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 CSV 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筌）

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なども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。

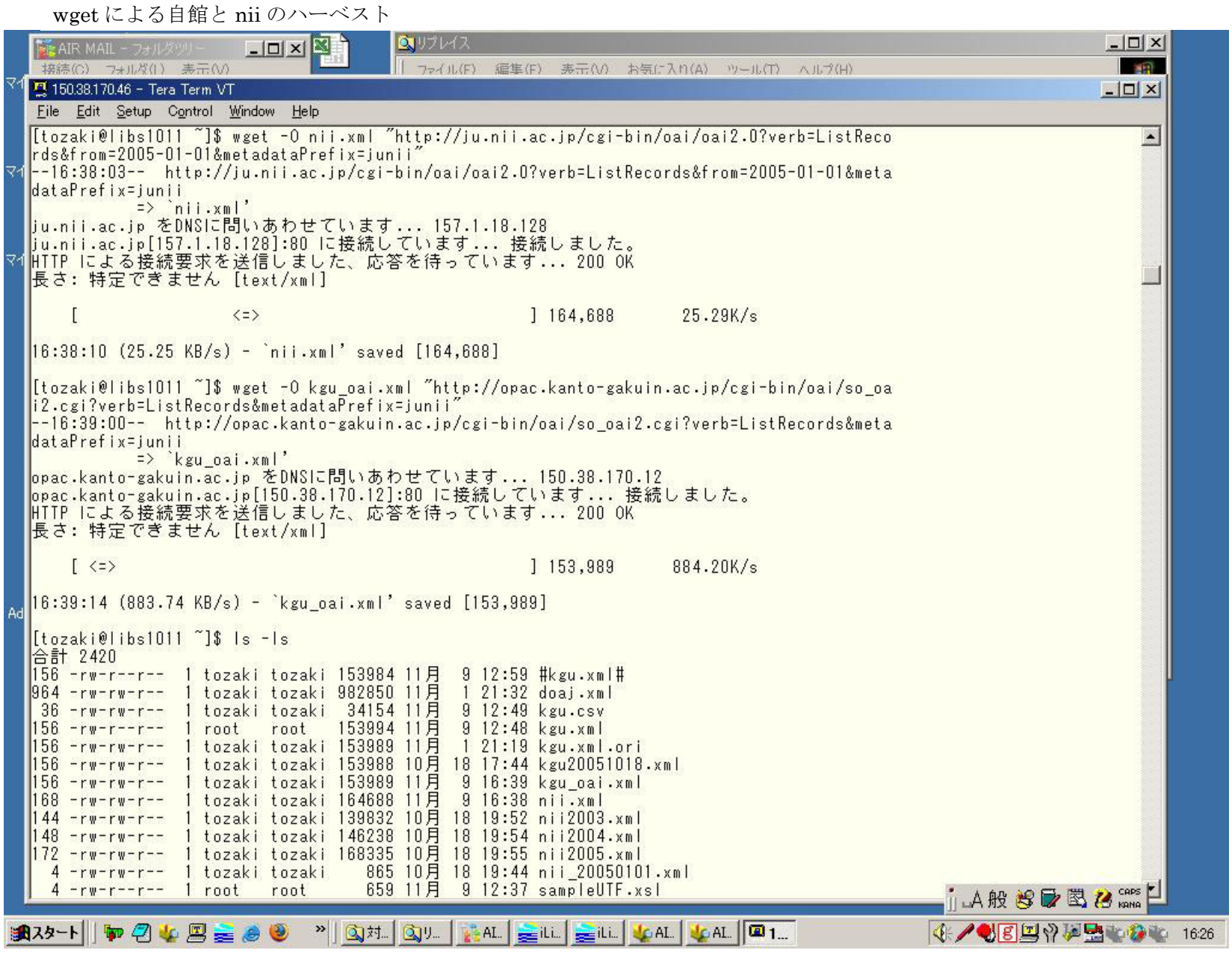
Portal構築用のオープンソースであるDspaceについては情報公開目的の要素が強いので、今回の演習ではチャレンジしておりません。Dspaceを含めて、オープンソースの活用を真剣に考えていきたいと思います。

