メディア教育研究　第1巻　第2号
Journal of Multimedia Aided Education Research 2005, Vol. 1, No. 2, 169－180

研究展望

Web上の高等教育用コンテンツの統合検索の課題と効果的検索への提案

篠原　正典[1)]・地蔵　真作[2)]

高等教育へのe-Learningの普及と共に、教育用コンテンツの共有・流通が世界規模で進められようとしている。そこでは、Web上に存在する有用なコンテンツの統合から、それらを検索するシステムの統合まで検討が始まった。この機に先んじて、本稿ではWebマイニングおよびメタデータを利用したコンテンツ統合から、Webサービスを利用した検索システムの統合に関して、現況と共に、近未来に予測される形態や課題について概観し、一部これらの技術の応用に際し効果的な検索を実現する提案を述べる。

キーワード
Webマイニング、セマンティックWeb、Webサービス、メタデータ、統合検索、教育用コンテンツ

1．はじめに

インターネット上には数十億ページという膨大な数の情報が掲載されているが、利用者にとってはあらゆる分野の情報が必要なわけではなく、自分のニーズに適合した情報が得られることが重要である。検索システムを使って情報を探すことが一般的なWebの利用法であるが、情報量が膨大となった現在ではWebページから必要とする的確な情報だけを収集した統合検索という考え方が注目されてきている。統合検索を情報統合からシステム統合まで広義に捉えると①既存のWebサイトから関連する情報のみをクローリングにより自動収集、あるいは収集した情報から必要な情報のみを自動分類して抽出するといったいわゆるWebマイニングの分野の研究、②関連する情報にメタデータを付与して、メタデータを元に情報を横断的に検索するいわゆるセマンティックWebに関連する研究、③検索エンジンを横断的に統合するいわゆるWebサービスに関連する研究が進んでいる。

Webマイニングは膨大なWeb情報から、ある特定の分野に関する情報を自動的に収集し、的確な検索を可能とするための手法であり、ベクトル空間に基づく方法、情報の規則性に基づく方法、識別学習に基づく方法など様々な自動情報分類の研究が行われてきている[5]。既に論文情報を集めた検索[6]やニュース記事を統合したサイト(注1)など、インターネット上でサービスに至っているものもある。セマンティックWebはWebの創始者であるTim Berner-Lee氏によって1998年頃から提唱されてきたものであるが、その全体像は階層構造を成し[2]、その基盤に位置するものはメタデータ技術である。メタデータの応用に関しては、既にRSS（Resource Description Framework Site Summary、あるいはRich Site Summary）のように、可能な限り簡略化して、必要なメタデータのみをXML（Extended Markup Language）で記述したものが急速な普及を見せ始め、新聞記事(注2)やWeblog(注3)などを中心として情報の統合検索が行われている。一方、Webサービスは2000年に使われだした言葉で、それまで存在していた分散プログラミング環境に類似した考え方であるが、セマンティックWebと同様にXMLを使うことにより、これまで、同じ言語を用いたプラットフォーム間の結合のみが可能であったものを、異なるプラットフォーム、例えばWindows系とUnix系の結合をも可能にしたことに大きな特長がある。

上記3つの技術については、現在でも活発に研究が進められていると共に、応用も実現されてきている。それぞれ特徴や利用形態が異なるため、いずれかの技術が他を淘汰するものではない。これらの技術の対象は特定の情報分野を対象とするものではないが、本稿では、現在、世界規模でWeb上の教育用コンテンツの統合化の検討

1) メディア教育開発センター
2) （有）リアクト

(注1) 例えば、http://news.google.com
(注2) 例えばhttp://www.asahi.com/information/rss
(注3) 例えば宮川達彦、伊藤直也、Blog Hacks－プロが教えるテクニック＆ツール100選、オライリ・ジャパン（2004）に技術面など体系化された内容が記載されている

が進められている(注4)ことや、国内でも教育用コンテンツの共有と流通の必要性が叫ばれていることから、特に高等教育で役立つコンテンツ(注5)の統合化に焦点を当て、上記の3つの情報統合技術について特徴や現況を概観すると共に、将来想定される形態について課題を述べる。また、特にWebマイニングに関しては、筆者が実際に分析した一部結果を基に、技術的な問題を抑えた効果的な情報収集・検索に対する提案も含める。

2．Webページの自動収集による情報統合

(1)　Webマイニングの応用

2003年で80億ページと見積もられるWebページ上の情報を検索するツールとして、間接的な世界シェアまで含めると70％を超えると言われる[11]Google(注6)などの優れた高速検索システムの利用により、即時に検索結果を得ることが可能となっている。Googleは広範囲な領域をまたがって情報を検索するのに非常に有効な検索システムである。200以上のサーバで動くクローラ（ロボットあるいはスパイダーとも呼ばれる）が世界中のWebサイトから約33億ものURLを収集し、全世界11箇所に散らばる2万台のPCで、1日あたり2億件の検索を処理している[13]ことから、そのシステムの性能の高さが伺い知れる。検索結果の表示順でも独自の方法であるPageRank[1, 7]というページへのリンク数による重み付けを考慮した計算手法により、確度の高い検索結果が得られるように工夫がなされている。しかし、それでも検索対象が膨大であるため、必ずしも検索者の希望に即した的確なページが検索結果の冒頭部分に出てくるとは限らない。例えば、人工知能に関する論文を検索したい場合に、「人工知能」&「論文」のキーワード入力で検索しても、検索システムはキーワードの「論文」を論文形式を意味するものとして認識しているわけではないため、目的と合致しない検索結果が多く出てくる。入力キーワードを工夫すれば適合率を向上できるが、工夫は検索者に依存する。しかし、これが「論文」だけを集めたカテゴリの中で「人工知能」というキーワード検索を行えば、目的と合致する結果が出る確率は非常に高い。

論文に限らず、ある分野に属する情報を人的に収集することは可能であり、カテゴリ検索と言われる検索法の中で、これまでも実際に行われている。しかし、情報量が増え続けていることやURLや情報の内容自体が変化することを前提としているWebページでは、収集や収集したデータの更新や修正を人手で行うことは容易ではない。この状況を改善すべく登場したのが、膨大な情報から計算機を用いて特定の分野の有用な情報だけを取り出そうとするWebマイニングである。Webページの収集においては、Googleにも使われているクローラと呼ばれるソフトウェアにより、インターネットから自動的にWebページを収集する方法が用いられる。この時、予め収集時に目的に適合する収集条件を基にフィルタリングして、関連するWebページだけを収集する方法と、対象となるWebサイトから全ての情報を収集した後、設定した条件でフィルタリングして必要なWebページのみを抽出する方法がある。

