SRU/SRW を用いた教育図書館資料の書誌検索システムの構築

A SRU/SRW-based Information Retrieval System for Bibliographic Data at Library of Education

江草由佳 *1，高久雅生 2
Yuka EGUSA, Masao TAKAKU

*1 国立教育政策研究所教育研究情報センター
Educational Resources Research Center, National Institute for Educational Policy Research
〒153-8681 東京都目黒区下目黒 6-5-22
E-mail: yuka@nier.go.jp
2 情報・システム研究機構新領域融合研究センター
Transdisciplinary Research Integration Center, Research Organization of Information and Systems
〒101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所内
E-mail: masao@nii.ac.jp

日本語書誌データを対象として SRU/SRW プロトコルに対応した書誌検索システムを開発した。国立教育政策研究所教育図書館が提供している「教育研究論文索引」の日本語論文書誌データ約 12 万件を対象とし，さらに同図書館 OPAC ログデータを元にしたテストクエリセットを作成し，これらの検索が SRU/SRW を通じて行えるかを確認するとともに，実運用規模における検索システムの可能性について考察した。

This paper reports an information retrieval system for Japanese bibliographic data based on SRU/SRW protocol. The system consists of a client which can access to our SRU/SRW server, and of a server which provides bibliographic data for educational papers at Library of Education. The server has about 120 thousands bibliographic records of "*Kyoiku-Kenkyu-Ronbun-Sakuin*", and are capable for basic retrieval functions in SRU/SRW, such as Boolean queries, queries for specific search indexes, etc. Its capabilities are confirmed by using two kinds of query datasets, functional test-set and OPAC log-based one, in a batch search experiment.

キーワード: SRU/SRW，情報検索システム，検索プロトコル，
SRU/SRW, information retrieval system, information retrieval protocol

1 はじめに

1980 年代末から 1990 年代を通じて標準化された情報検索プロトコル Z39.50[1] は，欧米の図書館業界を中心に普及が進んだ．しかし，その一方で，Z39.50 プロトコル自体は，現在のインターネット普及以前にその開発が始まり，OSI ベースの通信プロトコルとして標準化が進められ，かつ，その標準化作業が WWW 普及以前にほぼ終わっていたという 2 点から，プロトコル自体が他のオンラインシステムの開発において使われている環境から次第に外れた存在となっていった．また，多くの機能を盛り込んだ標準仕様であったため，その全ての機能を網羅することは難しく，また簡易な検索機能を実装するだけで良い場合にも，仕様全体のどの部分を実装すれば良いかを確認するのに苦労するなど，新たに実装を行なうことが難しい点もあった．とりわけ，WWW サーチエンジンが普及し，簡易な検索条件でのみ事足りるとする一般の利用者が増えたため，これらの要求に応える場合に Z39.50 を採用することはそのコストに見合わない

といったケースも出てきている.

これらの問題点を克服するために, Z39.50 標準管理機構において次世代プロトコルの検討が行われ, HTTP を通信プロトコルとし, XML を要求/応答メッセージの記述に用いるような, より現代的な方向に開発が進められた. この結果として開発されたプロトコルが SRU/SRW (Search/Retrieve via URL, Search/Retrieve Web Service) [2] の 2 つの次世代プロトコルである. SRU は REST (Representational State Transfer) フレームワークに基づくプロトコルとして設計され, SRW は SOAP 仕様に基づく検索要求を行うプロトコルである. これらのプロトコルは Z39.50 から離れ, 完全な別仕様として開発され, 互換性を犠牲にすることにより, より現代的なプロトコルとしてシステムを実装できるようになった. SRU/SRW は現在 1.1 版の仕様が公開されており, 1.2 版のリリースも近々予定されている. 日本国内でも, SRU/SRW の紹介記事 [3] が出ており, 国立国会図書館 [4], 筑波大学 [5] など SRU/SRW を用いた検索システムが提供されつつある.

一方で, 現在のところ, SRU/SRW プロトコルの実装は実験段階にとどまっており, これらが Z39.50 プロトコルと比して, どのようなインパクトをもたらすかは明らかではない. そこで, 筆者らはこれまで Z39.50 日本語検索システムを開発してきた経験 [6][7][8][9] を踏まえて, 本稿では SRU/SRW のクライアントおよびサーバの実装を通じて, その特性を考察する.

