オーディオブックを貸し出す図書館. Wired News. 2005-03-03. (online), available from <http://hotwired.goo.ne.jp/news/culture/story/20050307202.html>, (accessed 2006-03-28).

CA1596

メタサーチ技術と国立国会図書館デジタルアーカイブポータル

はじめに　～「メタサーチ技術」とは？～

インターネットの急速な普及に伴い，インターネット上を流通する情報も急激に増加している。そこで，このような情報を効率的に検索する方法が求められる。

インターネット上には，文字，画像，動画など色々な種類の情報があるが，このような情報について，探すべき対象を文字で整理，表現した「情報の情報」があると探し易くなるはずである。これは，「タイトル」「作成者」など情報の種類や内容などを特定できるよう構造化されたデータであると，より効率的に探し出せる。このような「情報の情報」を「メタデータ」と呼ぶ。「著者名」や「件名」，「一般注記」のような書誌データ，あるいは二次情報と呼ばれるものもメタデータに含まれる。

インターネット上の情報を検索するため，これまでは，メタデータ及びテキストファイルの本文の文字列を検索する「サーチエンジン」（又は「検索エンジン」）と呼ばれるプログラムを用いる方法が用いられてきたが，最近では自身の検索対象データベースを持つのではなく，ユーザが入力した検索条件を既存の複数のサーチエンジンに送信し，検索させる方法が出現した。これを「メタサーチ」(又は「メタ検索」)と呼ぶ。

本稿では，メタサーチ技術について，ごく一部ではあるが紹介し，これらを適用した例として「国立国会図書館デジタルアーカイブポータル（プロトタイプシステム）」についても簡単に紹介する。

これまでのメタサーチ技術

図書館の分野においては，まずカード目録に代わるものとして，これらの目録に記述された情報（書誌データ）の電子化がなされ，印刷版出力のためにデータを各々単独で稼働するコンピュータに入力し，検索する仕組みが備えられた。

その後，インターネットの普及に伴いユーザの PC が接続されるようになったことにより，そのようなデータを蓄積したコンピュータをサーバとし，複数のサーバ上の大量の書誌データを一度に検索したい，という要望が出てきたのは自然な流れだろう。

そこで，メタサーチの対象を書誌データとして適用することが考えられる。

米国議会図書館（LC）や OCLC などは，書誌データをもつサーバ間で目録情報を交換することを考えた。このような場合，やりとりするデータの種類，順番などをサーバ間で予め約束しておく必要があるが，この約束事（一般に「プロトコル」という）が Z39.50（CA1266，1386 参照）である。Z39.50 はその後，それぞれの機関が作成した書誌データを検索するための標準的な仕様となっていった。Z39.50 は，仕様を緻密に定めることで，単一のユーザインタフェースで様々なデータベースを対象に，精度の高い書誌データ検索を行うことが可能であった。

しかし，Z39.50 は，少なくとも日本国内では十分に普及したとは言い難い。その理由としては，以下のものが考えられる。

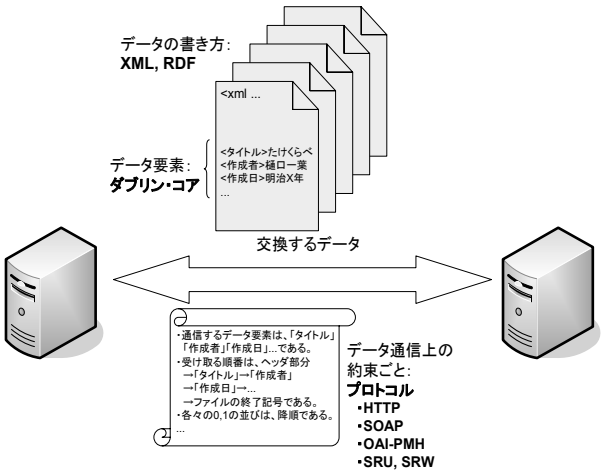
- 検索対象が図書館に特化していると思われる仕様であり，他の業界に普及しづらいものであったこと。
- 別途日本語化対応が必要であり，対応ソフトウェアの開発が遅れたこと。
- 緻密な仕様に対応したシステム構築とメンテナンスが必要であり，一定規模がないと人員面，予算面で厳しいと思われたこと。
- 分散型システムの短所として指摘されているように，各サーバの性能やネットワークの渋滞，帯域の影響を受けざるをえないこと。