９．備考、その他

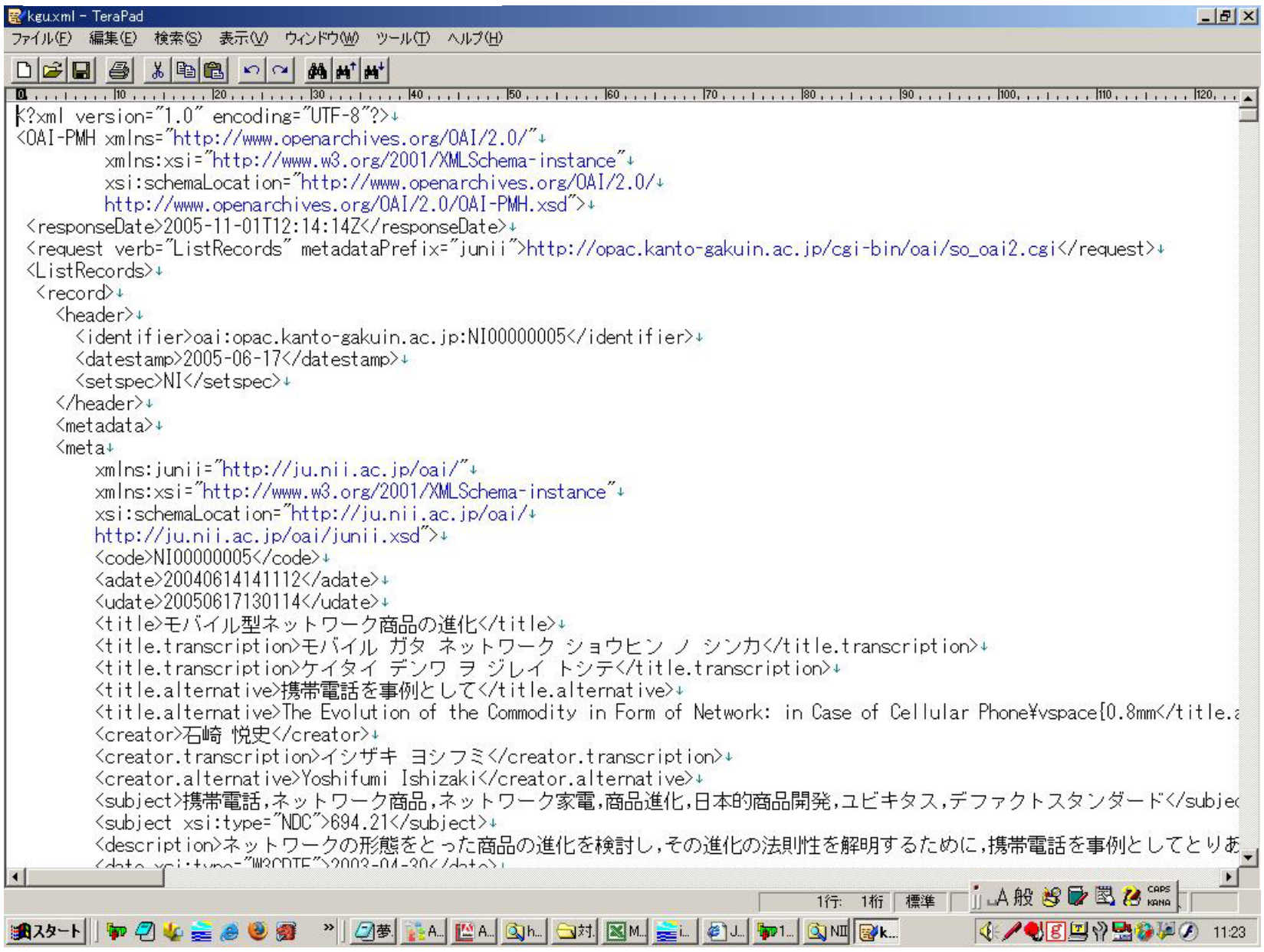
構想はあっても技量が伴わないため、ご迷惑をおかけしました。また、講師の皆様には人一倍、お世話になりました。講師、参加者、スタッフの皆さまのご支援により、無事に研修を終了し、何とか形に残る成果が得られました。この場を借りて御礼申し上げます。