2 SRU/SRW

SRU/SRW の検索要求は CQL (Common Query Language*1) を使用する. CQL は SRU/SRW の次世代プロトコルとともに新たに開発された. CQL はシステム間のやりとりのためだけでなく, 人による可読性も考慮に入れて設計されているために, 利用者が直接 CQL を入力しての検索もできるような検索式である.

SRU における要求を図 1 に示す. ベース URL は検索サーバを表わす. クエリパートは, プロトコルのバージョンや検索結果の返戻要求, 検索式などプロトコル上のパラメータを表現する. この例では「読書」で検索したヒット件数を要求している.

サーバからのレスポンスは XML で返される. 図 2 は, 本研究で開発した SRU サーバから 1 レコードを返した際のレスポンス例である.

3 書誌データ：教育研究論文索引

本研究で使用した書誌データは, 国立教育政策研究所教育図書館が作成し, Web 上で提供している「教育研究論文索引」[10] である.「教育研究論文索引」は, 教育図書館が受け入れた逐次刊行物の中から教育に関する記事を採録したものである. 本研究では, 1968 年および 1983～2005 年の 24 年分 122,510 レコードを使用した.

論題, 著者名, 著者名よみがな, 掲載誌名, 巻号, 掲載ページ数, ISSN, 請求記号, 発行年月, 文献番号, キーワード, 登録日, 冊子体の書誌項目がある.

4 開発システム

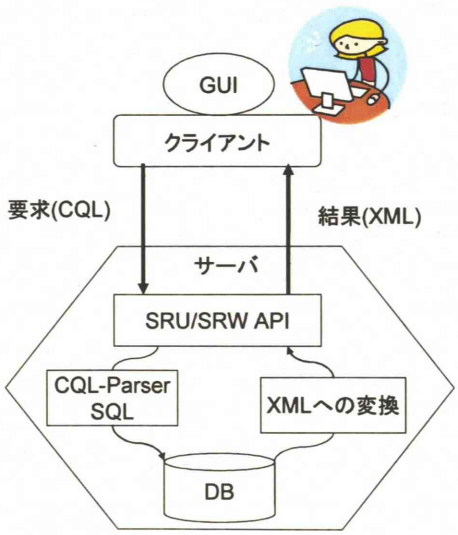

図 3 システムの構成図

*1 SRU 1.2 版からは Contextual Query Language という名称となる予定である.

```
URL 例：
http://kaede.nier.go.jp/epi?version=1.1&operation=searchRetrieve&query=%E8%AA
%AD%E6%9B%B8

ベース URL：http://kaede.nier.go.jp/epi
クエリパート：?version=1.1&operation=searchRetrieve&query=%E8%AA%AD%E6%9B%B8
```

図 1　SRU 要求: 検索式は「読書」を指定している．なお，URL 中での日本語文字列は UTF-8 として表現した上で URI エンコードするため，数値参照となっている．

```
<?xml version="1.0" encoding="UTF-8"?>
<zs:searchRetrieveResponse xmlns:zs="http://www.loc.gov/zing/srw/">
 <zs:version>1.1</zs:version>
 <zs:numberOfRecords>1</zs:numberOfRecords>
 <zs:records>
  <zs:record>
   <zs:recordPacking>xml</zs:recordPacking>
   <zs:recordData>
     <xml>
       <pubdate>1996.3</pubdate>
       <journal_id>370.5-59-29</journal_id>
       <page>p.1〜175</page>
       <regdate>1997.9.30</regdate>
       <author>林部一二，堀井啓幸，杉本真理子，佐藤晴雄</author>
       <keywords>教育内容・方法　図書館教育　社会教育・生涯学習　家庭教育</keywords>
       <volnum>29</volnum>
       <book_mapping>教育研究論文索引 1996 年版</book_mapping>
       <title>子どもの読書の実態と家庭における指導に関する調査研究</title>
       <journal>日本教材文化研究財団　調査研究シリーズ</journal>
       <paper_id>9603611</paper_id>
     </xml>
   </zs:recordData>
   <zs:recordPosition>1</zs:recordPosition>
  </zs:record>
 </zs:records>
</zs:searchRetrieveResponse>
```

