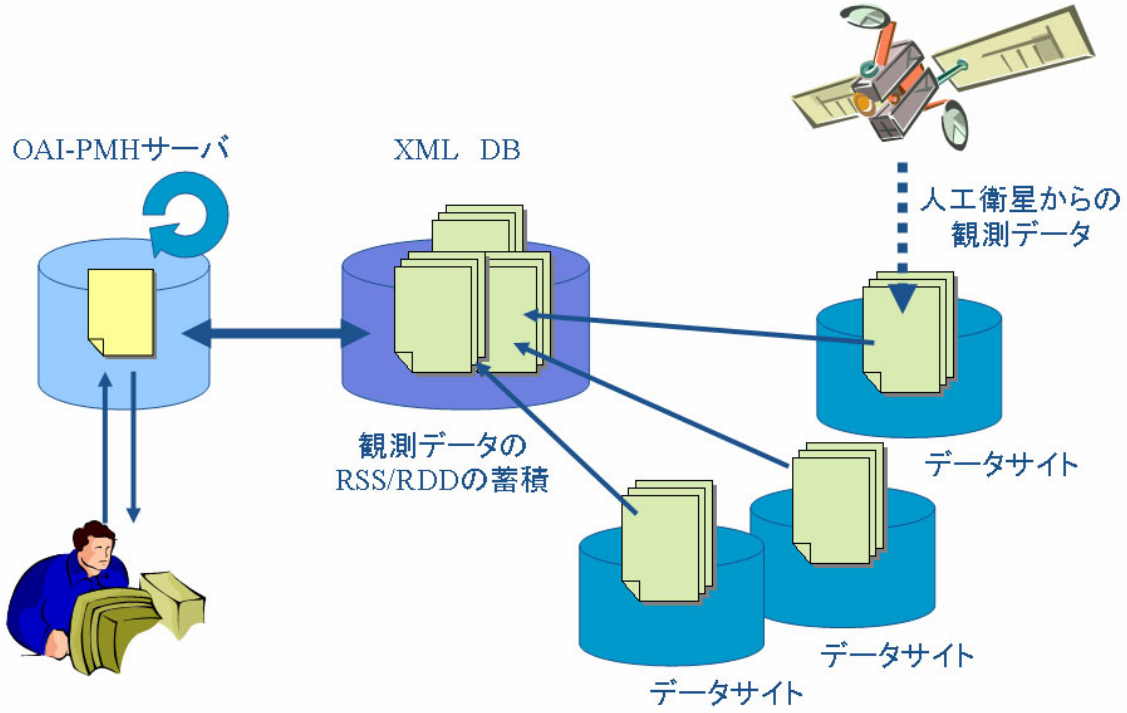
以上の背景と考察を踏まえて、研修の目標を下記のように設定した。

（１)独自のメタデータの配信の試み

OAI_DC 以外の具体的なデータとして、総合情報メディアセンターが研究支援活動の一環で開発している人工衛星観測データの検索サイト（STARS)で利用されている RSS/RDF ベースのメタデータを扱う。

（２)XML データベースとの接続

OAI-PMH ではハーベスタとのインタフェースのみ規定しており、リポジトリの実装については様々なものがある。学術情報、研究データは今後 XML データベースに蓄積していく計画もあるので、XML データベースと連動させることを試みる。

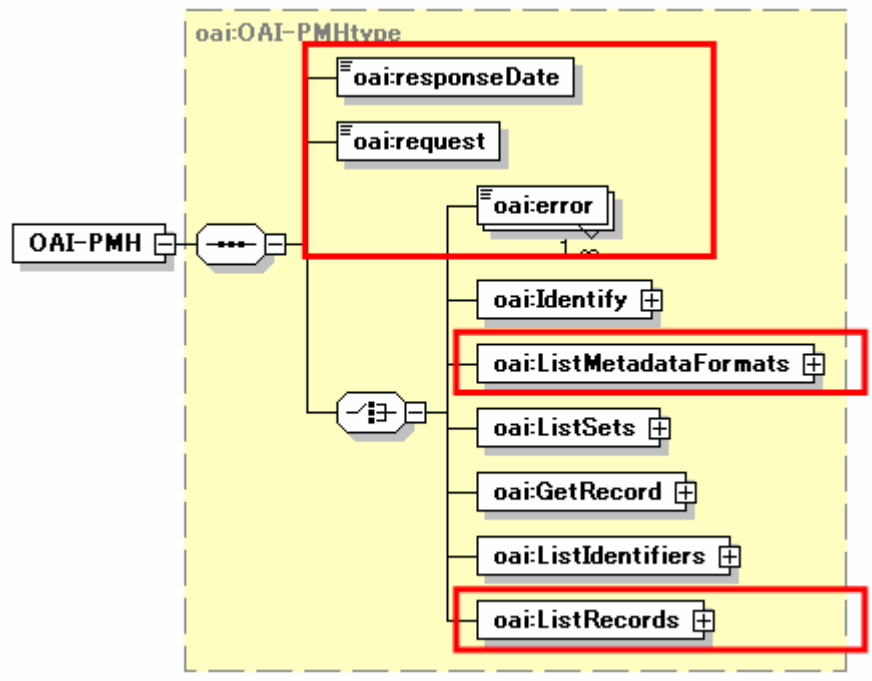


具体的には上図に示したように、人工衛星から観測されたデータがデータサイトにあり、それらの観測データの更新情報が RSS/RDF で愛媛大学に提供されている。そこで、この RSS 情報に埋め込まれた RDF/XML を今回の実習の為に構築した XML データベースに格納し、OAI-PMH の要求に合わせて適宜 XML によるクエリ、加工を行い、結果を返すような OAI-PMH サーバのプロトタイプ作成を行う。