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



自館からハーベストした結果 xml ファイル



XML から CSV へ変換する XSL 書式

sampleUTF.xsl * - TeraPad

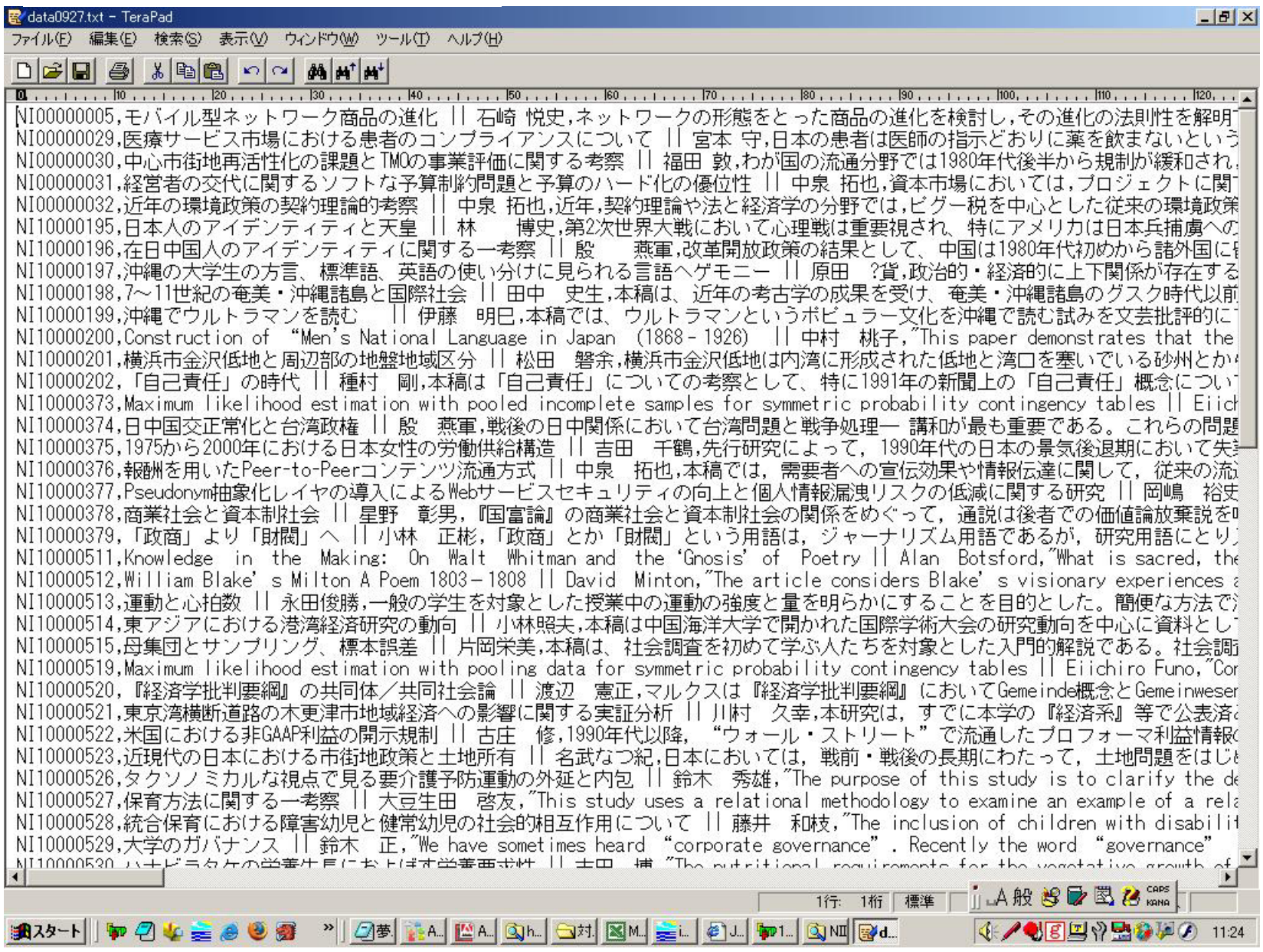
ファイル(F) 編集(E) 検索(S) 表示(V) ウィンドウ(W) ツール(T) ヘルプ(H)

```
<?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" />,↓
&amp;quot;<xsl:value-of select="description" />&amp;quot;↓
<xsl:text>↓
</xsl:text>↓
</xsl:template>↓
</xsl:stylesheet>↓
↓
[EOF]
```

17行: 1桁 標準

スタート 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