

平成 17 年度情報処理軽井沢セミナーレポート

関東学院大学図書館 電子情報係

外崎 みゆき

1. 作成した（取り組んだ）ポータル名 若しくは 課題

OAI-PMH によるメタデータのハーベストと、汎用連想計算エンジン GETA を使用した連想検索システムの構築

2. 1 の概要

今回の研修では OAI-PMH について知識を深める事を大きなテーマとした。

本学の場合、パッケージシステムにて紀要論文の OAI-PMH リポジトリを構築している。今回は、リポジトリとして提供内容を把握できている本学図書館（紀要論文）をターゲットとしてハーベスティングを行い、連想検索の環境を構築した。

- ・ OAI-PMH で提供している紀要論文をハーベスト
- ・ XSL 書式により必要データを抽出、CSV 形式のファイルに加工
- ・ 「description」項目を chasen システムにより構文解析
- ・ 解析結果より WARM を作成し GETA の検索対象とする

3. 演習とその成果 何を計画し、実装して、何ができたか。

※成果物のスクリーンショットも

(1) 演習第 1 日

OAI-PMH にてメタデータ・ハーベストが可能なサイトの調査を行った。次に、実際にハーベストのテストを「wget」コマンドにより試みた。また、前半の講義で紹介された、他の事例（内容）を調査するために、実際に運用されている下記のポータルサイトなどに接続して動作を確認した。

[鹿児島大学附属図書館・インターネット学術情報](#)

[千葉大学学術成果リポジトリ CURATOR](#)

[九州大学附属図書館：きゅうと LinQ](#)

[Directory of open access journals](#)

[ARL Scholars Portal Project](#)

(2) 演習第 2 日

chasen の環境を構築する作業を行った。参考になると思うので手順を示す。

JAVA と apache ant をダウンロードしてインストール、perl モジュールのインストールは CPAN を使用して行った。更に Dart-0.2、iconv-1.9.1 をインストールする。

Chasen および GETA は RedHat9 環境では問題なく環境構築が可能である。

- ※ 研修後、大学にて FedoraCore3 環境で構築した時は、コンパイル時にエラーが発生してビルドが出来なかったが、http://hoshizawa.no-ip.com/suzaku/manual_2.html に対処方法が詳しく掲載されているので参考にして、ビルドに成功した。
辞書は ipadic-2.7.0 を使用している。

(3) 演習第3日

今回、テストデータとして自学の紀要データを選択したので、wget を使用して自館のリポジトリからハーベスティングを行った。ハーベスティングした XML ファイルより XSLT の書式設定を利用して GSV 形式のファイルに変換し、「DESCRIPTION」項目を chasen で構文解析を行うための入力データとした。構文解析結果である頻度ファイルより WAM を作成し、GETA の環境を構築した。

4. 研修で学んだ技術及び知識

参考になった URL とその簡単な内容紹介（1行程度で）

<http://www.openarchives.org/Register/BrowseSites>（Registered Data Providers）

<http://www.kanzaki.com/docs/sw/rss.html>（RSS -- サイト情報の要約と公開）

<http://www.atmarkit.co.jp/fxml/dictionary/indexpage/xmlindex.html>（XML 用語事典）

<http://fuji.sakura.ne.jp/~yada/talk2000/perl.shtml>（CPAN 初級）

<http://geta.ex.nii.ac.jp/>（汎用連想計算エンジン（GETA） 公開HP）

<http://chasen.naist.jp/hiki/ChaSen/>（形態素解析システム茶釜）

5. 事前準備として必要と思われるもの

（不足していたソフトウェアや予習事項等）

演習に使用したノートパソコンには RedHat9 の環境を設定した。その他、必要と思われるオープンソースは事前にノート PC にダウンロードしておいたが、環境が準備できたのは perl モジュールのインストールのみだった。

