論　文

深層ウェブの実態とその要因：機関リポジトリに登録された文献を用いた調査*

宮田洋輔*2
安形輝*3
池内淳*4
石田栄美*5
上田修一*6

ウェブの規模が増大するにつれ，検索エンジンからアクセスできない状態，すなわち深層ウェブも増大していることへの関心が高まっている。マッカウンら（2006）とハーゲドーンとサンテッリ（2008）は，深層ウェブの規模をOAI-PMHを用いて収集した機関リポジトリに収録された文献のメタデータを用いて計測した。本研究では，2009年9月に，先行研究の手法を応用し，日本の機関リポジトリから収集した全文PDFファイルのURLを用いて，より大規模に深層ウェブの比率を計測した。その結果，Google，Yahoo!，Bingの3つの検索エンジンから検索できるウェブは72.0%に過ぎず，28.0%が深層ウェブとなっていることが分かった。1つの検索エンジンでは，最高でもGoogleの53.2%であった。また，PDFファイルとURLの特徴の調査から，動的なURLや長いURLが深層ウェブとなる要因であることが分かった。

目　次

* 2011年11月2日受付　2012年3月23日受理
*2 みやた　ようすけ　慶應義塾大学
*3 あがた　てる　亜細亜大学
*4 いけうち　あつし　筑波大学
*5 いした　えみ　九州大学
*6 うえだ　しゅういち　慶應義塾大学

1. はじめに

1.1 深層ウェブとは

「深層ウェブ」(deep Web) とは，端的に言えば，検索エンジンからはアクセスできないウェブコンテンツを指す。検索可能な表層ウェブ(surface Web) の対義語である。「深層ウェブ」という語の提唱者であるバーグマン (Bergman, Michael K.) によれば，他のウェブコンテンツからリンクでは到達不可能なコンテンツのことであり，データベースのようにパラメータ付きクエリに対し動的に生成されるページ，スクリプトや Flash によって動的に生成されるリンクからしか到達できないページなどがこれに該当するとされていた[1]。さらに，バーグマンは，深層ウェブになっているコンテンツの特徴として，表層ウェブのコンテンツに比べて，より深遠なコンテンツを持ち綿密であり，より高価値であるとしている[2]。BrightPlanet は，このような深層ウェブは検索エンジンを通してアクセスできるウェブの 500 倍の規模があると報告した[3]。

検索エンジン提供者も，深層ウェブのコンテンツを看過してきたわけではない。たとえば，Google は，HTML フォームに対してテンプレートを用いた処理を行うことなどによって，動的に生成されるページも収集している[4]。一方で，ウェブの規模が増大するにつれ，公開されていても検索エンジンのロボット[5] による収集がされず，検索できない静的なページも増加していると思われる。

ウェブコンテンツに関するアクセス可能性に基づく分類は，様々に議論されており，定義をしておく必要がある。図 1 に公開と検索可能性に基づくウェブ全体の構造を図示した。

まずウェブ全体は「公開されているか」から公開ウェブ (public Web)[6] と非公開ウェブ (private Web)[7] に分けられる。公開ウェブは，コンテンツの URL (URI) さえ分かれば，アクセスできるようになっているものである。一方，非公開ウェブは，パスワードや IP アドレスなどによってアクセス制御されているものである。非公開ウェブは，アクセスに認証を必要とするものであり，第三者による不正なアクセスは，2000 年施行の「不正アクセス行為の禁止等に関する法律」によって違法行為と見なされている[8]。