高等教育用のコンテンツ収集においてもWebマイニングの研究は進められており、一例として大学のシラバスの自動収集の報告[3, 12]がある。これは収集時にフィルタリングして条件に合致したページのみを収集する方法を取っている。htmlで記載されたWebページにはリンクや語彙などの特徴的な構造を有するものがあり、シラバスページは「科目」、「担当者」、「講義目標」など、ある程度共通した語彙項目を持ち、かつ同じシラバス群では固定の様式を有するという特徴がある。報告[3, 12]ではシラバスのページのhtmlを分析して、共通語彙から構造を抽出し、またシラバス一覧ページとそこからリンクされる個々の科目を説明するページを評価して、シラバスページであるか否かを決定している。そして、4つの大学におけるシラバス収集を行い、26,270件のシラバス収集に成功している。

このようにシラバスや冒頭で述べた論文など、一部の分野に関してはWebページの構造から規則性が抽出されてはいるものの、規則性は対象とするコンテンツの分野に依存することから、個々の分野において異なる研究アプローチが必要となる。そして、該当する分野の情報を網羅する割合（再現率）と、収集した情報が的確である割合（適合率）を如何に高くできるかが課題となる。これらは相反する条件であるため実現は容易ではない。

(2)　URL表記に対する一つの提案

前述したようにWebページの構造性から規則性を抽出することは容易ではなく、ましてや構造のないWebページに対しては難易度がより高くなる。そこで、本章ではメタデータがデータの内容を明示するのと同じように、WebページのURLがそのページの内容を明示する方法を提案する。シラバスや統計資料、あるいは研究内容や研究成果、教材などは個々のサイトから検索して得るよりは、それらのコンテンツを統合して検索した方が効果的であることはいうまでもない。このように統合化することにより利用効果が高まるコンテンツに対して、そのWebのURL内にページの内容を特徴付ける語彙を付記することを提案する。この語彙は他の情報と明らかに混在せずに利用でき、かつ、情報の内容を特徴付ける

(注4) 例えばMerlotが提供している統合検索サイトhttp://fedsearch.merlot.org/main/search.jsp や CORDRA（Content Object Repository Discovery and Registration/Resolution Architecture）のサイトhttp://cordra.lsal.cmu.edu/cordra/を参照

(注5) 本稿では教材だけでなく、研究成果や統計資料、シラバスなどの情報を総称してコンテンツと表記する

(注6) http://www.google.co.jp/

語彙である。例えば、「syllabus」という語彙は、語彙自体がシラバスそのものを意味する語彙であることから、シラバスのWebページを特定する語彙に該当する。現在でもGoogleのようにURLに含まれる語彙からそのWebページを検索することが可能であるように、URLに「syllabus」を含むWebページを抽出することは容易に成し得る。つまり、それがシラバスページの抽出に相当するわけである。さらに抽出したWebページを対象としたキーワード検索を実現すれば、分類化された中からの検索が可能となる。

実際に、Webページの内容を意識したURL表記がどの程度なされているかについて参考までに調べてみた。

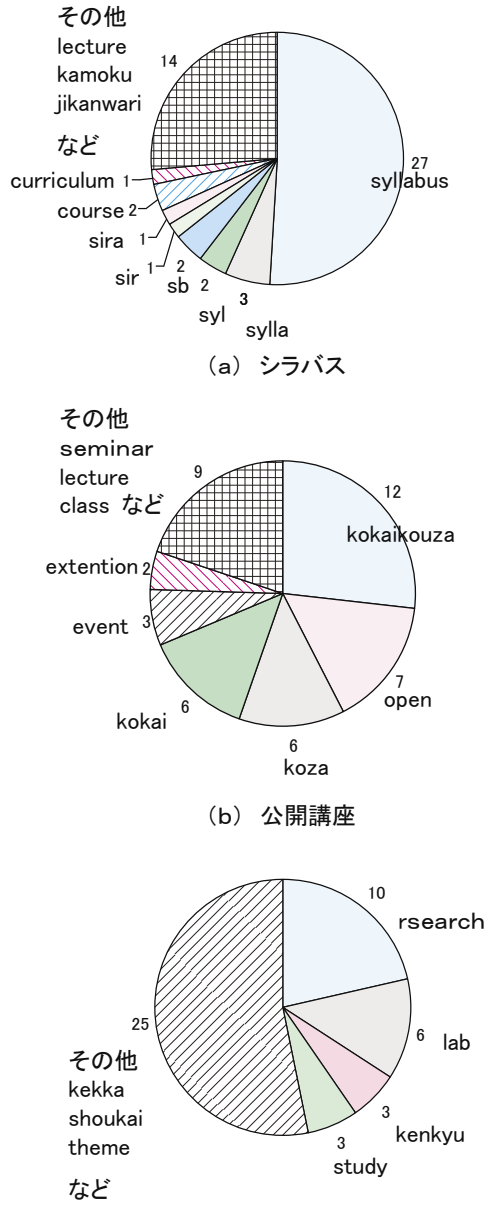


図1　シラバス、公開講座、研究内容のページのURLに含まれる語彙

調査対象は全国の国立の大学法人39校のWebサイトである。対象とした情報は横断的に統合して検索することに対して効果的と思われる一部であるシラバス、公開講座、研究内容である。これらのWebページで特徴的と思われる語彙と、それらが含まれる割合を図1に示す。調査対象とした大学の全てがこれらの情報を掲載しているわけではなく、また大学の中には、大学院、学部、学科等によってWebページに含まれる語彙が異なる場合があるが、図のそれぞれに記載されている数は調査対象機関の中で上記の情報の掲載が確認されたサイト数[注7]を示している。図1の結果を見て分かるように、シラバスのページでは、比較的その特徴を現していると思われる「syllabus」、「sylla」、「syl」などの語彙を含むサイトが約70％存在するが、その他の分野の情報に対しては様々であり、特徴的な語彙を使った傾向は見当たらないというのが現状である。

本提案は、特徴的な語彙を該当するURLの中に含めることであり、Webサイトの構造は個々の機関により様々であるため、どの階層に含めるかは自由で構わない。少なくともURLのどこかに特徴的な語彙が含まれていればよい。ただし、下記に例を示すようにWebページが階層構造を成す場合には、その構造を反映することが望ましい。表1は横断的に統合して検索することにより、その効果的な検索が期待できる情報に対して、含めるべき語彙を提案したものの一部である。この中でシラバスや研究成果などはさらに科目別に細分化されるであろうから、例えば日本10進分類法、あるいは本稿の「2. メタデータを用いた統合検索」で使う分類メタデータに即した語彙をURLに付記することにより、より詳細な分

表1　URLに付記する語彙案

Webページ内容	URL付記語彙例
シラバス	syllabus
研究内容	research-subject
研究成果	research-result
公開授業	extension-course
学習教材	teaching-material
プレスリリース	press-release
研究会・講演会	workshop
統計資料	statics
報告書	report
調査報告書	survey
白書	white-paper