開発環境：

武蔵野嵐山での研修会場に持ち込んだノートパソコン
WindowsXP Professional
IIS6.0 + ASP.NET 2.0
SQL Server 2005
.NET Framework 2.0
C#,XQuery/XPath,XPathNavigator/XPath Document(Visual Studio 2005)
Altova XML Spy 2005

演習第１日目：



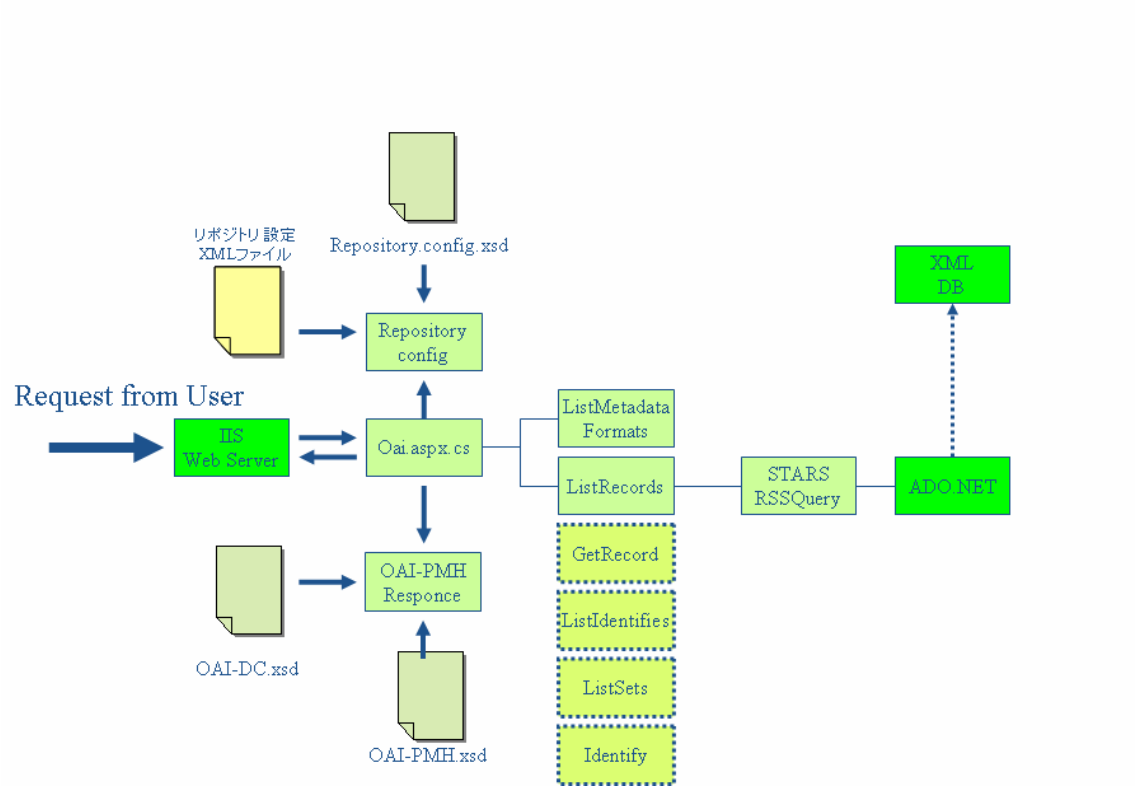
1.OAI-PMH サーバの実装の選択

講習会で学んだ OAI-PMH の概要をもとに、講師の助言を受けながら XML DB での OAI-PMH サーバの実装を調査した。DSpace、ePrint では要望する事項を満たせず、イン

ターネットから OAI-PMH に関する実装を調査した。その結果、特定の用途に向けた OAI-PMH の実装は様々に紹介されている（http://www.openarchives.org/tools/tools.html）が、受講者の希望する用途に沿った形での利用を想定しているものはなかった。そのため、OAI-PMH サーバのプロトタイプを作ることを試みた。OAI-PMH サーバは Web アプリケーションとして実装できるため、cgi や Java サーブレットなどの開発経験があればプロトタイピングレベルでの実装は比較的容易であると思われた。そこで、IIS+ASP.NET2.0 上での実装を試みることにした。