図 2　SRU/SRW からの返戻レコードの一例: 教育研究論文索引独自のスキーマにもとづく XML 形式

開発システムの構成を図 3 に示す．システムは SRU/SRW クライアントと SRU/SRW サーバに分かれる．

4.1　SRU/SRW サーバ

サーバは基本的な検索要求およびレコード返戻要求に応える機能を有している．

クライアントからの検索要求があったら，まず検索要求に含まれる CQL 検索式を解析し，内部のデータベースで用いている SQL に変換する．内部データベースにおいて SQL を用いた検索が行われたら，その結果を XML 形式に変換してクライアントに返す．この間，検索式にエラーがある場合や，対応していない検索式を渡された場合，対応していない返戻レコード形式を要求された場合には，SRU/SRW 仕様に沿ったエラーメッセージを返す．

キーワード検索，フレーズ検索，and，or，not 演算子を使用した論理検索，検索項目を指定した検索，Dublin Core（DC）の検索項目を指定した検索などの基本的な CQL 検索式に対応した．

検索結果として返されるレコードは教育研究論文索引用の独自 XML スキーマに基づくものを採用した．図 2 に返戻レコードの例を示す．

4.2　SRU/SRW クライアント

図 4 に本研究で開発した SRU/SRW クライアントのスクリーンショットを示す．

クライアントはシンプルな構成として構築しており，Web ブラウザ上のフォームで利用者に CQL を入力してもらい，検索をサーバ側に渡す構成となっている．

サーバからの応答として返ってきた XML

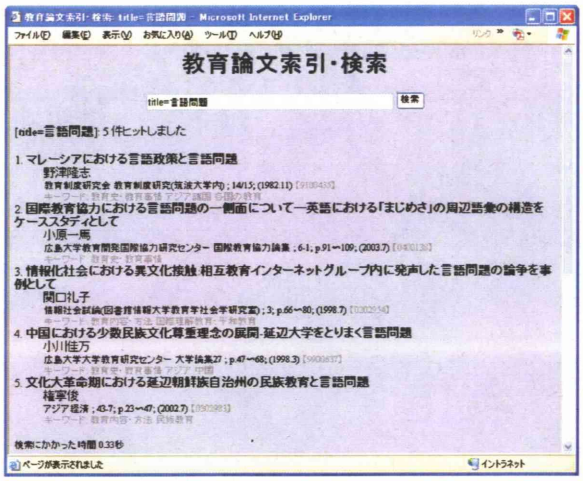

図 4 SRU/SRW クライアント

メッセージは内部でパースして，100 件毎に表示する．

5 テスト用クエリ集合の開発

開発したサーバに対する検索実験として，以下の 2 種類の観点から，テスト用クエリ集合を作成し，検索が行えるか確認した．

1. 機能テストセット: CQL の様々な機能に対してどの程度対応しているかどうかチェックする目的のもの
2. ログベーステストセット: 現実の検索要求に対してどれだけ検索可能かをチェックする目的のもの

機能用テストセットについては，既存の CQL 関連ソフトウェア，仕様書で提供されているテストセット[*2][*3]を元に作成した．ログベーステストセットについては，教育図書館蔵書検索（OPAC）[*4]の検索ログを使用して作成した．この OPAC では主に書籍を対象としており，今回対象としたデータは論文を対象としている点は異なるものの，教育分野に関する情報ニーズである点と教育図書館が所蔵している資料の検索という観点から類似しているため，擬似的な検索要求とみなすことができると判断し，この検索ログを使用することとした．OPAC のログデータから収集した検索式は CQL 形式に変換した．これらテストセットの検索式は，1 行に CQL で表現された 1 つの検索要求を記述するという形式で作成しておき，それをバッチで検索実行する形で使用した．

機能用テストセットとしては 128 の検索式を生成し，その全てについて検証を行なった．これら 128 件の検索式のうち，6 件は意図的にエラーとなるように作成した検索式であり，これらがエラーを返すことを確認した．また，10 件については，未対応のためエラーを返すことも確認している．なお，この他 15 件の近接演算を用いた検索式には未対応であるが，正しくエラーが返せていない．これら以外の 97 件の検索式においては検索要求を受け付けられることを確認した．ログベーステストセットとしては，8216 件の検索式を生成し，その全てについて検索を実行した．それぞれのテストセットのヒット件数の分布は表 1 に示す．