事前に準備できる環境は極力整えておいたほうが演習時に時間が有効に活用できる事を痛感した。私の場合は Linux や perl など各ソフトに関する基礎知識が不足していたので準備したパソコンの環境設定が充分では無かった。

6. 今後の課題（職場で更に調査する必要のあるもの等）

- (1) ハーベスト、データ加工、構文解析、WAM の作成と各ステップを自動化する方法
- (2) ハーベスティングが可能なリポジトリの調査
- (3) FedoraCore3 環境でのシステム構築
- (4) 多言語への対応

(5) 今回の演習では構築を断念した Dspace 環境の構築

7. 今後の計画（実際のポータル構築計画等）

- (1) 今回の演習ではハーベスト、データ抽出、構文解析、WAM の作成と各ステップごとに手動で作業をすすめて GETA の検索環境を作成した。リポジトリはデータが追加されるため、この処理をスクリプトまたはバッチ処理として組込むことで自動化し、定期的にハーベスティングすることを考えたい。
- (2) テストモデルとして自館からハーベストを行っているが、NII のリポジトリをターゲットとして、XSL にて論文を条件としてデータ抽出を行い、WAM を構成することで、検索対象が自館から国内の論文となり収録内容が充実する。
- (3) ハーベスティングが可能なリポジトリを調査し、ターゲットに加えることで収録論文を増やしたい。
- (4) 職場に戻り、FedoraCore3 環境で復習をかねて同様のシステムを構築している。環境が整えば DOAJ などターゲットとして検討したい。

8. 演習の感想

一緒に研修会に参加した方たちの意識の高さに圧倒されました。図書館の有り方について視野を広げると共に技術力のアップが必要だと痛感しました。判らない事が質問できるという天国のような環境で演習に取り組ませて頂いて、とてもよい勉強になりました。今回の研修で得た知識を有効に活用できるよう努力したいと思います。

Portal 構築用のオープンソースである Dspace については情報公開目的の要素が強いので、今回の演習ではチャレンジしておりません。Dspace を含めて、オープンソースの活用を真剣に考えていきたいと思っています。

9. 備考、その他

構想はあっても技量が伴わないため、ご迷惑をおかけしました。また、講師の皆様には人一倍、お世話になりました。講師、参加者、スタッフの皆さまのご支援により、無事に研修を終了し、何とか形に残る成果が得られました。この場を借りて御礼申し上げます。

仕事の合間に手を動かすので、遅々として進展しない状況ではありますが、大学に戻ってから FedoraCore3 環境で再構築を試みています。研修では文字コードを EUC-JP で扱いましたが、今回は UTF-8 です。今後の Linux 環境、入手可能なオープンソースを考えると、文字コードは UTF-8 が主流になるかと思っています。「連想検索用インターフェイスのサンプル CGI」が UTF-8 をサポートしていないので、何処までの工程を UTF-8 で扱うかを悩みながら作業を進めています。