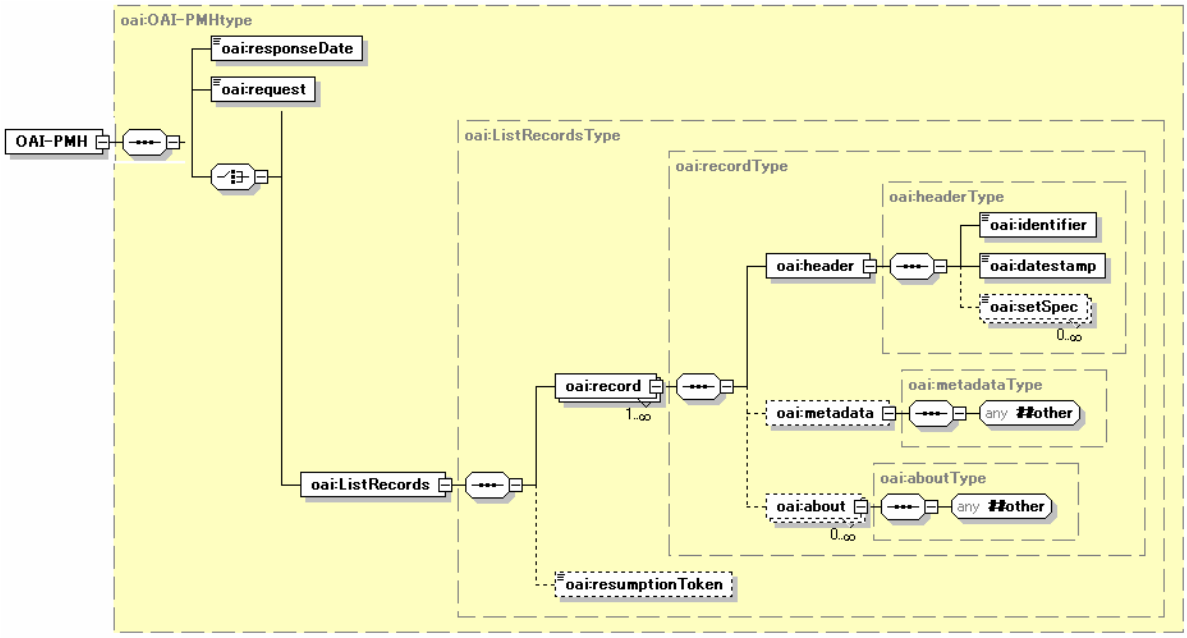
Web サーバへの実装について

OAI-PMH に関する資料は国立情報学研究所によって日本語の資料が整備されている（http://www.nii.ac.jp/metadata/oai-pmh2.0/）。
「Open Archives Initiative Protocol for Metadata Harvesting」において OAI-PMH の要求、応答、エラー処理についての規定がなされている。この要求、応答に関する規定を実装するためのガイドラインを XML Schema 形式で定義しているのが、3.2.1 XML Schema for Validating Responses to OAI-PMH Requests 節で参照されている OAI-PMH.xsd スキーマである。
(http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd)
そして、その最小メタデータセットとして定義されているのが、OAI_DC（http://www.openarchives.org/OAI/2.0/oai_dc.xsd）である。
そこで、OAI-PMH の 6 種類の verb をパラメータとして指定した HTTP GET リクエスト要求に対処し、上記 XML スキーマに従って応答を構築していけば、OAI-PMH の要求を返す OAI-PMH サーバが出来ることになる。
　以上の事柄をまとめ、下図に示すように設計を行った。
ユーザーから IIS Web Server は HTTP GET 要求を受け取り、自作 Web アプリケーションである Oai.aspx において、verb のパラメータを抽出する。そして、それぞれの verb に応じてその verb に対応する処理をするルーチンを呼び出し、その結果を OAI-PMH の XML スキーマに埋めこんで、XML 応答として返す。時間的な制約のため、6 種類の verb のうち、ListMetadataFormats/ListRecords/Identity を実装することとした。このうち、XML データベースと実際に通信し、応答を返すのは ListMetadata の処理部分である。
　1 日目はこのうち、HTTP GET 要求の verb ごとのディスパッチ処理と、OAI_PMH の XML スキーマに沿ったファイルを生成するクラスの生成を行った。