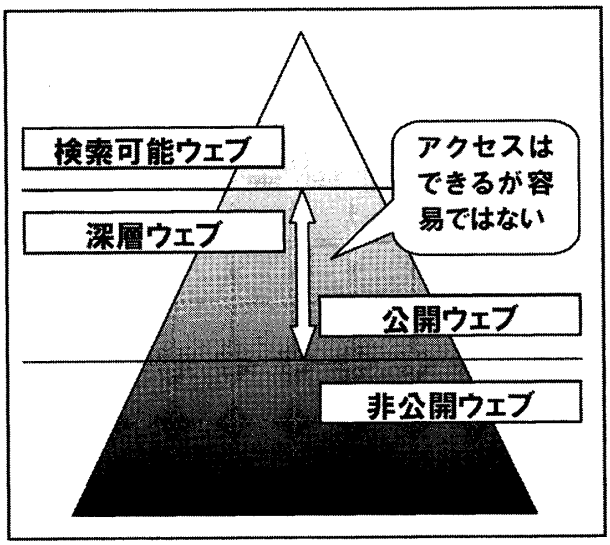

図 1 公開と検索可能性に基づくウェブ全体の構造

非公開ウェブにあるコンテンツは，認証によって限定された範囲以外からのアクセスを期待するものではない。そのため，非公開ウェブに存在するコンテンツに関しては，検索エンジンなどの手段ではアクセスすることができない。本研究は公開されたウェブに関して，検索エンジンからどの程度アクセスが可能であるか，を対象とし，非公開ウェブについては扱わない。

公開ウェブは，さらに検索エンジンなどの手段で「検索できるか」という点から，検索可能なウェブと検索できないウェブに分割される。ここで，検索可能なウェブ (indexable Web) は一般的な検索エンジンから検索してアクセスできるコンテンツ群のことを指す。検索可能なウェブは，見えるウェブ (visible Web) と呼ばれることもあり，バーグマンの議論における「表層ウェブ」と同義であると考えられる。ローレンスとジャイルズ (Lawrence, Steve and Giles, C. Lee) の調査では 1999 年の 2 月の時点で，公開されて検索可能なウェブのサイズは 800 万ページであるとされていた[9]。2011 年 11 月現在，約 128 億ページが検索可能なウェブになっていると推定されている[10]。Google は，2008 年 7 月の時点で 1 兆の URL を把握しているとしている[11]。

一方の公開されているものの検索できないウェ

ブが「深層ウェブ」である。コンテンツの公開が，アクセスを期待したことであると考えると，公開したはずのコンテンツが検索エンジンという一般的な経路からアクセス出来ないことは望ましい状態ではないといえる。先述したとおり，深層ウェブとなる原因には，他のコンテンツからリンクでは到達不可能なコンテンツ，データベースのようにパラメータ付きクエリに対し動的に生成されるページ，スクリプトやFlashによって動的に生成されるリンクからしか到達できないコンテンツなどが考えられる。

深層ウェブと非公開ウェブは，検索エンジンによるアクセスできないという点は共通しているものの，深層ウェブは非公開ウェブの場合と異なり，アクセス自体が制御されているわけではなく，URLが把握できていれば一般的なウェブブラウザなどを通してアクセス可能である。深層ウェブは，見えざるウェブ（invisible Web）[12]や隠されたウェブ（hidden Web）[13]，などと呼ばれる場合もある[14]。そのため，本研究では「深層ウェブ」を，検索エンジンで検索することができないウェブコンテンツであると定義する。

1.2 学術情報流通と深層ウェブ

日常の探索行動の中でGoogleなどの検索エンジンに頼る度合いは，ますます強まりつつある。仕事や生活上で生じる問題に必要な情報の入手は，検索エンジンによる検索から始まることが多い。また研究活動においても，検索エンジンは，重要性を増している。

コナウェイ（Connaway, Lynn S.）らは，2005年から2009年までに英国研究情報ネットワーク（RIN）と英国の情報システム合同委員会（JISC），さらにOCLCが行った12種類の情報源探索行動の利用者調査の体系的レビューを行っている[15]。この中では，いずれの調査結果にも共通する発見として，研究者も学生も研究行動については分野による違いが存在すること，研究遂行のどの段階でも電子ジャーナルの重要性が増していること，Googleやその他の検索エンジンが重要であることを示す多くのエビデンスがあること，Googleは，電子ジャーナルのコンテンツを見つけ，利用するために使用されていることなどがあげられている。

上記の体系的レビューにも含まれている英国研究情報ネットワークの2006年の報告では，文献を見つけるのにとてもよく用いるツールとして，研究者や図書館員の6割は検索エンジンを挙げた。一方，書誌データベースをよく使うのは，2割だった[16]。