wgetによる自館とniiのハーベスト

```
AIR MAIL - フォルダツリ
リブレイス
150.38.170.46 - Tera Term VT
File Edit Setup Control Window Help
[tozaki@libs1011 ~]$ wget -O nii.xml "http://ju.nii.ac.jp/cgi-bin/oai/oai2.0?verb=ListRecords&from=2005-01-01&metadataPrefix=junii"
--16:38:03-- http://ju.nii.ac.jp/cgi-bin/oai/oai2.0?verb=ListRecords&from=2005-01-01&meta
dataPrefix=junii
=> nii.xml
ju.nii.ac.jp をDNSに問いあわせています... 157.1.18.128
ju.nii.ac.jp[157.1.18.128]:80 に接続しています... 接続しました。
HTTPによる接続要求を送信しました、応答を待っています... 200 OK
長さ: 特定できません [text/xml]

[ <=> ] 164,688 25.29K/s

16:38:10 (25.25 KB/s) - `nii.xml' saved [164,688]

[tozaki@libs1011 ~]$ wget -O kgu_oai.xml "http://opac.kanto-gakuin.ac.jp/cgi-bin/oai/so_oai2.cgi?verb=ListRecords&metadataPrefix=junii"
--16:39:00-- http://opac.kanto-gakuin.ac.jp/cgi-bin/oai/so_oai2.cgi?verb=ListRecords&meta
dataPrefix=junii
=> `kgu_oai.xml'
opac.kanto-gakuin.ac.jp をDNSに問いあわせています... 150.38.170.12
opac.kanto-gakuin.ac.jp[150.38.170.12]:80 に接続しています... 接続しました。
HTTPによる接続要求を送信しました、応答を待っています... 200 OK
長さ: 特定できません [text/xml]

[ <=> ] 153,989 884.20K/s

16:39:14 (883.74 KB/s) - `kgu_oai.xml' saved [153,989]

Ad
[tozaki@libs1011 ~]$ ls -ls
合計 2420
156 -rw-r--r-- 1 tozaki tozaki 153984 11月 9 12:59 #kgu.xml#
964 -rw-rw-r-- 1 tozaki tozaki 982850 11月 1 21:32 doaj.xml
36 -rw-rw-r-- 1 tozaki tozaki 34154 11月 9 12:49 kgu.csv
156 -rw-r--r-- 1 root root 153994 11月 9 12:48 kgu.xml
156 -rw-rw-r-- 1 tozaki tozaki 153989 11月 1 21:19 kgu.xml.ori
156 -rw-rw-r-- 1 tozaki tozaki 153988 10月 18 17:44 kgu20051018.xml
156 -rw-rw-r-- 1 tozaki tozaki 153989 11月 9 16:39 kgu_oai.xml
168 -rw-rw-r-- 1 tozaki tozaki 164688 11月 9 16:38 nii.xml
144 -rw-rw-r-- 1 tozaki tozaki 139832 10月 18 19:52 nii2003.xml
148 -rw-rw-r-- 1 tozaki tozaki 146238 10月 18 19:54 nii2004.xml
172 -rw-rw-r-- 1 tozaki tozaki 168335 10月 18 19:55 nii2005.xml
4 -rw-rw-r-- 1 tozaki tozaki 865 10月 18 19:44 nii_20050101.xml
4 -rw-r--r-- 1 root root 659 11月 9 12:37 sampleUTF.xml
```

自館からハーベストした結果 xml ファイル

```
kgu.xml - TeraPad
ファイル(F) 編集(E) 検索(S) 表示(V) ウィンドウ(W) ツール(T) ヘルプ(H)
|10 |110 |120 |130 |140 |150 |160 |170 |180 |190 |200 |210 |220 |
|?xml version="1.0" encoding="UTF-8"?>+
|<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" +
|  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" +
|  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/+
|    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">+
|<responseDate>2005-11-01T12:14:14Z</responseDate>+
|<request verb="ListRecords" metadataPrefix="junii">http://opac.kanto-gakuin.ac.jp/cgi-bin/oai/so_oai2.cgi</request>+
|<ListRecords>+
|  <record>+
|    <header>+
|      <identifier>oai:opac.kanto-gakuin.ac.jp:NI00000005</identifier>+
|      <timestamp>2005-06-17</timestamp>+
|      <setSpec>NI</setSpec>+
|    </header>+
|    <metadata>+
|      <meta>+
|        xmlns:junii="http://ju.nii.ac.jp/oai/" +
|        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" +
|        xsi:schemaLocation="http://ju.nii.ac.jp/oai/+
|          http://ju.nii.ac.jp/oai/junii.xsd">+
|          <code>NI00000005</code>+
|          <adate>20040614141112</adate>+
|          <update>20050617130114</update>+
|          <title>モバイル型ネットワーク商品の進化</title>+
|          <title.transcription>モバイル ガタ ネットワーク ショウヒン ノ シンカ</title.transcription>+
|          <title.transcription>ケイタイ デンワ ヲ ジレイ トシテ</title.transcription>+
|          <title.alternative>携帯電話を事例として</title.alternative>+
|          <title.alternative>The Evolution of the Commodity in Form of Network: in Case of Cellular Phone</title.alternative>+
|          <creator>石崎 悦史</creator>+
|          <creator.transcription>イシザキ ヨシフミ</creator.transcription>+
|          <creator.alternative>Yoshifumi Ishizaki</creator.alternative>+
|          <subject>携帯電話, ネットワーク商品, ネットワーク家電, 商品進化, 日本の商品開発, ユビキタス, デファクトスタンダード</subject>+
|          <subject xsi:type="NDC">694.21</subject>+
|          <description>ネットワークの形態をとった商品の進化を検討し, その進化の法則性を解明するために, 携帯電話を事例としてとりあ
```