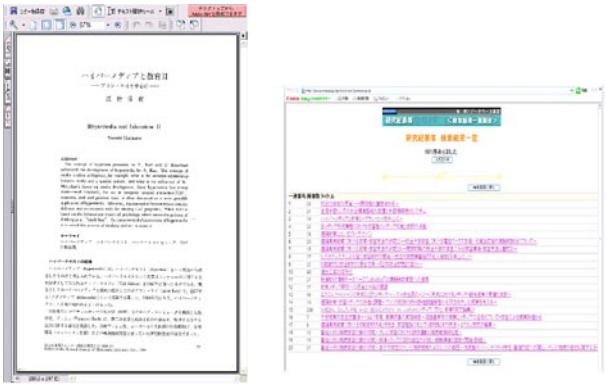
[注7] 例えば調査した範囲内で大学のシラバスページに記載されているURLにsyllabusなどの特徴的な語彙が1種類である場合の数は1、一つの大学でも学部によってsyllabusやsirabasuなど2種類の特徴的な語彙が使われている場合の数は2、特徴的な語彙がない場合はその他として1カウントとしている

類化にも対応できる。例を挙げれば、科目分類が物理学すなわちphysicsであるとすれば、http://www.AAA.ac.jp/syllabus/physics/……/というURL表記は「syllabus」のディレクトリに含まれる「physics」ということから物理学のシラバスに属することを示す(注8)。この方法はURLに特徴的な語彙を含めるだけでよいため、同分野の情報を分類化するのに少ない稼動で情報の統合を可能とするものである。しかし、本提案を実現するには本稿の2章で述べるメタデータ付与と同様に、URLに表記する語彙の標準化とWebサイト運用者の協力を得る必要があるという課題は残る。

(3)　検索結果表示に対する一つの提案

前節は特定のWebページを分類するための提案であるが、ここでは、検索結果として得たWebページを利用者にとってさらに理解しやすくするための方法を提案する。高等教育で役立つ情報と言えるシラバス、研究成果、統計資料、論文など、多くの情報は、今ではWebの中ではhtml形式だけでなくpdfやxlsあるいはppt、docなどの形式で書かれている場合が多くなっている。これに呼応して、これらのファイルを直接検索できるようにGoogleではファイル形式を指定して検索することが可能となっている。すなわち、例えば、検索時に国の省庁Webサイト(注9)からxlsファイルを指定することにより、高い適合率で統計データが検索できる。これらのファイル形式のページはWebサイトのディレクトリ構造上では、末端部(注10)に位置する場合が多い。

図2(a)に一例としてメディア教育開発センターのWebサイト(注11)の中に存在するpdf形式のページを示す。ブラウザ上に表示されたpdfのページ内にはリンクが無く、これより先にリンクで辿ることができない構造となっている。このようにpdfやxlsファイルは一般的にリンク情報をページ内に含まない(注12)。そのため、トップページから順にページを辿ってpdfページに辿り着いた場合には、ブラウザの戻りボタンでpdfページのリンク元のページに戻ることはできるが、Googleなどの検索システムを使ってpdfやxlsなどを検索した場合を想定すると、URLを直接アクセスしていることからブラウザの戻りボタンは効かず、またページの中にはリンク情報が含まれないため、検索後このファイルから先にリンクを辿れないという問題が起こる。



(a) Pdfページ　　(b) (a)のpdfページへのリンク元ページ

図2　pdf形式のページとそのリンク元のページ

pdfやxlsのページはそれ自体がコンテンツそのものであるため、有益な情報が含まれている可能性が高いが、そのコンテンツがどのようなコンテンツの一部であるのかを判断することが難しい場合がある。特に表などのデータはデータだけを見ても内容が理解し難いものがある。このようなときに理解を助ける情報を与えてくれるのが、そのページのリンク元となるページである。そのページの情報を得て、初めて書かれた内容が理解できるものも多々ある。図2(b)に図2(a)のpdfページのリンク元のページを示すが、このページには多くのpdfページの一覧が掲載されており、図2(a)のpdfがどのような情報群の一部であるか、また図2(a)以外の関連する情報の存在を知ることができる。このようにリンク元ページはxlsやpdf形式のデータ群をまとめた一覧ページや目次であったり、あるいはそのデータの群の説明を含むページであったりするなど、有用な情報を含んでいる。

pdfやxlsのURLを元に、それらのページを掲載している機関名を推定することは比較的容易である。例えば、http://www.nime.ac.jp/it2004/img/program.pdfというpdfファイルがあった場合、このURLから、このpdfファイルが掲載されているWebサイトは冒頭の階層のURL、すなわちこの場合にはhttp://www.nime.ac.jp/をアクセスすれば、通常はWebサイトのトップページが得られ、先のpdfファイルの掲載機関が分かる。ところが、このページはpdfやxlsなどのページをディレクトリ構造内のどこかに含むトップページを示すものの、一般的にはそれらとのリンク関係はない。また、http://www.nime.ac.jp/it2004/img/program.pdfの上位階層、すなわちhttp://www.nime.ac.jp/it2004/img/がpdfのリンク元となる確率は低い。このようなことからリンク元のページをpdfやxlsページのURLから推定することは非常に困難である。

(注8) 厳密に議論すればsyllabusが記載されたページにはシラバスそのもののページの他に、シラバスの一覧を表示したページ、さらにはシラバス作成に関する理念を記述したページなどがあるだろうが、これらも重要な情報を含むと判断すれば同一カテゴリに含めればよく、さらに細分化したい場合には一覧ページをsyllabus-catalogueなどと細分化することも可能である。いずれにしても2章で述べるメタデータ同様に標準化していくことが重要である

(注9) URLにgo.jpを含むものを指定して検索

(注10) 本稿ではこれよりさらにリンク先がないという意味で使用している

(注11) http://www.nime.ac.jp

(注12) pdfの中には同じファイル内のページにリンクする目次を有するものはあるが、他のファイルやURLの異なるページへのリンクはほとんどない

表2　pdfやxlsページとそのリンク元ページ例

PDF、XLS、CSV形式のページURL	左記のリンク元ページURL
http://www.zaimu.japanpost.jp/tokei/pdf/sogo0202.pdf	http://www.zaimu.japanpost.jp/tokei/soindex.html
http://www.nta.go.jp/category/toukei/tokei/jikei/1594/deta/15.xls	http://www.nta.go.jp/category/toukei/tokei/jikei/1594/01.htm
http://www.meti.go.jp/statistics/downloadfiles/ha2gsm2j.csv	http://www.meti.go.jp/statistics/data/h2afdldj.html
http://www.zaimu.japanpost.jp/tokei/1989/excel/ca890003.xls	http://www.zaimu.japanpost.jp/tokei/1989/kw89.html
http://www.mhlw.go.jp/toukei/itiran/roudou/chingin/jittai/02/xls/fuhyo.xls	http://www.mhlw.go.jp/toukei/itiran/roudou/chingin/jittai/02/index.html
http://www.mhlw.go.jp/toukei/itiran/roudou/chingin/kouzou/02/xls/12.xls	http://www.mhlw.go.jp/toukei/itiran/roudou/chingin/kouzou/02/index.html
http://www.meti.go.jp/statistics/downloadfiles/h2flc171j.pdf	http://www.meti.go.jp/statistics/data/h2flc01j.htm
http://www.meti.go.jp/statistics/kougyou/1998/k1/h10k4nin.xls	http://www.meti.go.jp/statistics/kougyou/1998/k1/index.html
http://www.meti.go.jp/statistics/kougyou/1998/k5/h10-bun.pdf	http://www.meti.go.jp/statistics/kougyou/1998/k5/h10-bun.pdf
http://www.meti.go.jp/statistics/tokusabi/2001k/h13-t-02.xls	http://www.meti.go.jp/statistics/data/h2v2000j.html http://www.meti.go.jp/statistics/data/h2v2001j.html
http://www.nta.go.jp/category/toukei/tokei/menu/chousyu/h11/data/01.pdf	http://www.nta.go.jp/category/toukei/tokei/menu/chousyu/h11/data/01.pdf