演習第 2 日・3 日目（徹夜を含む）

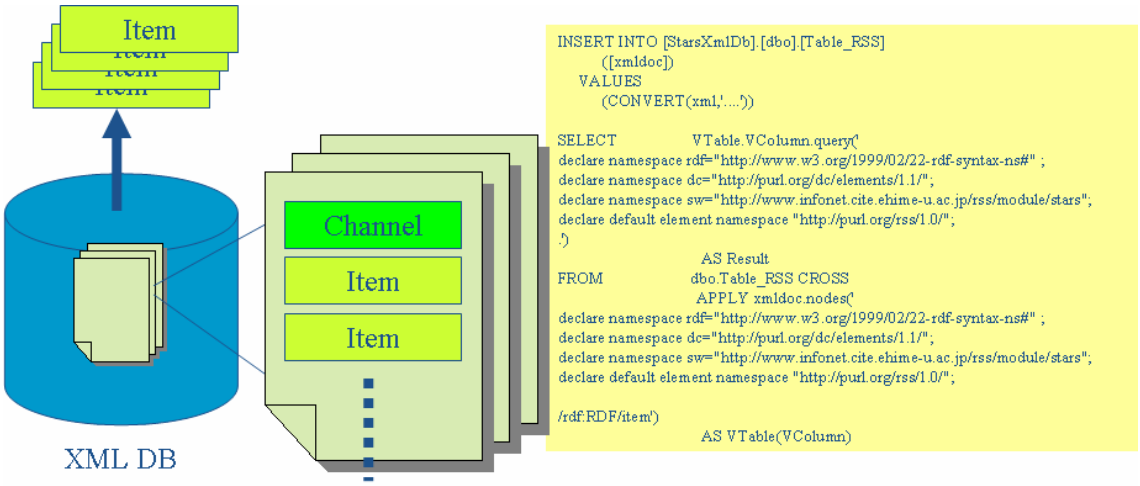
1 日目で OAI-PMH を処理する大枠が完成したので、2 日目は、OAI-PMH のスキーマにそって処理する実装に入った。



上図の通り、OAI-PMH のスキーマは、oai:resonseDate,oai:request そして、それぞれの verb ごとの内容が含まれるが、今回は特に oai:ListRecords を選択した。oia:responseDate

には要求時のシステムの時刻を代入し、oai:ListRecord 以下には oai:record と oai:resumptionToken を入れた。oai:resumptionToken は多くのレコードがあり、一回の要求では応答しきれないときに分割応答をするための区切りとして返すものであるが、プロトタイプの為に割愛した。oai:record 以下には oai:header/oai:metadata があるが、この oai:metadata そのものが、OAI-PMH の metadataPrefix で指定されたメタデータ形式で返される XML 形式のメタデータである。

このメタデータとして、XML データベースに格納された RSS ファイルからの抽出結果を格納するようにした。

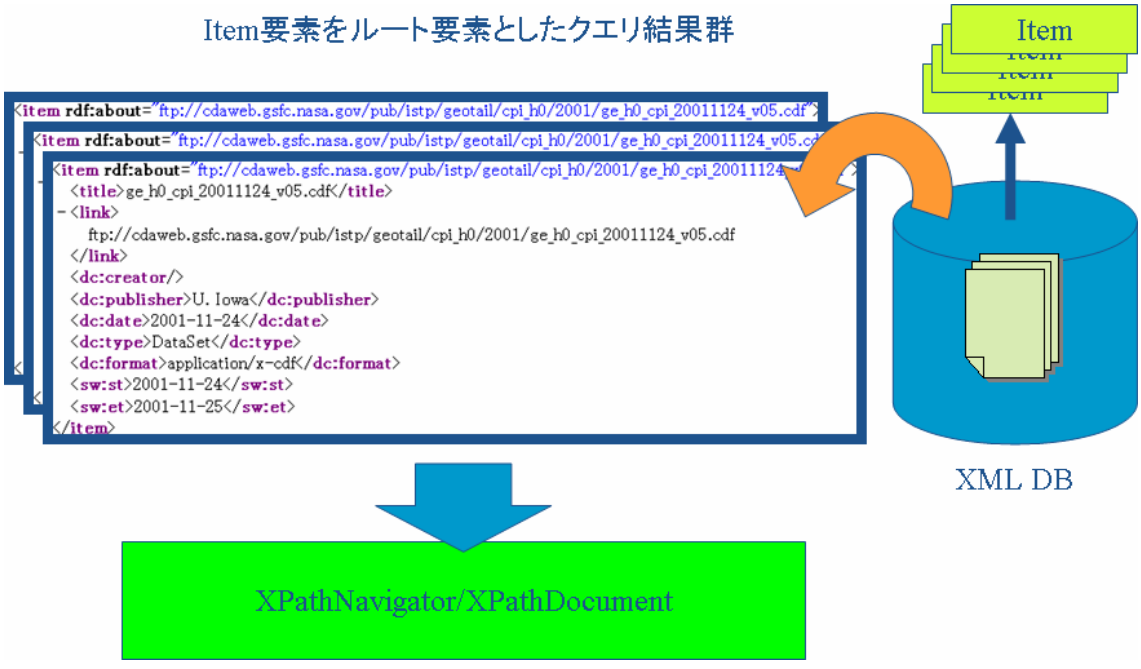


上図に示すように、XML データベースには、複数の RSS フィードの結果がそれぞれの XML 形式文書として格納されている。それぞれの RSS には、最新のコンテンツのメタ情報が RSS の Item 要素として記載されている。そこで、RSS として記述された XML 文書群から、OAI-PMH 要求に対する応答として RSS の Item 要素をメタデータとして返すための Query を構築した。