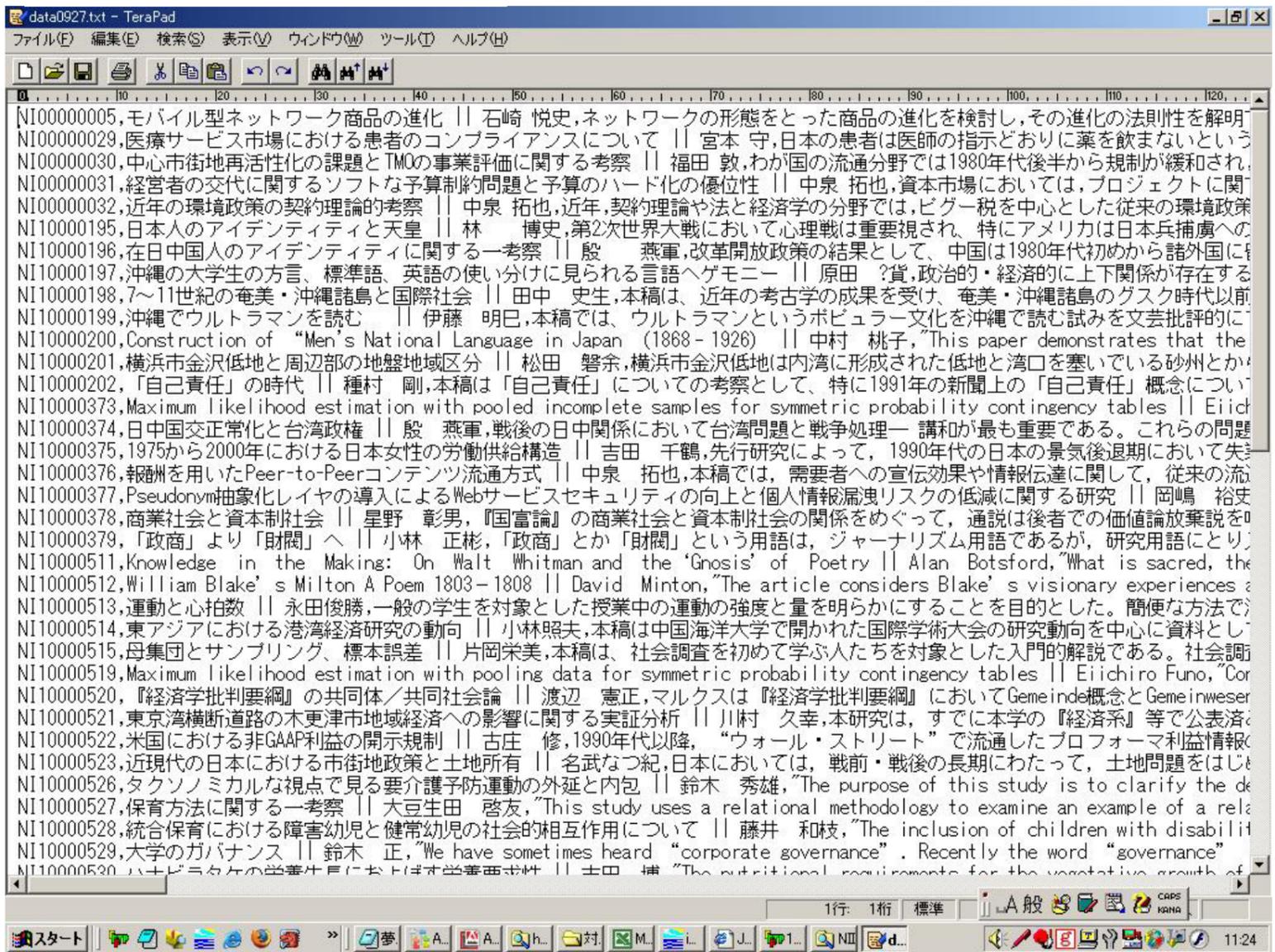
XML から CSV へ変換する XSL 書式

```
sampleUTF.xml * - TeraPad
ファイル(F) 編集(E) 検索(S) 表示(V) ウィンドウ(W) ツール(T) ヘルプ(H)
[Icons]
110 120 130 140 150 160 170 180 190 200 210 220
<?xml version="1.0" encoding="UTF-8"?>+
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" >+
<xsl:output method="html" encoding="EUC-JP" />+
↓
<xsl:template match="/">+
<!--+
<xsl:template match="/">code,creator,description+
<xsl:text>+
</xsl:text>+
-->+
↓
    <xsl:apply-templates select="OAI-PMH/ListRecords/record/metadata/meta"/>+
</xsl:template>+
↓
<xsl:template match="OAI-PMH/ListRecords/record/metadata/meta">+
<xsl:value-of select="code" />,<xsl:value-of select="title" /> || <xsl:value-of select="creator" />,+
&quot;<xsl:value-of select="description" />&quot;+
<xsl:text>+
</xsl:text>+
</xsl:template>+
</xsl:stylesheet>+
↓
[EOF]
```