そこで、本稿では検索結果に表示された個々のページ(注13)から、そのページのリンク元になるページに移動できる方法を提案する。Webページを収集するクローラはWebページからリンク情報を辿ってリンクされたWebページを順次収集していく。このとき、リンク元のURLとそこからリンクされたページのURL情報をデータとして有している。表2は例として日本の省庁のWebサイトにあるpdfやxlsのページとそのリンク元となるページのURLを示したもので、このようなURLのリンク関連情報をクローラで収集したデータから作成し、データベースとして検索システムに蓄積しておく。このデータベースを使うことによりpdfやxlsのページを検索した後にリンク元ページに戻る方法が実現できる。

実際に検索システムに機能開発する場合を考えてみる。表2を見るとpdfやxlsページにリンクしているページには単数のものと複数のものが存在する。単数の場合にはリンク元ページを一意に決定できるため、リンク元ページへの戻り機能開発に問題ない。しかし、複数のリンク元ページが存在する場合には、どのURLをリンク元ページとして同定するかを決めておかないと機能を具現化できない。そこで、リンク元ページを同定するために、実際のWebサイト上でページのリンク関係を調べてみた。国内の政府の11機関のWebサイトに掲載されているpdf、xlsおよびdoc形式の統計資料約34,000ページのリンク分析を行った(注14)結果を表3に示す。これらのファイル形式のリンク元ページはほとんどが単数であることがわかる。この場合にはリンク元ページを一意に同定できる。しかし、割合は少ないものの表3に見られるようにリンク元ページが複数存在するものもある。そこで、表3のリンク元ページが複数存在する場合において、的確なリンク元と判断されるページが存在するディレクトリを調べた結果を表4に示す。同一フォルダ内に存在する場合や、1階層上のディレクトリに存在する場合、異なる階層に存在する場合などがあることがわかるが、上記の調査範囲内のページにおいては、全て同一サイト内に存在することがわかる。すなわち、同一Webサイト配下のディレクトリ内に存在するページであれば、どれでも的確なリンク元ページであると判断できることを示している。これらの規則性を元に、例えば、pdfやxlsなどを検索結果として表示させた後、「リンク

表3　xls、pdf、doc形式ページにリンクするページの数

ファイル形式	リンク元ページ総数	リンク元ページが単数であるもの	リンク元ページが複数あるもの	リンク元ページが単数である割合
xls	31071	30959	112	99.6%
pdf	2830	2818	12	99.6%
doc	36	36	0	100%

表4　リンク元ページが複数存在する場合における的確なリンク元ページの位置

ファイル形式	リンク元ページが複数存在する数	的確なリンク元ページが存在するディレクトリ			
		同一フォルダ内	1階層上のフォルダ内	同じサイトの他の階層のフォルダ内	的確なリンク元ページが同一サイト内に存在する割合
xls	112	96	4	12	100%
pdf	12	0	0	12	100%

(注13) 上記までの説明では個々のページをpdfとxlsファイルを中心に説明したが、htmlやjpeg、imageなどどのような形式のページでも構わない

(注14) Webサイト構造を分析するツールWebExplorer http://www.vector.co.jp/soft/win95/net/se247055.htmlを利用して分析した

元ページへの移動」なるボタンを検索結果画面と異なるフレーム内あるいは別画面で表示させ、そこから表2に示すリンクデータを基に、リンク元ページへ移動させることは容易に実現できる。

３．メタデータを用いた統合検索

①メタデータによる統合

メタデータはセマンティックWebの中でも基盤となるもので、コンテンツから特定の情報を抽出するために付与する項目や語彙のことである。例えばメタデータの中には「タイトル」や「概要」といった項目が定義され、それらはXMLで記述される。そのため、「タイトル」というメタデータタグのついた語彙のみを検索することにより、その情報につけられた題名からの検索といったことが可能となる。あるいは「著作権」に関する情報だけをメタデータを利用して表示させることもできる。統合化においては共通のメタデータ利用が必須であることから、その項目やXMLによる表記法（スキーマ）が標準化されてきている。

この技術は、メタデータのみが統合検索システムに登録され、メタデータが付与されるコンテンツは外部からアクセス可能であれば、インターネット上のどこのサーバに置かれてもよいという特徴を有する。メタデータを利用した検索では、基本的にはメタデータに書かれた内容を検索して、メタデータの中に含まれるURL情報を元にコンテンツへリンクされる仕組みになっている。そのため、コンテンツそのものは著作権を有する個々の運用者のサーバに格納されるため、統合検索側による著作権侵害の問題が避けられ、また、メタデータは単なるテキストデータであるため、データベース容量が肥大するというシステム面の問題も小さいという利点がある。さらにメタデータを利用することにより、本来検索対象とする情報に含まれない著作権の利用許諾に関する情報や、情報間の相互関連などの情報をメタデータの中に組み入れることにより、検索利用者に付加的な情報を与えることができる。これはWebマイニングではできない点である。メタデータを付与する元となる情報は学習教材など比較的改変頻度の少ない情報が多い(注15)一方で、簡易なメタデータで構成されるRSSを使って、新聞記事やWeblogなど頻繁に内容が更新される情報にメタデータを付与することにより、日付別やタイトル別、あるいは内容別に検索することが可能となっている。

メタデータの中でも例えば、教育分野のメタデータはLOM（Learning Object Metadata）と呼ばれている。日本国内の初等中等教育においては既にLOMが作られ[3, 9]、実際に教育情報ナショナルセンターのポータル検索サイト(注16)で運用されている。その中では、タイトル、キーワード、対象とする学習者（学年など）、利用言語、概要など必要とされる項目が設定されている。著者が所属するメディア教育開発センターでも高等教育用のコンテンツをLOMで横断的に検索するシステム[9]を構築中である。

②LOM登録の課題

メタデータ付与はコンテンツの特徴を表す的確なデータとなり得、かつ標準化されたXMLで表記されるため統合化に有用な方法であるが、課題もある。裏返せば、メタデータは的確な内容を示す必要があること、また、メタデータには多くの項目が含まれているため、それを人手で付与することはコンテンツ保有者に負担をかけることになる。これが普及を妨げる要因ともなる。そのため、全ての項目を付与するのではなく、LOMにおいても付与すべき必須項目と推奨項目、またその他の項目に分けられており、付与の負担を抑えられるようにしている。