表 1 テストセットのヒット件数の分布

ヒット件数	テストセットクエリ数 機能	ログ
0	49	7231
1～100	24	885
101～1000	12	53
1001～	27	47
計	97	8216

6 考察

日本語書誌データを使用した SRU/SRW 検索システムにおける留意点を挙げ，それぞれについて考察する．

6.1 日本語検索

Z39.50 と SRU/SRW は，欧米を中心にプロトコル仕様の開発が進んだために，日本語のような膠着語で分かち書きの必要な言語，複数の文字コードが使用されている言語についてはあまり考慮されていない．日

*2 SRWLucene http://wiki.osuosl.org/display/OCKPub/SRWLucene
*3 CQL::Parser http://www.textualize.com/cql-parser
*4 http://opac.nier.go.jp/

本語検索式への対応という面においては，これらの指定をどのようにして日本語環境において実装していくか，その手法についての議論は十分に行われたとは言いがたい．この点は，既に Z39.50 にも同様の問題が指摘されていた [11] が，これらの問題は SRU/SRW にも引き継がれている [12][13].

現在，比較的一般に行われているのは形態素解析器等を用いた辞書，連接ベースの分かち書きであるが，これらは書誌レコードや検索式などの短い文字列に対しては，必ずしも高精度な分かち書きを行える訳ではないという欠点があり，検索精度の低下，検索漏れの発生といった懸念もある．ただ，SRU/SRW および Z39.50 では，指定がなされない限り，データベース内部でどのようなマッチング手法を使用しているかは大部分データベース側で自由に設定できるよう，抽象化されている．そのためプロトコル仕様（この場合は CQL）自体が欧米語を想定していたとしても，ある程度自由な設定が効くはずである．一方で，CQL の=や all, any といったリレーションと呼ばれる文字列マッチングもしくは項目マッチングの指定項目を，どのようにデータベースにおけるマッチングに適用させていくかは，日本語の場合は一定の方式とされるほどの方式は存在せず，また，効率性の面からも様々な実装方式に分かれることと思われる．この点については，実装方式のガイドライン制定や，仕様に対するフィードバックを行っていく必要性があるように思われる．

なお，本報告において開発したシステムにおける実装では，検索要求で送られたキーワードが中間一致するレコードをヒットするように開発したため，そもそも「ワード」概念を意識せずに検索が可能となっており，利用者が入力した文字列が含まれるレコードは必ずヒットする．そのため，対象としている言語が日本語だけの場合はこのような実装も一手法といえる．

6.2 検索要求における文字コード

SRU/SRW では，XML の結果返戻を前提にしている点などから，文字コードは XML における Unicode 指定を基本として設計されている．さらに，SRU では簡易な URI 中における検索要求の引き渡しのために検索要求中の文字コードは UTF-8 に限定されている．仕様として文字コードが指定されたことから，日本語のように様々な文字コードを使用している環境でもシステムは XML 処理および URI 中の UTF-8 指定にのみ対応すればよく，システム間での文字コードにおけるネゴシエーションなどは必要なくなった．この点は Z39.50 と異なる．

とりわけ SRU においては使用文字コードを UTF-8 に限定しているため，日本語文字等を使用して検索したい場合には，必ず UTF-8 への変換を行う必要がある．既存の環境が UTF-8 環境ではなく Shift_JIS や EUC 環境である場合，サーバ側であれば，レコードデータもしくは検索サーバにおける対応が必要になり，また，クライアント側であれば表示および入力時の UTF-8 対応が必要となってくる．

今回開発したクライアントでは，エンドユーザの入力からサーバへの受け渡しまで全てを UTF-8 で行うシステムとして開発した．一方，サーバ側では，クライアントからの要求は UTF-8 を始めとする要求に応えるものの，内部処理は元々の日本語データとの関係から EUC-JP を用いた．このため，要求が UTF-8 で来た場合，内部で一旦文字コード変換を行ってからデータベース検索を行い，検索結果を再度 UTF-8 に戻すなどの処理を行った．

6.3 サーバ側における選択の明示

SRU/SRW においては，様々な機能や CQL において，指定が何も無い場合はサーバ側における選択（serverChoice）にゆだねるという仕様になっている．これには，サーバを開発する際に，トレードオフを考慮して実装の手間を減らして開発工数を下げたり，サーバ側での工夫で多様なマッチング手法等を採用できるなどの点で貢献があると考えられる．特に，検索のアクセスポイントに関しては，DC などの既存の標準スキーマとのマッピング等は必須ではないため，現