17行: 1桁 標準 LA 般 CAPS LANG

スタート 夢 A. A. h... 村 M... L... J... 1... NI s... 11:21

XML より XSL で変換した CSV データ



女性の労働条件と育児への影響 を検索した結果

Live Association Central - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

戻る 検索 お気に入り メディア リンク

アドレス http://150.38.170.41/cgi-bin/assoc.cgi

Live Association Central

[HOME]

Search for Go

Show: Measure: [LONG QUERY]

Select some items and in

Show: Measure:

Top 10 of 16 found documents in kiyo data.

- (1.00) [1975から2000年における日本女性の労働供給構造](#) || 吉田 千鶴
- (0.32) [近現代の日本における市街地政策と土地所有](#) || 名武なつ紀
- (0.32) [商業社会と資本制社会](#) || 星野 彰男
- (0.19) [統合保育における障害幼児と健常幼児の社会的相互作用について](#) || 藤井 和枝
- (0.18) [沖縄の大学生の方言、標準語、英語の使い分けに見られる言語ヘゲモニー](#) || 原田 ?貨
- (0.15) [東京湾横断道路の木更津市地域経済への影響に関する実証分析](#) || 川村 久幸
- (0.15) [7~11世紀の奄美・沖縄諸島と国際社会](#) || 田中 史生
- (0.14) [在日中国人のアイデンティティに関する一考察](#) || 殷 燕軍
- (0.12) [医療サービス市場における患者のコンプライアンスについて](#) || 宮本 守
- (0.12) [タクノミカルな視点で見る要介護予防運動の外延と内包](#) || 鈴木 秀雄

Topic words to summarize the result (Top 30).

的 こと 影響 者 化 政策 日本 効果 の 社会 統合 労働 幼児 本稿 もの 力 地域 障害 分析 論 結果 介護 中国人 研究 言語 運動 予防 形成 価値 ため

ページが表示されました

あ般 夢 AL A pa 対 Mi 平 L iLi J

11:13