表5にIEEE LOM Ver1.0(注17)の内容とメディア教育開発センターが行おうとしているLOMの必須項目と推奨項目およびその他の項目[9]を参考に示す。メタデータによる検索がコンテンツのテキスト全文に対する全文検索に勝るところは、メタデータに書かれたキーワードや焦点が明確に記述された概要などを検索対象とすることにより、ニーズに的確なコンテンツの検索を可能にすることや、全文検索では得られない情報を利用者に与えてくれるところである。対象としている利用者が小学生や大学生かあるいは成人かなどについては5.6 Education-Contextや5.7 Education-Typical Age Rangeで判断でき、著作権に関連する利用許諾範囲なども6.3 Right-Descriptionの内容で判断できる。欲しいコンテンツが図やグラフか、試験か講義かなどのタイプを5.2 Education Learning Resource Typeで条件選択することも可能となる。今後、高等教育機関においてITを活用したe-Learningの普及が進み、教育用コンテンツの作成とその統合化が必要となるであろうが、その場合には、一つ一つのコンテンツが学習素材として取り扱われるものや、コンテンツが教育コースの一部になっているものなどの効果的な検索も考慮する必要が出てくる。特にコースの中の一部コンテンツ中には、決められた学習プロセスに沿って学習するものがあり、コースの中のコンテンツ間の強い関連性が重要となり、それが利用者にもわかるように、7.2.2 Relation-Descriptionなども推奨項目として含める必要が出てこよう。

手動でのLOMの付与稼動を抑える一方で、コンテンツへのLOMの自動付与の研究もなされている。初等中等教育における教育用コンテンツに対して「1.2 Title」、

(注15) 国内では初等中等教育において教材用コンテンツとして政策的に作られた教育用動画像集などに付与されている
(注16) http://www.nicer.go.jp/
(注17) http://ltsc.ieee.org/wg12/

表5　メタデータの比較

IEEE LOM Ver 1.0			Dublin Core	RSS
1		General		
1.1		Identifier	Identifier	
1.1.1		Catalog		
1.1.2		Entry		
1.2	必須	Title	Title	Title
1.3		Language	Language	Language
1.4	推奨	Description	Description	Description
1.5	推奨	keyword	Subject	
1.6		Coverage		
1.7		Structure	Coverage	
1.8		Aggregation Level		
2		Life Cycle		
2.1		Version		
2.2		Status		
2.3		Contribute		
2.3.1		Role		
2.3.2		Entity	Contributor, Creator, Publisher	Publisher, Creator
2.3.3		Date	Date	Date
3		Meta-metadata		
3.1	必須	Identifier		
3.1.1		Catalog		
3.1.2		Entry		
3.2		Contribute		
3.2.1		Role		
3.2.2		Entity		
3.2.3		Date		
3.3		Metadata Scheme		
3.4		Language		
4		Technical		
4.1		Format	Technical Format	
4.2		Size		
4.3	必須	Location	Identifier	Link
4.4		Requirements		
4.4.1		Or-Conposite		
4.4.4.1		Type		
4.4.4.2		Name		
4.4.4.3		Minimum Version		
4.4.1.4		Maximum Version		
4.5		Installation Remarks		
4.6		Other Platform Requirements		
4.7		Duration		
5		Education		
5.1		Interactivity Type		
5.2	推奨	Learning Resource Type	Type	
5.3		Interactivity Level		
5.4		Semantic Density		
5.5		Intended End User Role		
5.6	推奨	Context		
5.7	推奨	Typical Age Range		
5.8		Difficulty		
5.9		Typical Learning Time		
5.10		Description		
5.11		Language		
6		Rights	Rights	Right
6.1		Cost		
6.2		Copyright and Other Restrictions		
6.3	推奨	Description		
7		Relation	Relation, Source	
7.1		Kind	Relation	
7.2		Resource		
7.2.1		Identifier		
7.2.1.1		Catalog		
7.2.1.2		Entry		
7.2.2		Description	Relation	
8		Annotation		
8.1		Entity		
8.2		Date		
8.3		Description		
9		Classification	Subject	
9.1		Purpose	Subject	
9.2		Taxon Path		
9.2.1		Source	Subject	
9.2.2		Taxon	Subject	
9.2.2.1		Id		
9.2.2.2		Entry	Subject	
9.3		Description		
9.4		keyword	Subject	

本表には記載していないが、メディア教育センタではIEEE LOM Ver.1.0に含まれないサムネールも推奨項目としている

「1.3 Language」、「1.4 Description」、「1.5 Keyword」、「1.7 Structure」、「3.3 Metadata Scheme」、「4.1 Format」、「4.2 Size」、「7 Relation」、「9.2.1 Classification」の項目におけるLOMの自動付与技術の開発も行われている[4]。しかし、残念ながら自動付与の精度はまだ低いという結果を示している。さらにこの研究の対象に含まれていない著作権情報や教育分野、対象年齢などの情報までも含めて自動的に付与することを想定すると、かなり難しくなる。これらの情報はコンテンツから自動的に抽出することが難しいために、人的に処理せざるを得ず、自動化にも限界がある。

上記のようにメタデータ登録における稼動の問題があ

るにも関わらず、一方では、メタデータを利用するRSSが新聞やWeblogを中心に急速な広まりを見せている。RSSはそれだけに留まらずその他の幅広い分野でも使われ、RSSで配信されているリストまでもできている(注18)。この急速な普及の理由は、メタデータが簡素化されていることや、RSSの構造を知らなくても自動で作成するツール(注19)や、RSSで記載されたコンテンツを収集して読み出すツールが多数登場し、かつメタデータで設定された内容だけでも十分に満足する利用者が多いためである。そこで、メタデータ付与の簡便さの必要性から、LOMをRSSのように取り扱うことができないかを考えてみる。RSSは基本的にはタイトル、アドレス、見出し、要約、更新時刻などのメタデータがつけられ、多数のWebサイトを横断的にしかも効率的に把握できるXMLのフォーマットである。RSSで使われているメタデータを参考までに表5に併記する。一見してもわかるようにRSSは非常に簡易なメタデータの使用に留まっているのがわかる。しかし、実はRSSは簡易なメタデータを使うだけではなく、同様な内容を並列に記述できる特長がLOMと大きく異なっている。