ニウ（Niu, Xi）らによる米国の5大学の研究者2,063名を対象とした研究活動の調査で，文献探索を始める時にまず使うのはGoogleか図書館のページかを尋ねたところ，ほぼ半々の結果だった。ただし，3つの大学では図書館のウェブサイトよりもGoogleの方がよく使われていた[17]。

検索エンジンの利用が特に好まれる理由について，ヘミンガー（Hemminger, Bradley M.）らは二つの理由をあげている。Googleなどが好まれるのは，第一に，検索エンジンには，メタサーチ，すなわち多数の情報源を一括して検索する機能[18]があるためであり，第二に，検索結果から必要とする情報そのものをたやすく入手できることがあるためである[19]。このように検索エンジンは，研究者によく使われ，学術情報流通の中で次第に重要な位置を占めつつある。

セルフアーカイブや機関リポジトリへの登録によって，ウェブ上に研究成果を公開する利点に関する研究も行われている。例えば，デイヴィス（Davis, Philip M.）は，オープン・アクセスの論文とそうでない論文を比較し，オープン・アクセスの論文がより多くの引用を集めるわけではないが，より多くダウンロードされたことを明らかにし[20]，公開することの利点を示した。

公開された学術情報へのアクセスの経路の多くも，検索エンジン経由であることも明らかになってきている。オルガン（Organ, Michael）によるウロンゴン大学のリポジトリの調査では，リポジトリの全文ファイルのダウンロードの95.8%がGoogle経由であった[21]。また，佐藤と逸村による日本の4つの機関リポジトリのアクセスログの分析からも，3つのリポジトリで検索エンジンからのアクセスが最も多く，またその中でもGoogleからのアクセスが多かった[22]。

このように，学術情報へのアクセスと検索エンジンとの関係はますます強くなってきている。そ

のため，検索エンジンで検索できない深層ウェブがどの程度存在するのかを知ることは重要である。深層ウェブの実態を解明することとともに，深層ウェブが検索エンジンにどのような影響を与えているのかを検討することは新たな課題となっている。また，学術情報における検索エンジン利用の実態を考えると，深層ウェブがどの程度存在するか，深層ウェブが検索エンジンにどのような影響を与えているかを検討することは，学術情報流通の研究にとっても，重要な意義があると考えられる。

1.3 機関リポジトリを用いた深層ウェブの調査

ウェブ全体に占める深層ウェブの規模を把握するためには，ロボットを用いてすべてのウェブを収集する必要がある。しかし，リンク切れやインターネットの規模とその成長の速さによってすべてのコンテンツを収集調査することは事実上不可能である。

そこで，機関リポジトリを深層ウェブの調査に用いる研究が現在までにいくつか行われてきた。機関リポジトリに収録された文献はOAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）[23]によってすべて収集可能である。OAI-PMHを使って網羅的に収集したコンテンツ群をウェブ全体のサンプルとして用いることで，ロボットでウェブ全体を収集するのとは異なる方法で，深層ウェブの調査が可能になる。

マッカウン（McCown, Frank）らと，ハーゲドーンとサンテッリ（Hagedorn, Kat and Santelli, Joshua）によって，機関リポジトリに収録された文献のメタデータを用いた深層ウェブの調査が実施されている。

マッカウンらは，2005年6月にOAI-PMHで取得した文献データについて，検索エンジンのカバー率の調査を実施した[24]。検索エンジンによるカバー率とは，調査対象URL全体に対する検索可能であったURLの比率である。

マッカウンらの調査の手順は以下のとおりである。

1) OAI-PMHに準拠したリポジトリのリストの作成
2) リポジトリからのOAI-PMHによるメタデータレコードの網羅的収集
3) メタデータレコードから調査対象URLの抽出
4) 抽出したURLを検索エンジンで検索し，検索結果を記録する