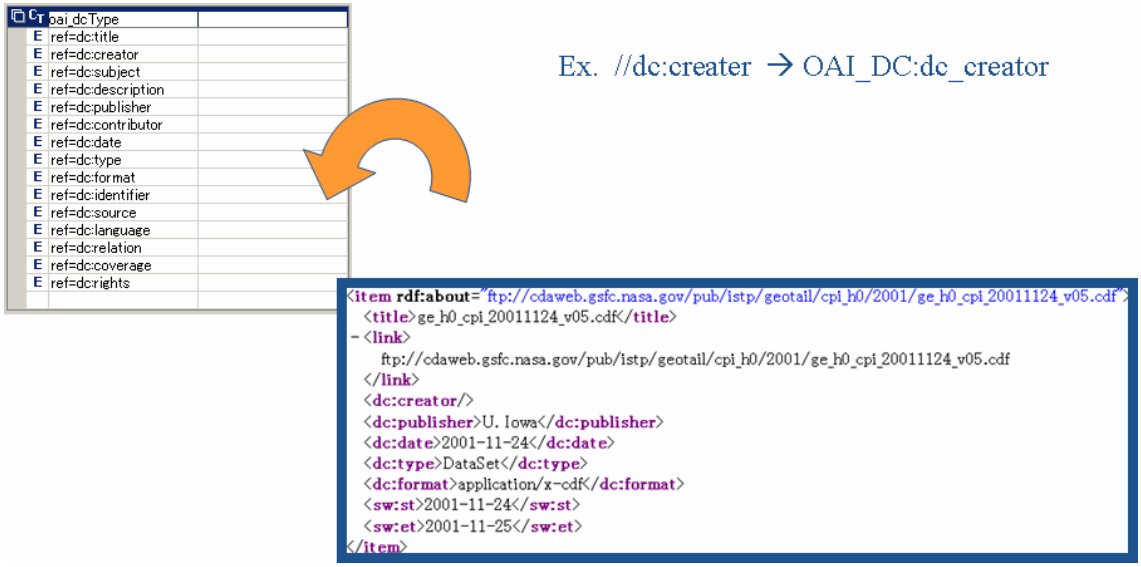
ここで、SQL 様式のクエリーを XML データベースに対して行うときに名前空間を適切に指定してクエリしないと、期待する応答が得られないことに気づかず、クエリ構築に半日をロスした。 まず、RSS 文書から Item 要素部分のみを引き出すが、この時点では、Item 要素群が文書ごとに連結した結果となって帰ってくる（下図）。

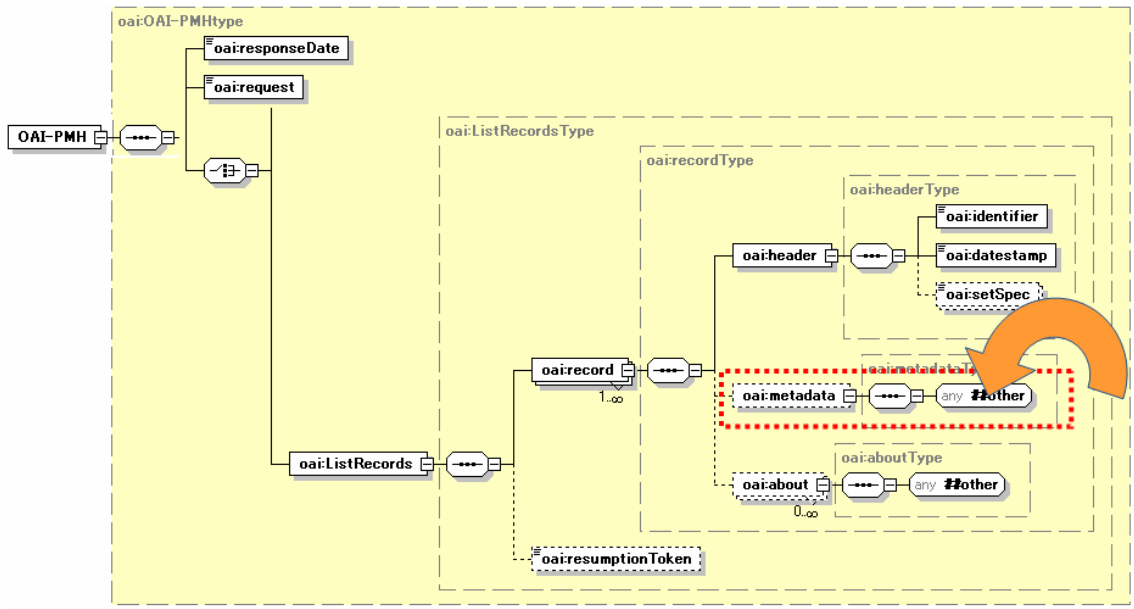
```
- <rdf:RDF>
  + <channel rdf:about="http://www.infonet.cite.ehime-u.ac.jp/STARS/rss/"></channel>
  - <item rdf:about="ftp://cdaweb.gsfc.nasa.gov/pub/istp/geotail/cpi_h0/2001/ge_h0_cpi_20011124_v05.cdf">
      <title>ge_h0_cpi_20011124_v05.cdf</title>
    - <link>
        ftp://cdaweb.gsfc.nasa.gov/pub/istp/geotail/cpi_h0/2001/ge_h0_cpi_20011124_v05.cdf
      </link>
      <dc:creator/>
      <dc:publisher>U. Iowa</dc:publisher>
      <dc:date>2001-11-24</dc:date>
      <dc:type>DataSet</dc:type>
      <dc:format>application/x-cdf</dc:format>
      <sw:st>2001-11-24</sw:st>
      <sw:et>2001-11-25</sw:et>
    </item>
  - <item rdf:about="ftp://cdaweb.gsfc.nasa.gov/pub/istp/geotail/cpi_h0/2001/ge_h0_cpi_20011125_v05.cdf">
      <title>ge_h0_cpi_20011125_v05.cdf</title>
    - <link>
        ftp://cdaweb.gsfc.nasa.gov/pub/istp/geotail/cpi_h0/2001/ge_h0_cpi_20011125_v05.cdf
      </link>
      <dc:creator/>
      <dc:publisher>U. Iowa</dc:publisher>
      <dc:date>2001-11-25</dc:date>
      <dc:type>DataSet</dc:type>
      <dc:format>application/x-cdf</dc:format>
      <sw:st>2001-11-25</sw:st>
      <sw:et>2001-11-26</sw:et>
    </item>
  </rdf:RDF>
```