図3にRSSを使ったメタデータの一部を示すが、〈Channel〉タグで示す情報の提供元や大枠の情報を示す他に、個別の情報を〈Item〉タグの中に併記できるようになっている。そのため、新聞記事や日記情報などのように複数の同様な内容を併記でき、また、発行日などのデータを入れることも可能であるため、日ごとあるいは時間ごとに更新されるデータに対応できるようになっている。また、図3のメタデータに固定されているわけではなく、〈Description〉などの中にサブカテゴリなどを設けることも可能である。

```
＜Channel＞
  ＜Title＞タイトル＜／Title＞
  ＜Link＞リンク先URL＜／Link＞
  ＜Description＞概要説明＜／Description＞
  ＜Language＞言語＜／Language＞
  ＜Copyright＞著作権者＜／Copyright＞
  ＜Publisher＞発行者＜／Publisher＞
                                  } 下記の個々の情報提供に関するデータ
  ＜Item＞
    ＜Title＞タイトル＜／Title＞
    ＜Link＞リンク先URL＜／Link＞
    ＜Description＞概要＜／Description＞
    ＜PubDate＞発行日＜／PubDate＞
    ＜Category＞ジャンル＜／Category＞
                                  } 個々のデータ 例えば 新聞記事1
  ＜／Item＞
  ＜Item＞
    ＜Title＞タイトル＜／Title＞
    ＜Link＞リンク先URL＜／Link＞
    ＜Description＞概要＜／Description＞
    ＜PubDate＞発行日＜／PubDate＞
    ＜Category＞ジャンル＜／Category＞
                                  } 新聞記事2
  ＜／Item＞
```

図3　RSSを使ったメタデータの例

それではLOMをRSSに対応させることは可能か、あるいは効果があるかについて検討してみる。RSSが本来簡易なメタデータで構成されていることから、LOMを適用する場合でも、必須あるいは多くても推奨項目までを対象とすればRSSのような構造にすることは可能である。特に複数のコンテンツからなるコースを登録する場合にはRSSのようなメタデータ構造は有効であるように思える。と言うのは一つのコース教材の中に複数の教材が含まれるわけであるから、それらを〈Item〉の中に登録すれば、コースとしての全体の構造および個々の教材の存在も理解しやすくなる。また、例えば株価や為替、さらには日別や月別で変わる統計データなどを利用した教育コンテンツを対象とした場合には、時間的要素をメタデータに入れることができるためRSSを利用する利点は大きいと思われる。現時点では、表5に示すようにIEEE LOM Ver1.0の中のいくつかの項目で教育用コンテンツを登録することに対して問題は無い(注20)ことから、RSSを導入する必要性は無いが、今後、前述したような教育用コンテンツまで対象を広げると、LOMにRSSの思考を取り入れた検討も重要になると思われる。

③LOMの国際連携における課題

国内に限らず海外の教育コンテンツの統合検索も検討され始めてきている。しかし、同じ教育分野といっても標準となるLOMは一つではなく、IEEE/LOMとDoublin Coreが定めたメタデータ(注21)が使われている。また、海外の中には独自のメタデータを使っている機関もあることから、国際間の統合検索においてはメタデータスキームの相互互換が必要となる。表5にワシントン大学のStuart SuttonによるIEEE LOM Ver.1.0とDublin CoreによるLOMとの関係を示す(注22)。上記に示した必須項目と推奨項目だけを捉えるとDublin CoreのLOMにもほとんど適用できていることから、これらの項目で相互互換性を持たせることはそれほど難しくないと思われる。しかし、「9 Classification」の中の細かな分類まで互換性を持たせるとなると、コンテンツの分類法にまで考察の必要性が出てくるため容易ではない。個人的な見解であるが、初等中等教育においても高学年になるほどカテゴリ検索からキーワード検索が多く使われていること[4]から推測すると、高等教育における利用者もほとんどがキーワード検索を利用するのではないかと想定される。

(注18) 一例としてhttp://rss-jp.net/rss_list.html
(注19) Web上にいくつか出ている。http://www.webdevtips.com/webdevtips/codegen/rss.shtmlもその一つである
(注20) というより、まずは必要最小限のLOMデータを使って横断検索を構成していこうという動きである
(注21) 図書館、博物館、美術館などの組織によって検討が進められてきたメタデータ仕様であるhttp://dublincore.org/groups/education/
(注22) http://www.ischool.washington.edu/sasutton/IEEE1484.html

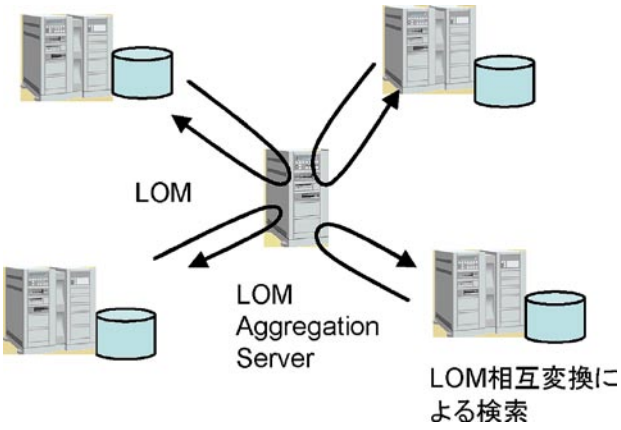


図4　LOM共有による情報統合

そうであれば、敢えてClassificationの中まで海外と互換性を持たせる必要はないのではないか。それよりはむしろキーワードとして入力が予想される専門用語の多言語対応の必要性を感じる。

次に、LOMを利用した検索サイトを分散させたいとき、あるいは海外の検索サイトを統合する場合に、メタデータ検索を利用することを考えてみる。一例として、図4に示すようにメタデータアグリゲーションによる方法が考えられる。それぞれの検索サイトが有するLOMをLOM Aggregationサーバに登録し、それを定期的に自サーバに収集・格納して検索する仕組みである。LOMがサーバに新規登録あるいは変更された場合に、サーバに接続されたそれぞれの検索サイトに自動送信する機能があれば、個々の検索サイトが常に最新のLOMを維持できる。一箇所の検索サイトに全てのLOMを集約して、その検索サイトを介してコンテンツを検索する場合と比較すると、この構成の特徴は既存の検索サイトが使えることから、例えば海外の検索サイトを統合した場合においても、自国に適したLOMの利用と検索サービスができることである。ただし、LOMに記述された内容そのものを著作権上公開できず、LOMを他サイトのサーバに格納することに問題がある場合には、次章のWebサービスによる方法が望ましい。

本章の冒頭でも述べたように、メタデータによる検索ではメタデータの内容が検索対象となり、実際のコンテンツはメタデータに書かれたURLにリンクされる仕組みがとられている。ここで留意すべきは、一部の検索サイトで実施されているように、ページのURLを動的に生成する場合にはメタデータによる検索は対応できないことである。すなわち、URLを直接ブラウザに入力してアクセス可能なコンテンツに限ることになる。この課題を解決するものは、次章で述べるWebサービスを用いた検索統合である。