マッカウンらは機関リポジトリに関する4つのレジストリから776のリポジトリのURLを取得した。つぎに776のリポジトリのうち475のリポジトリに対してOAI-PMHによってメタデータの収集を実施し，9,843,451件のメタデータレコードを取得し調査対象とするデータ集合を作成した。調査対象から抽出した4,376,271件の重複のないURL（ダブリンコアのidentifiers）から無作為抽出した1,000件のURLについてYahoo.com, MSN, Googleの3つの検索エンジンにおけるカバー率を調査した。

その結果，Yahoo.comのカバー率が最も高く65%であり，Googleのカバー率は44%，MSNは7%しかカバーしていなかったことを報告している。そして，URLの21%は，いずれの検索エンジンでもカバーされておらず，深層ウェブとなっていることが明らかになった。

つぎに，ハーゲドーンとサンテッリは2008年6月にマッカウンらの調査の追試として，OAISterから抽出したレコードに対するGoogleによるカバー率に関する調査を実施した[25]。

これは先行研究の追試ではあったが，調査の手順には異なる点があった。マッカウンらの調査では，調査対象となったリポジトリを5グループに分類していたのに対し，ハーゲドーンとサンテッリの調査では，収録しているレコード数に基づいて4グループに分類している。また，マッカウンらの調査では，グループごとに1,000件のレコードのみを対象としていたのに対して，ハーゲドーンとサンテッリの調査では各グループ収録レコードの10%を無作為抽出し，合計147,305件のURLを調査対象としている。さらに調査に用いた検索エンジンは，Googleのみである。

ハーゲドーンとサンテッリは調査の結果，Googleのカバー率は44.35%であることを明らかにした。この結果から，3年前に実施されたマッカウンらの調査のGoogleのカバー率44%と殆ど違いはなく，Googleは深層ウェブを検索できる

ようにしていないとしている。

これらの調査では，深層ウェブの規模は明らかにされたものの，それらが何故深層ウェブになっているかについての要因は調査されていない。

1.4 研究の目的

様々な領域でウェブの重要性が増すに連れ，ウェブ上に存在する情報資源のどの程度が検索可能かどうか，を理解することの重要性は高まっている。本研究の目的は，ウェブ全体に占める深層ウェブの規模を測定し，コンテンツが深層ウェブとなっている要因を明らかにすることである。

そこで，日本の機関リポジトリから網羅的に収集した全文ファイルのURLをウェブ全体のサンプルとして用いて，深層ウェブの規模を調査した。さらに深層ウェブの調査結果に基づいて，コンテンツの特性（ファイルの特徴，URLの特徴）との関係を分析し，深層ウェブとなっている要因についての調査を行った。

2. 深層ウェブの調査

2.1 調査方法

2.1.1 調査対象機関リポジトリ

今回の調査対象とするウェブコンテンツは，国立情報学研究所（NII）が学術機関リポジトリポータルJAIROでのメタデータ収集のために用いているjunii2メタデータ・フォーマット[26]に対応した日本の92機関リポジトリで公開されている全ての全文ファイルのURLとした。

機関リポジトリの全文ファイルを対象とした理由は，(1) OAI-PMHによる収集により，リンクを辿るロボットによるクローリングとは別の手法で，全てのコンテンツを収集できること，(2) 研究成果の公表を主たる目的とするためロボット排除プロトコルの設定を厳しくしていないことが推測できること，(3) 学術情報の全文ファイルは機関リポジトリの中心的なコンテンツであること，である。

日本の機関リポジトリに限定した理由は，junii2メタデータ・フォーマットに対応するデータ収集を行うためである。このフォーマットは，OAI-PMHでサポートが義務付けられているoai_dcのような他のメタデータ・フォーマットと比較して，情報量が多く，収集後の詳細な分析が可能となる。

機関リポジトリからの網羅的なメタデータのハーベスティングは2009年4月11日に実施した。メタデータから全文ファイルのURLを抽出した結果，合計で404,431件が得られた。