最近のメタサーチに使われている技術要素

以下の説明は，厳密には正確でない箇所があるが，馴染みのない方にイメージを伝えることを意図しているため，ご容赦いただきたい。

最近では，様々なサーバや PC の間で，より自由にデータを交換するための技術が出現している。このよ

図 1　データ交換技術要素



うな技術を適用することにより，効率的にメタデータを収集したり，検索したりすることができる。

文字列で様々なデータを扱えるようにその書き方を定めているのが“XML”(eXtensible Meta Language)と“RDF”(Resource Description Framework)である。XMLはプログラムで様々に処理可能な文字列データの書き方を定義しており，RDFはXMLに追加してデータ構造の書き方や，プログラムがデータ部分を認識するための目印の名前を定義している。最近，サイトのニュースなどを自動配信する“RSS”(RDF Site Summary；CA1565参照)という技術が普及しているが，これはRDFを簡略化したものと捉えることができる。

XMLやRDFで記述するメタデータの内容については，どのようなデータでも持ちうる「作成者」「タイトル」「作成日」のような要素を含む「ダブリン・コア」と呼ばれる15要素について，合意が得られており，メタデータの記述に広く使われている。

インターネット上のデータ通信には，基本として“HTTP”と呼ばれるプロトコルが定められており，既に広く使われているが，XMLデータをサーバ間でやりとりするには，HTTPに追加して，予めデータ構造などのプロトコルを決めておく必要がある。最近普及してきたのが“SOAP”と呼ばれるものである。SOAPを用いることにより，プログラムは異なるサーバの機能を自動的に呼び出し，活用することができる。このような機能を“ウェブサービス”と呼ぶ。

さらに，メタサーチに関連したプロトコルとして，OAI-PMH（CA1513参照）や，“SRU”(Search/Retrieve URL service)，“SRW”(Search/Retrieve Web service)と呼ばれるものが出現してきている。OAI-PMHを用いると，対象のサーバから多くのメタデータを収集することができ，Z39.50の次世代として開発されたSRUやSRWを用いると，対象のサーバのメタデータを検索することができる。書誌データへの特化はZ39.50よりも緩やかであるため，他業界におけるメタデータ要素についても，より柔軟に扱うことが可能である。

以上のような技術要素は，国際標準としてISO規格で定められているか，事実上の標準(デファクト・スタンダード)として広く認知されているものである。

海外でのメタサーチ調査研究

メタサーチに関する調査研究は，海外の国立図書館や国立公文書館でも行われている。LCではSRU及びSRWの開発普及を推進しており，欧州図書館（The European Library）のプロジェクト（CA1556参照）では保有コンテンツについて検索可能なポータルサイトを構築している(1)。また，雑誌にはSRU及びSRWとOAI-PMHの親和性について議論した論文が投稿されている(2)。

「国立国会図書館デジタルアーカイブポータル」の概要

国立国会図書館（以下「当館」）では，「国立国会図書館電子図書館中期計画2004」に基づき，「デジタルアーカイブポータル」の構築を進めている。

「ポータル」とは，元々「玄関」「入口」を意味し，転じて「ここに来れば内部の各部屋を回らずとも，少ない労力で必要な手続きに案内するサービス」を指す。当館の「デジタルアーカイブポータル（以降「ポータル」という。）」は，利用者が求める様々なデジタル情報などを，ワンストップで的確に利用可能とする仕組みの構築を目指している。

従って，ポータルでは対象とするデジタル情報などを検索し提供するサービスが求められ，この機能にメタサーチ技術を適用する。

「国立国会図書館デジタルアーカイブポータル（プロトタイプシステム）」に適用したメタサーチ技術

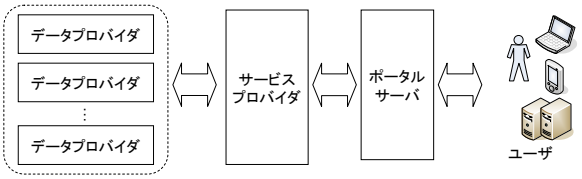
当館ではポータルの構築に先立ってプロトタイプ（実験システム）を構築し，本格稼動に必要な機能の検証を行っている(3)。本稿では，プロトタイプに適用したメタサーチ技術について紹介する。

図2　プロトタイプ画面
<http://www.dap.ndl.go.jp/>

プロトタイプは，大きく次の3種類のサーバ群から成る。

・デジタル情報を保有するデータプロバイダ
・デジタル情報などを利用可能とするサービスを提供するサービスプロバイダ
・ユーザがWebブラウザ上でサービスを利用できるようにするポータルサーバ