そのため、それぞれの文書ごとの Item 要素の集合体のクエリから、さらに Item 要素をルート要素とした集合体として得るために再帰的にクエリをかけた。こうして得られた結果はそれぞれの Item 要素をルート要素とした集合体としてアクセスできることになる。



そこから、OAI-PMH に要求されている最小セットとしての OAI_DC 形式にマッピングしていくことになる。当初 XSLT での変換を試みたが、名前空間などの問題もあり、うまく行かずに挫折し、プログラムでの逐一の変換を試みた。.NET 2.0 Framework で開発された新しい XML ドキュメントへのアクセス手法として、XPath によるクエリが出来る XPathNavigator を使い、特定のノードを切り出し、それを XML スキーマに従って当てはめていった。

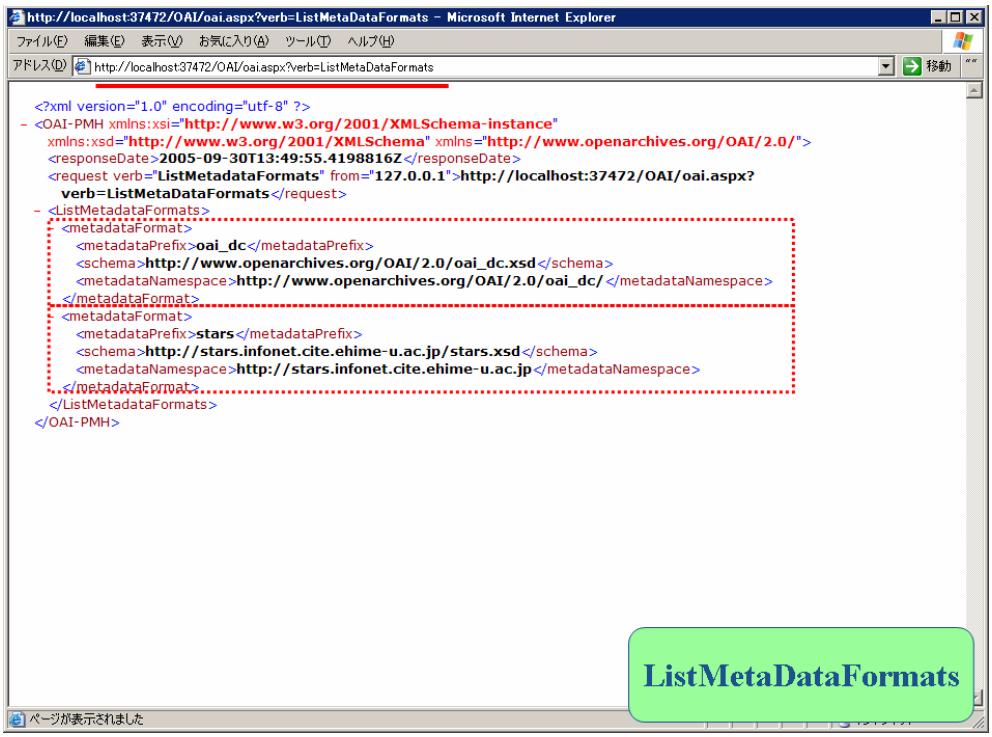




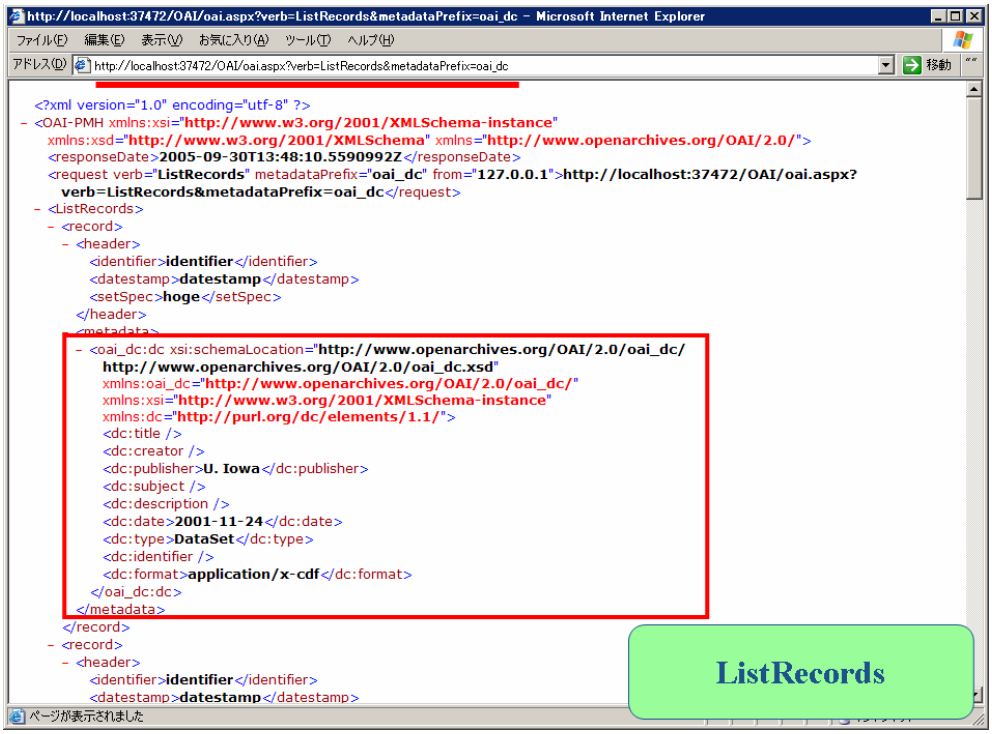
以上のように RSS の Item 要素のうち、dc:creater --> OAI_DC:dc_creator というように順次当てはめていき、そしてその結果を OAI-PMH スキーマの oai:metadata の中に埋め込み、その処理を RSS の Item 要素の数の分だけ oai:record 要素を繰り返す形で応答 XML 文書を作成していく。

このようにして、作成した結果の画面を以下に転載する。

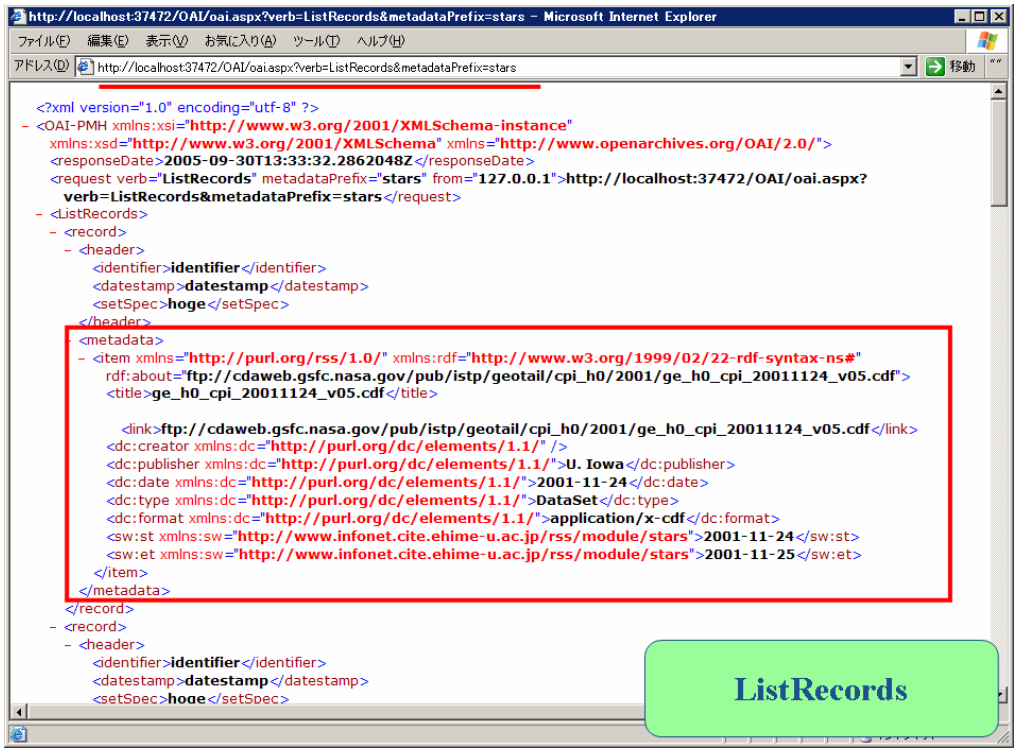
ListMetaDataFormat 要求に対する応答



ListRecord に対する応答(oai_dc 形式での応答)