2.1.2 検索エンジンを対象とした調査

調査対象とした検索エンジンは，Google，Yahoo! Japan（以下，Yahoo!とする），Bingとした。選定の際には，(1) 検索エンジンシェアの上位であること[27]，(2)国内だけでなく世界的なサービスを展開していること，(3) 検索エンジンをプログラムから利用するためのAPIを公開していること，(4) 任意のURLからの検索を実行可能であること，を条件とした。

調査に用いた検索APIは，GoogleはGoogle AJAX Search API，Yahoo!はYahoo!デベロッパーネットワークのウェブ検索Web API，BingはBing API 2.0である。各検索エンジンAPIには検索式として調査対象URLを渡し，検索結果が0件であれば検索できないURLとし，1件以上であれば検索可能URLとした。なお，Yahoo!は全角文字の一部が含まれるURL（10件）について，検索API経由の検索，手動での検索の両方において，「PDF」という検索式に自動的に変換されてしまい，適切な検索ができなかった。したがって，それらのURLは検索されなかったものと見なした。

2009年9月6日～8日にかけて，404,431件の全文ファイルのURLを検索式として，検索エンジンに対する調査を実施した。

2.1.3 カバー率と重複率

検索エンジンで検索可能なウェブの比率を示すカバー率と，検索エンジン同士の重複率を算出することで，深層ウェブがどの程度あるかを明らかにすることができる。

カバー率は以下のような式で算出した。

$$\text{カバー率} = \frac{\text{検索可能URL数}}{\text{調査対象URL数}}$$

また，複数の検索エンジンで検索可能な重複部分の比率を算出した。

2.2 調査結果

2.2.1 検索エンジンのカバー率

表1は，検索エンジンごとに，調査対象URLのうちの検索可能URL数，さらにその値からカバー率を算出したものである。

検索エンジンのなかで単独でのカバー率が最も高いのはGoogleであるが，Googleだけでは調査対象URLの53.2%の範囲しかカバーできていないことがわかった。Google，Yahoo!，Bingを組み合わせた場合には72.0%の範囲までカバーできていた。逆に，全体の28.0%は主要な3つの検索エンジンを組み合わせたとしても検索できず，深層ウェブとなっている。

2.2.2 検索エンジンの重複率

表2は検索エンジンの重複率を算出したものである。

表1 検索エンジンのカバー率

	Google	Yahoo!	Bing	合計
検索可能URL	215,359	174,805	115,679	291,024
カバー率	53.2%	43.2%	28.6%	72.0%

調査対象数 404,431

表2 検索エンジンの重複率

	Google	Yahoo!	Bing
Google		54.2%	37.9%
Yahoo!	66.7%		39.1%
Bing	70.5%	59.2%	

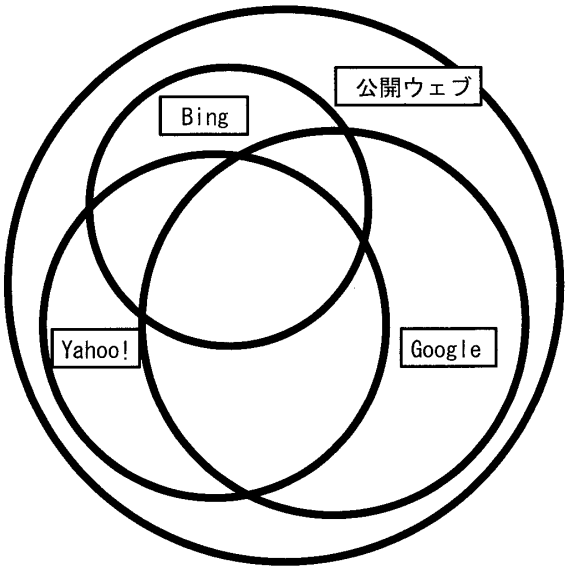

図2 検索エンジンのカバー率と重複率

この表は，表頭の検索エンジンのカバー範囲が，表側の検索エンジンのカバー範囲と重複する比率を示している。例えば，Googleの54.2%はYahoo!と重複し，Yahoo!の66.7%はGoogleと重複している。