図 3　ポータルの概要

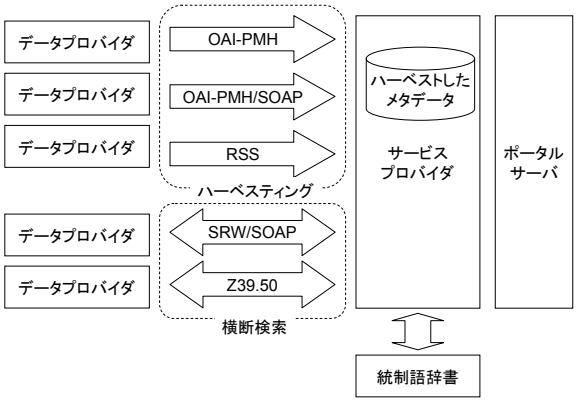


従って，メタサーチ技術が適用されるのは主にサービスプロバイダであるが，デジタル情報のメタデータなどのデータ交換を行うために，データプロバイダのサービスプロバイダとの通信機能も含まれることになる。

データプロバイダは，自身が保有するデジタル情報などを管理するためにそのメタデータを保有している。サービスプロバイダとのデータ交換に使用するメタデータ要素は，最低限ダブリン・コアに定義されたものは可能とする。また，プロトコルに関しては，次の2種類のデータプロバイダが存在する。

・大量のメタデータを機械的に収集すること（ハーベスティング）が可能な仕組みを備える。プロトコルはOAI-PMHだけでなく, SOAPやRSSの場合についても実験を行う。
・メタデータハーベスティングには対応しないが，横断検索には対応する。プロトコルはSRWを想定するが，Z39.50の場合についても実験を行う。

図4　データプロバイダとサービスプロバイダ間のプロトコル

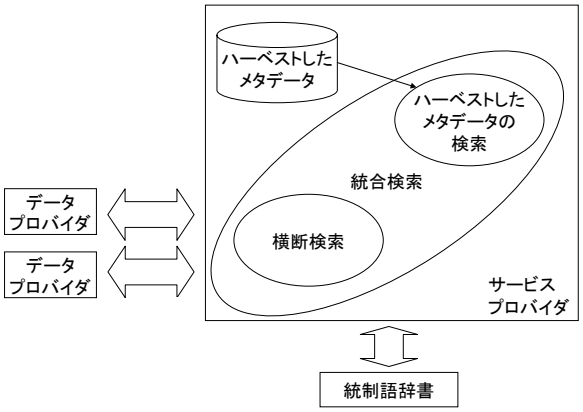


ポータルの代表的な機能として，「統合検索」が挙げられる。統合検索では，利用者が多様なデジタルアーカイブの所在，態様を意識することなく，一つの検索窓から情報を検索でき，検索結果から可能な限り，求めるデジタル情報そのもの（デジタルコンテンツ）へアクセスできるようにしている。例えば,「近代デジタルライブラリー」と「青空文庫」から，明治期の著作の画像と本文を一度に検索できる，といった具合である。メタサーチ技術を適用することで，このような機能の実現が可能となる。

この統合検索において，全文検索を行う部分は形態素解析方式とし，形態素解析を行うソフトウェアChaSenと，解析後に抽出された語からインデクスを作成し全文検索機能を提供するソフトウェアNamazuを拡張して開発を行っている。これらは実績の多いフリーウェア（無料のソフトウェア）である。

この統合検索機能の実現には，対象となるデータプロバイダから，各々が保有するデジタル情報のメタデータを検索可能とする必要がある。このため，サービスプロバイダは，前述したデータプロバイダの種類に対応して，ハーベストしたメタデータの検索と，横断検索の2種類の方法を使い分ける。そして，統合検索とは，これら2種類の方法での検索結果を統合し，結果としてユーザはこれらを同じように検索したものとして表示できるようにする機能を指す。統合検索機能ではさらに，検索結果として得た情報からデジタル情報そのものへ導くリンクを提供する。

図5　統合検索



統合検索では，統制語辞書を引き，辞書収録の語を検索の補助とする機能を備えた。

またキーワードから「連想」される語を介した検索を行う「連想検索」機能を備えた。この機能についても，フリーウェアであるGETAを用いて実現した。

現在，データプロバイダが保有するメタデータ要素は各データプロバイダに依存した定義により作成され，プロトコルも統一されていない状態である。そのようなメタデータは，他のデータプロバイダで扱われているメタデータ要素の差異を調整し，データプロバイダの違いを意識せず，統合的に扱うことが難しいものとなっている。