4．Webサービスを用いた検索システムの統合

①SOAPを使った検索統合

Webサービスというと、言葉からではオンラインショッピングなどのようなWebを使ったサービスを連想するが、そうではなくネットワーク上で動いている異なるサービスを疎に結合することを指す。疎が意味することは、システム間のハードウェアやOS、実装プログラムに対する依存性が小さく、今まで個別であったシステムがインターネットを介して結合でき、さらにプログラム言語の異なるWindows系のシステムとUnix系のシステム間でも結合できることである。その実現のためには相互運用性を保証する技術が必要であり、結合する二つのシステムの送信側と受信側でデータを交換する仕組みが必要になる。その仕組みとして、送信先URLや転送方式を規定するSOAP（Simple Object Access Protocol）とデータの型、構造、表現方式などを規定するWSDL（Web Service Description Language）が開発されてきた。SOAPはシステム間で構造化されたデータをXMLを用いて情報を交換することを目的としたプロトコルであり、図5に示すような構成となっている。

検索システム間でSOAPメッセージがやり取りされる場合、まずWSDLでデータのやり取りの方法を確認する。受信システムが処理を行う情報はSOAP本体にXML形式で記述される。すなわち、送信システムから送信された検索窓に入力するキーワードを受信側で認識し、それを受信側の検索システムに入力して、検索された結果を、SOAP本体に記述されたレスポンスにしたがってXML形式で送信側に返す。入力するキーワードはXMLで記述され、この記述のし方は、例えば〈keyword〉…〈/keyword〉タグで挟まれた語彙を入力キーワードとするといったように、先のWSDLによるやり取りの中に規定されている。

図5　SOAPメッセージの構造

②検索統合の課題

本稿の「メタデータを用いた統合検索」で述べたように、メタデータがXMLで記述されていることから、Webサービスによる統合はメタデータを利用した検索システムの統合に適した方法であり、既に教育用コンテンツの統合検索でもサービスが行われている(注4)。既存の検索サイトを連携する場合に、1箇所をゲートウェイとしてみなし、そのゲートウェアからそれぞれの検索サイトにアクセスして、検索された結果を統合してゲートウェイ上のサイトに表示する方法と、個々の検索サイトがゲートウェイ的な役割を果たす方法が考えられる。国内で複数の検索サイトを統合する場合には前者の方法で十分であろうが、海外の主要なコンテンツ検索サイトを統合する場合には、アクセス負荷を分散するため、および各国が自国の検索サービスを活かせるようにするためには、後者の構成が良いと思われる。前述したMERLOTの統合検索(注4)は前者の構成となっているが、将来は後者の構成が取られていく可能性もある。

図6に後者の構成例を示す。図のA、B、Cは個々の検索サイトであり、それぞれが統合検索のゲートウェイとなっている。図6の検索サイトAを取り上げると、WSDLによりBおよびCのサイトが認識され、これらと通信が可能となり、同時にそれらのサイトとのデータのやり取りが確認される。BおよびCはWSDLの約束に従って、SOAPに書かれたデータの中から検索サイトに入力するキーワードを抽出し、自検索サイトで検索し、その検索結果をWSDLに書かれた返信条件にしたがってAに返す。例えば、検索結果の中に含まれるコンテンツのタイトルや概要などが「Title」や「Description」などのXMLで規定され、戻り値として返される。XMLが共通

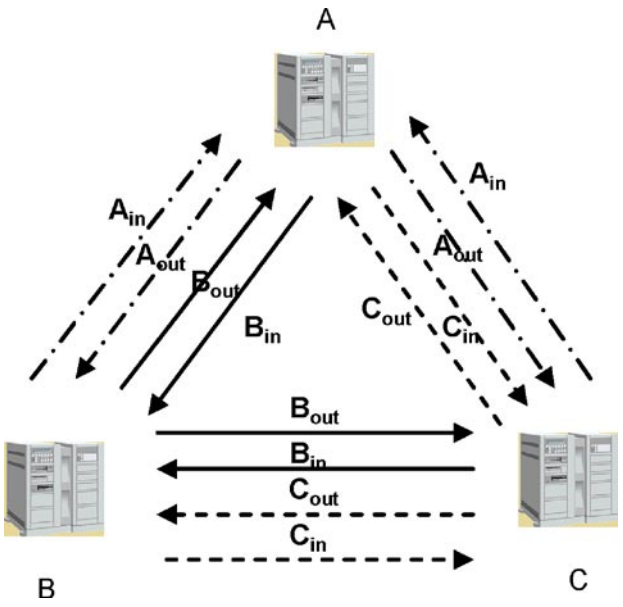


図6　Webサービスによる検索サイトの統合例

であれば、BおよびCから「Title」というXMLで規定された内容をタイトル名としてみなすことができる。AはBおよびCから収集した検索結果を、戻り値を元に自検索サイトの表示条件にしたがって表示する。A、B、Cで使われるSOAPの構造を規定することにより、共通利用が可能となる。A、B、Cの個々の検索サイトがそれぞれ検索結果に表示する項目を変えたいときには、個々のサイトへの戻り値の約束事を変えることにより対応は可能である。

SOAPを利用した検索統合では、個々の検索サイトが検索した結果を統合して表示するため検索時間がかかるという問題がある。また、自サイト内だけで検索する場合には検索結果の表示順は何らかの規則を元に優先順位を定めて決定すればよいが、表示条件の異なる検索サイト間の優先順位を決定するのは難しく、表示結果の順番をどのように設定するかも課題となる。SOAPは構造化されたデータをXMLを用いて交換することを目的としていることからメタデータを利用した検索サイトを統合するのに適していると述べたが、一方では、メタデータによるカテゴリ検索に必ずしも適していない。個々の検索サイトのキーワード入力欄に語彙を入力して検索するキーワード検索は可能であるが、カテゴリに沿って階層を追って検索していくことに対しては、階層を追うごとのデータのやり取りが発生するため、対応が困難である。またカテゴリ別に設定された入力欄(注23)が備わっている検索サイトを統合する場合には、どの入力欄にどの語彙を入力するかを定義し、それに対応するように検索システムの改良が必要となるため、技術的には可能であるが、個々の検索運用者間で稼動を厭わない連携が必要となり、実現は容易ではない。

5．まとめ

表6に上記三つの情報統合技術の特徴を示し、実用上の課題を以下にまとめる。

コンテンツに関しては、既にWeb上に散在するコンテンツを統合する場合にはWebマイニングが適する。Webマイニングによる Webページの自動収集や分類および検索は、既存のWebサイトの運用者にとっては検索システムからのクローリングが定期的に自サイトに訪れるだけで、特別な稼動負担は発生しない。収集側からみると、設定された収集条件や抽出条件の基で収集が自動的に行われるため、収集対象となるWebページの内容更新にも対応可能である。しかし、如何に収集条件や抽出条件の精度を上げるかが大きな課題であることはWeb