また，表1，2の値を用いて，公開ウェブに対する検索エンジンのカバー率の大きさを円の大きさで，エンジン同士の重複率の大きさを円の重なり具合で示したものが図2である。

この図からは，カバー率の最も高い検索エンジンだけではウェブ全体を検索できるわけではないこと，各検索エンジンがカバーする範囲は重複しつつも異なる範囲であることがわかる。

3. 深層ウェブとなる要因の調査

3.1 ロボット排除プロトコルの調査

検索エンジンによるコンテンツへのアクセス可能性は，ロボットがウェブのどの範囲を収集し索引化するのかという挙動に大きく依存している。ロボット排除プロトコル（Robots Exclusion Protocol）は，ウェブサイトの管理者が，自らのサイトへのロボットのアクセスをコントロールするために広く用いられているルールであり，多くの倫理的な振る舞いをするロボットは，この強制的でない慣習に従っている。具体的には，サーバのルートディレクトリに，robots.txtというテキストファイルを設置し，アクセスを許可しないロボットの名称やサーバ内のディレクトリ名を記述する[28]。ロボット作成者は，ロボットがサーバにアクセスする際に，まずこのファイルを読み込み，そのアクセスの範囲を確認するよう実装することが求められる。

これまで，robots.txtがどの程度利用されているかという実態調査がいくつか行われてきた。たとえば，ケリーとピーコック（Kelly, Brian and Peacoc, Ian）は，1999年に，英国の163大学のウェブサイトを調査し，32.5%にあたる53サイトがrobots.txtを設置していることを明らかにするとともに，各々のファイルサイズやこのルールが実際にどのような使われ方をしているかについて言及している[29]。また，ドロット（Drott, Carl M.）は，2000年と2001年に，Fortune Global 500の上位

30社と下位30社の計60社について，robots.txtやMETAタグの内容を調査している[30]。2000年の結果では，robots.txtを設置していたのは16サイト（26.7%）であったのに対し，2001年調査では11サイト（18.3%）と減少している。また，ヨーロッパやアジアの企業と比較して，北米の企業の方が多く設置しているものの，全体としてrobots.txtは広く普及しているとは言えないと結論づけている。このほか，サン（Sun, Yang）らは，2005年から2006年にかけて，計5回の調査を行い，Open Directory Project（DMOZ）の「教育」「ニュース」「大学」という三つのカテゴリから抽出した計7,593サイトを対象としてrobots.txtを自動的に収集・分析している[31]。robots.txtを設置しているサイトは，調査期間を通じて，35.1%～38.5%へと増加しており，なかでも，ニュースサイトと米国の大学の設置率が高く，企業やアジアの大学の設置率は低い。このほか，robots.txtで頻繁に記載されるロボット名，ファイルサイズ，アクセス間隔の指定，ロボット排除プロトコルの誤用や矛盾といった点について調査を行っている。

ここでは，NIIによる機関リポジトリ一覧[32]に掲載された115のサイトを対象として，robots.txtの調査を行った。サーバが一時的にアクセスできない状態となっている可能性を考慮して，2009年9月12日，9月14日，および，9月29日の三回にわたって，同様の調査を実施した。その結果を表3に示した。その結果，robots.txtを設置しているのは，45サイト（39.1%）であった。これは既往調査の結果とほぼ同様の結果であったと言える。

ロボット名を指定するUser-agentフィールドに着目すると，42サイトはすべてのロボットを意味する「*」を用いていた（表4）。特定のロボット名を記述しているのは3サイトのみであり，Slurp（Yahoo! Search Technologyのロボット），MSIE（Microsoft Internet Explorer），Googlebot（Googleのロボット），msnbot（MSN Searchのロボット），baiduspider（中国の検索エンジン百度のロボット），Yeti（韓国の検索エンジンNAVERのロボット），Ocelli（フリーのロボット）などがあった。また，robotx.txtファイルは存在するが，

表3　robots.txtの有無

内容	件数	比率
robots.txtあり	45	39.1%
robots.txtなし	66	57.4%
アクセス不可	4	3.5%
合計	115	100.0%

表4　robots.txtの記述内容

内容	件数	%
User-agent:*のみ	39	88.6%
特定のUser-agentを指定	3	6.8%
何も記述していない	2	4.5%
合計	44	100%