そこで，標準的な仕様を定め，この仕様に基づくデータプロバイダを構築することが重要な意義をもつ。この点で，当館が標準となり得る仕様を提案し，これを普及させていくことが必要と考えている。各機関，個人の御協力をお願いする次第である。

その他，統合検索に関する様々な問題の詳細は，プロトタイプに掲載しているので，ご覧頂きたい。プロトタイプにより細かな問題点が洗い出されたが，メタサーチ技術の適用により，複数のコンテンツを一元的に検索することが可能であることが検証されたことは有意であったと考えている。

おわりに

ここまで述べてきたとおり，メタサーチ技術は現在の大量のデジタル情報を検索するのに有用である。これまでにZ39.50などの開発の経緯を経ており，最近では様々な技術要素を用いることで物理的距離，業種や分野の壁，さらには時間的な距離を超え次世代の利用者をも想定し，様々な情報を検索できる可能性を模索している。

メタサーチ技術を用いた例として，当館が構築しているポータルのプロトタイプの統合検索機能を紹介した。実際に操作していただくと，メタサーチ技術を用いることで，多様なコンテンツを一元的に検索できることを体感していただけると思う。

メタサーチ技術の更なる進展，また読者や当館を含むそれぞれの活動により，ユーザがデジタル情報を一層活用できる社会が期待される。

（総務部企画課：吉田（よしだ）　暁（さとる））

(1) Veen, Theo van et al. Search and Retrieval in The European Library, D-Lib Magazine. 10(2), 2004. (online), available from <http://www.dlib.org/dlib/february04/vanveen/02vanveen.html>, (accessed 2006-05-15).
(2) Sanderson, Robert et al. SRW/U with OAI, D-Lib Magazine. 11(2), 2005. (online), available from <http://www.dlib.org/dlib/february05/sanderson/02sanderson.html>, (accessed 2006-05-15).
(3) デジタルアーカイブポータル（プロトタイプシステム）. (online), available from <http://www.dap.ndl.go.jp>, (accessed 2006-05-15).

CA1597

動向レビュー

電子ジャーナルのアーカイビング
―海外の代表的事例から購読契約に与える影響まで―

1. はじめに

電子ジャーナルの興隆とともに，その保存の重要性を裏づけるデータが増えている。例えば，7,000 人以上の研究者を対象とした米国の調査では，将来の利用に備えて電子ジャーナルを保存することが「非常に重要である」と回答した者が 83%に上ったという(1)。

ところが，電子ジャーナルは保存性に難点があるのが実情だ。そのひとつがライセンス契約による利用形態である。購読者は，出版社のサーバにアクセスするライセンスを取得するだけであり，基本的に購読コンテンツを保存することはできない。それは出版社頼みになる。万一倒産や企業合併，災害に出版社が見舞われた場合は，電子ジャーナルの保存を保障するものは何もないことになる。

よって，電子ジャーナルについては，図書館による購読と保存が一続きであった従来の資料と異なり，アクセスを保障するための保存（以下ではこれをアーカイビングと呼ぶ）体制を意識的に設ける必要がある，と言える。本稿ではまず，代表的な取り組みを概観する。その後，上で述べたような万一の事態への備えとしてだけでなく，望ましい内容の契約を購読者が出版社と結ぶためにも電子ジャーナルのアーカイビングは重要である，という点も確認しよう。

2. 事例

2.1. オランダ国立図書館

はじめにオランダ国立図書館（Koninklijke Bibliotheek：KB）の取り組みに言及したい。同図書館は複数の出版社（オランダの国外に本社を置く出版社を含む）と協定を結び，電子ジャーナルの公的アーカイブ機関の役割を担っている。

アーカイビングの流れは次のとおりである。まず，出版社が電子ジャーナルを KB に無償で提供する。続いて KB が，電子出版物のアーカイビングシステム「e-Depot」（KB の依頼を受け，2002 年に IBM 社が開発）を用いて，アーカイビング業務を遂行する。かかる経費はオランダ政府が拠出している。しかし，将来は（ちょうど，包装紙の費用を商品の価格に含める場合があるのと同じ具合に）アーカイビングのコストを電子ジャーナルの購読料金にあらかじめ算入する必