(注23) 例えば検索語彙入力部分に「タイトル入力欄」、「対象学年入力欄」などが備えられており、個々の項目で検索可能となっているような場合

表6　それぞれの情報統合技術の特徴

	統合化の対象	データ収集法	統合条件	特長	課題
既存Webサイトからの必要な情報の抽出と統合検索（Webマイニング）	Webページ内情報	クローリング	クローラアクセス可	個々の情報へのメタデータの付与負担がない（頻繁に内容の変わる情報の統合に適合） 既存のWeb情報の改変に対応可能 情報に対する全文検索が可能	的確な情報の自動収集（情報の抽出条件の高精度化） Webからのリンク情報から辿れないURLを有するページの収集が不可
情報へ付与するメタデータによる情報の統合検索（セマンティックWeb）	情報に付与したメタデータ	メタデータ登録	メタデータスキームの統一	分散したデータベース的な分野の情報統合に適合 メタデータへ著作権利情報などの付加情報の付与が可能 画像や映像などテキストを含まない検索に対応	メタデータの付与負担（メタデータの自動付与） 異なるメタデータの互換のためのオントロジー開発が必要
Webサービスによる検索システムの統合（Webサービス）	連結した検索システムの検索結果	SOAPやWSDLなどの規定に従った相互運用	システム間のデータ交換のためのデータ転送方式（SOAP） システム間のデータ交換のためのデータ様式（WSDL） XML表記法	サイト上に存在する既存の検索サイトを連結したメタ検索が可能	検索所要時間 検索結果表示順の決定 LOM検索の特徴であるカテゴリ検索には不適

マイニングの宿命である。

一方、これからコンテンツを開発しようとする場合においてはメタデータによる情報統合が有効である。条件として開発に合わせてメタデータ登録の協力が得られることが前提ではある。この場合、メタデータのタイトルや概要、URLだけをコンテンツからそのままテキスト抽出して登録したのでは意味が無い。クローラによるテキスト収集と大差が無いからである。LOMで記述された必須条件、推奨条件までのメタデータ付与が重要となる。

個々の運用者により検索システムが運用されている場合はどうであろうか。クローラにより個々のコンテンツを収集できない可能性が高いためWebマイニングは不適である。検索システムの個々のコンテンツのURLが静的に固定しているのであれば、メタデータの共有により情報統合は可能である。この場合、コンテンツへのメタデータ登録の稼動が課題となることから、それぞれの検索システムがメタデータ検索に対応していることが必須と考えられる。

コンテンツに直接アクセスできず検索によってのみ表示が可能な検索システムや、メタデータ検索によらない検索システムを統合する場合にはWebサービスが適する。ただし、WSDLやSOAPを共有することが必須であることから、運用者に検索インタフェースの部分の開発協力が必要となる。コンソーシアムのように協力体制が確立されて進める場合には適している。必ずしもコンテンツはメタデータが付与されている必要はないが、検索キーワード入力や複数の検索結果を一つの表示形式にあわせる場合に、XML形式で表示内容の整合を図る必要がある。このようにキーワード検索の統合は可能であるが、個々の検索システムにカテゴリ検索機能が含まれる場合には、その機能まで含めた統合化のハードルは高くなる。

参考文献

[1] Brin S, Page L., The anatomy of a large-scale hypertextual Web serch engine, Proceeding of 7th WWW Conference (1998)

[2] 荻野達也、神原顕文、清水　昇、豊内順一、細見　格、津田　宏、白石展久、韋　慶傑、セマンティックWebとは、情報処理、43巻、7号、pp.709-717、(2002)

[3] 伊藤栄典、山田信太郎、廣川佐千男、Webシラバス統合のためのレコード解析、人工知能学会研究会資料、SIG-SWO-A201-05、pp.1-7（2002）

[4] 情報化人材育成プラットフォーム（情報技術活用学習基盤システム開発）」報告書　第6編「コンテンツ検索プ

ラットフォーム開発・実証実験及び関連調査」、ALIC、（2004）
［5］ 北　研二、津田和彦、獅々堀正幹（2003）、情報検索アルゴリズム　共立出版　東京
［6］ Lawrence S., Giles L. and Bollacker K., Digital libraries and autonomous citation indexing, IEEE Computer, Vol. 32, No. 6, pp.67-71 (1999)
［7］ L. Page, S. Brin R. Motwani and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Proc. of the 7th WWW Conf., pp.161-172 (1998)
［8］ M. Shinohara, M. Okamoto, Y. Okui and T. Tanaka, Effective Retrieval for Educational Resources using Learning Object Metadata for K-12 Schools in Japan, The 8th International Dublin Core Metadata Workshop, pp.255-258, (2001)
［9］ 清水康敬、高等教育におけるe-Learningの支援と教育コンテンツの共有、メディア教育研究、Vol. 1、No. 1、pp.1-10（2004）
［10］ 篠原正典、徳畑香菜、岡本麻由美、三宅丈夫、永野和男、小・中学校教育用学習素材検索システムの開発と児童・生徒の検索時における検索過程、教育システム情報学会、Vol. 18、No. 2、pp.200-209、（2001）
［11］ 住　太陽、検索エンジン業界勢力地図、情報の科学と技術、54巻、2号、pp.72-77（2004）
［12］ 山田信太郎、松永吉広、伊藤栄典、廣川佐千男、Webシラバス情報収集エージェントの試作、電子情報通信学会論文誌、D-I、Vol. J86-D-I、No. 8、pp.566-574（2003）
［13］ 山名早戸、検索エンジンのアーキテクチャ、情報の科学と技術、54巻、2号、pp.84-89（2004）

篠原（しのはら）　正典（まさのり）
1954年鹿児島県生まれ。1977年鹿児島大・工・電子卒。同年電電公社入社。武蔵野電気通信研究所、厚木研究開発センターで化合物半導体結晶材料、量子構造デバイスの等の研究に従事。1990年工学博士（東京大学）。1996年よりNTTで教育の情報化プロジェクト「こねっとプラン」を推進。2004年2月より現職。日本教育工学会、電子情報通信学会、日本通信教育学会、応用物理学会、情報コミュニケーション学会各会員。

地蔵（ちくら）　真作（しんさく）
1965年生。1989年愛知教育大学卒業。同年株式会社ウイズキッズに設立メンバーとして参加。教育関連のパッケージソフトウェア開発に従事。1995年独立。1997年、有限会社リアクト設立、取締役に就任。現職。インターネット上での検索システムのソフトウェア開発、システムインテグレーション、アジャイルソフトウェア開発手法の調査実践を行う。ACM、情報処理学会会員。

Issues of Federated Search of Higher Educational Resources on the Web and Some Proposals towards Effective Search

Masanori Shinohara[1)] ・ Shinsaku Chikura[2)]

Necessity of sharing and delivery of educationally useful contents on the Web has been urged in the IT-advanced nations. Investigation of federated search and sharing of contents through the Internet are going to be launched and discussion has just started in some council to make a standard. Seizing this opportunity, this material describes the current situation and deals with several issues of integration of contents including linking several search systems from the point of Web mining, semantic web and Web services. Moreover, some methods are proposed towards more effective search based on the actually examined data.

Keywords
Web mining, semantic Web, Web service, metadata, federated search, educational contents

[1)] National Institute of Multimedia Education
[2)] React Inc.