ListRecord に対する応答（stars_rss 形式での応答）



まとめ

3 日間の実習を通して、OAI-PMH サーバの部分的な実装を行い、RSS に埋め込まれたメタデータを OAI-PMH によるハーベスティングが出来ることを確認した。本来、RSS で完結するものをわざわざ OAI-PMH で集める必要はないが、データリポジトリに動的にデータが蓄積され、かつ XML データベースから検索しやすい題材として RSS を利用したのであり、今後の XML 様式のメタデータの蓄積、検索、収集といった試みの足がかりを作れたといえる。OAI-PMH の XML Schema を基に作成していったので、基本的な構文は満たしており、XML　DB からクエリした結果を OAI-PMH のハーベスト要求に対して、OAI_DC、STARS の 2 種類の metadataPrefix で提供できることを確認した。

様々な教育コンテンツや計測データのメタデータを XML DB で管理し、OAI-PMH で配信する事への足がかりを得ることが出来た

今回の研修で得た OAI-PMH に関する見識を分かりやすく紹介、啓蒙していくとともに、具体的な OAI-PMH で使うメタデータと利用モデルを考えていく必要があると考えている。

研修で学んだ技術及び知識

・OAI-PMH の基本的知識及びメタデータ、リポジトリに対する理解

・メタデータの収集と加工に対する基本的な理解
・GETA（汎用連想検索エンジン）についての概要

事前準備として必要と思われるもの

・UTF-8 を使えるエディタ、XML エディタ
　XML を扱うときに UTF-8 に対応していないと正しく編集できなかったり、ゴミが入ってしまうことがある。また XML を編集する機会が多いので、XML 文法を認識できるエディタやツールがあると望ましい。

・最低一つの開発言語に慣れておくこと
　３日間という非常に短い期間なので、開発に慣れていないと言語習得だけで終わってしまう可能性もある。自身が自由自在に使えるという開発言語を身につけておいたほうが良い。

・XSLT や XPath などに対して慣れておくこと。
　実はこれができていなかったために、メタデータを取って加工するということが出来ず、３日のうち２日はこれで潰れてしまった。特に名前空間が絡むと、ややこしくなる。XML に関する基本的理解以上に名前空間の理解や、それに伴う XSLT や XPath の扱い方に慣れておくこと。これが出来ていないとかなりのタイムロスがおきると思う。

・XML、開発書籍関連の書籍
　今は Google 巡りで情報のつまみ食いが出来るが、実習に入ったときには受講者が一斉にアクセスしたせいもあってか、かなりネットワークが遅いことがあった。また Google についても情報が玉石混交なので、書籍を２～３冊持っていくのが良いだろう。事前に目を通して、携行していきたい書籍を見定めたほうが良いと思われる。前半の座学のときに、XSLT や XPath の知識がないと実習がおぼつかないことに気がつき、座学が終わって実習に入る前の時間に東京の書店に立ち寄って購入した。かなり準備しておかないとつらいと思います。私は最後の２日間は徹夜で自室に篭って取り組んでいました。

今後の課題,計画

　OAI-PMH についての基本的理解とプロトタイプ実装が出来たことが有意義であったが、実際に動かしてみて、やはりメタデータの定義と処理、扱いについての考察を続けていく必要があると感じた。OAI-PMH はセミナー受講以前では抽象的な概念が多く、理解が難しかったのであるが、こうして実際に動かしながらやってみると、意外とシンプルであるこ

とに気づかされた。そして問題は OAI-PMH そのものの習得が壁なのではなく、その先にある、メタデータの定義、運用にある、ということを改めて気づかされた。

今回は RSS を題材として XML データベースを構築したが、愛媛大学学内で SQL Server を利用して XML データベースを構築し、2006 年の夏頃の公開を目指していくつもりである。

演習の感想

受講以前は OAI-PMH の文書や Web サイトを読んでも、抽象的なことが多く、要点を理解できなかったのであるが、前半の座学、そして後半の実習を通して OAI-PMH について理解し、プロトタイプまで進めたことに驚きとセミナー参加の意義を感じている。

やはり、独学でやるよりは遥かに吸収の効率が良いことを感じた。

一方、セミナーの期間が短く、それぞれの課題設定と、「動くもの」を要求されるために、相当の準備と当日の努力が要求され、今まで受けたセミナーの中では要求水準が高いものと感じた。決して受動的な態度では完遂できないセミナーであると思う。

逆に言えば、事前の十分な準備が出来ていれば、それだけジャンプできる機会が与えられるセミナーであると思う。

軽井沢ではなく、武蔵野嵐山になったのは残念であるが、いざ、受講してみると外に出てくつろぐ暇もなく、外の環境は実際どうでもよくなるぐらい、集中して作業し続けなければならない状態に追い込まれた。

GETA については時間的制約から取り組まなかったが、解散後の帰りの電車で研究室の方といろいろなお話や意見交換が出来、非常に充実した時間を過ごすことが出来た。

講師の先生方、国立情報学研究所のスタッフ、そして参加者の皆様にこの場を借りてお礼を申し上げます。