在運用しているデータベースを SRU/SRW 対応にして提供していく際にそのまま XML 形式における提供を検討するだけで良いなど，便利な点もある．しかし，横断検索等のニーズを考えた場合，DC などのよく使われている標準とのマッピングを行なって，検索可能にすることを検討するべきである．

さらに，どのようなサーバ実装が選択されているか分からない場合は，利用の際に不都合が起こるケースが想定されるため，サーバについての説明を利用者に対して提供する必要が出てくる．これへの対策としては，SRU/SRW のサーバ情報を提供するための Explain 機能を用いて提供することが可能なのではないか．その際にはどのような情報が利用者にとって必要なのかについて検討する必要がある．

7 おわりに

本研究では，SRU/SRW のクライアントおよびサーバの開発を行った．実装にあたっては国立教育政策研究所教育図書館が提供する教育研究論文索引の主に日本語論文書誌データ約 12 万件を対象とし，さらに同図書館 OPAC ログデータを元に，擬似的な検索要求としてテストセットを作成し，これらの検索が SRU/SRW を通じて行えるかを確認し，実運用規模における検索システムの可能性を示した．また，特に SRU/SRW の日本語環境における対応状況の確認も行ない，あわせて考察した．

謝辞

本研究は科学研究費補助金（課題番号: 18800080）の助成を受けた．

参照文献

[1] Z39.50 Maintenance Agency. ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50) : Application Service Definition and Protocol Specification. 1995, 156p. (online), available from http://www.loc.gov/z3950/agency/1995doce.html.

[2] The Library of Congress. SRU: Search and Retrieve via URL (Standards, Library of Congress). available from http://www.loc.gov/standards/sru/.

[3] 平山 亮. 情報検索および図書館相互貸借の標準規格. 情報の科学と技術. Vol.56, No.7, 2006, p.307-311.

[4] 国立国会図書館. 国立国会図書館デジタルアーカイブポータル (ndl-dap) - SRW とは？-SRW-システム紹介 http://www.dap.ndl.go.jp/home/modules/pukiwiki/?SRW（参照 2007-04-19）.

[5] 筑波大学. 知的コミュニティ情報システム仕様書. http://www.kc.tsukuba.ac.jp/spec060621.pdf（参照 2007-04-18）.

[6] 宇陀 則彦, 江草 由佳, 高久 雅生, 石塚 英弘. Z39.50 による日本語書誌データ検索システム. 情報知識学会誌. Vol.9, No.2, 1999, p.1-15.

[7] 江草 由佳, 真野 泰久, 宇陀 則彦, 石塚 英弘. Z39.50 プロトコルによる日本語書誌データ情報検索システム. 第 6 回研究報告会講演論文集. 情報知識学会. 東京, 1998-05. 情報知識学会, 1998, p.29-36.

[8] 高久 雅生, 江草 由佳, 宇陀 則彦, 石塚 英弘. “Z39.50 による書誌データ検索システムの構築: Dublin Core を共通スキーマとして”. ディジタル図書館. No.16, 1999, p.97-106.（オンライン）, 入手先 http://www.dl.slis.tsukuba.ac.jp/DLjournal/No_16/12-masao/12-masao.html.

[9] 江草 由佳, 高久 雅生, 宇陀 則彦, 石塚 英弘. Z39.50 データベース選択支援 環境. 情報知識学会誌. Vol.11, No.2, 2001, p.1-10.

[10] 江草 由佳. 教育図書館における複数コレクションの提供. ディジタル図書館. No.32, 2007, p.23-33.

[11] 安齋 宏幸. Z39.50 の技術解説. 情報の科学と技術. Vol.48, No.3, 1998, p.134-139.

[12] 宮 澤 彰. CQL/ZING-SRW. http://www.jsa.or.jp/stdz/instac/committe/H15report/report-contents/aidos/wg1-23/23-10.ppt（参照 2007-04-18）.

[13] 宮澤 彰. CQL と非ヨーロッパ書法の検討. http://www.jsa.or.jp/stdz/instac/committe/H15report/report-contents/aidos/wg1-26/26-05.ppt（参照 2007-04-18）.