何も記載していないものも2サイトあった。このほか，Sitemapフィールドを指定しているものが2サイト，ロボットのアクセスの間隔を指定するCrawl-delayフィールドを指定しているものが1サイト存在した。User-agentフィールドに「*」，Disallowフィールドに「/」を記述して，全てのロボットのクローリングを完全に排除しているものが4サイトあった。

3.2　全文URLの調査

学術情報の全文URLの定量的特性が検索エンジンへの登録を阻害している可能性が考えられる。URLと検索エンジンによる索引化との関係は，検索エンジン最適化（Search Engine Optimization, SEO）の分野ではしばしば「クローラビリティ」として言及される。

「クローラビリティ」は，検索エンジンが，ウェブ上のコンテンツを収集する際に動かしているロボットによる収集のしやすさを意味している。SEOやウェブマーケティングの世界では，ウェブページ・コンテンツのクローラビリティを高めるための方策がしばしば議論されている。検索エンジンを運営するGoogleやMicrosoftなどもSEOに関する情報を提供している。

そこで，junii2形式で収集したメタデータのfullTextURL要素から抽出した，404,431件の全文URLを用いて，検索エンジンへの登録と全文URLの特徴との関係の分析をおこなった。

全文URLを分析する観点として，「URLの長さ」，「URLが動的であるか」，「ディレクトリの

深さ」の3つの観点から，検索エンジンへの登録との関係を分析した。

「URLの長さ」はURLの文字数によって計測した。HTTP1.1の仕様に関するRFC2616「Hypertext Transfer Protocol – HTTP/1.1」では，無制限の長さのURIを扱えるようにするべきであるとされているが，255バイト以上の長さのURIをサポートしていないクライアントも存在するため注意が必要とされている[33]。また，各ブラウザで，様々ではあるが，マイクロソフトが提供するウェブブラウザInternet Explorerで処理できるURLの最大文字数は2,083文字[34]であるように，URLの長さに対する上限が設定されている。このようにウェブコンテンツの扱いやすさとURLの長さとの間に相関があることが推測できる。

動的なURLとは，データベースなどから情報を得る際に，検索インターフェイスに対して利用者が入力した内容によって最終的なURLが決定するURLである。動的なURLである場合，検索エンジンによって索引化されない可能性の高いことが以前は指摘されていたが[35]，近年は検索エンジンのアルゴリズム性能の向上によって，2ないし3程度のパラメータを含んだ動的なURLは静的URLと同じように扱われているとされている[36]。ここでは，URL中にパラメータを付加するための「?」が含まれている場合，動的URLであると判断した。

URLの「ディレクトリの深さ」もしばしば検索エンジンからのクローラビリティに影響を与える要素である。例えば，Googleの「検索エンジン最適化スタートガイド」においても，検索エンジンに登録されやすくするために，“ディレクトリ構造を簡潔にしよう”や“自然な階層構造を作ろう”など，ディレクトリ構造に言及している[37]。ここでは，「ディレクトリの深さ」をURL中に含まれる，ドメイン名以降の「/」の数によって測定した。

図3にURLの分析の例を示した。図のURLは，文字数が98文字なのでURLの長さは98，URL中に?が含まれており，動的なURLである。また，ホスト名の「koara.lib.keio.ac.jp」以下に，4つの/があることから，ディレクトリの深さは4となる。

http://koara.lib.keio.ac.jp/xoonips/modules/xoonips/download.php?koara_id=AN00003152-00000056-0043

URLの長さ: 下線部の文字数
動的URL: 囲み線部
ディレクトリ階層: 2重線部

図3 全文URLの分析例

全文URLの特徴と検索エンジンからのアクセス可能性との関係を表5に示した。

URLの長さは偏りが大きいことから，URLの長さを四分位数ごとに区切り，全体を四つのブロックに分けて，アクセス可能性との関係を分析した。まず，最もURL長の短い下位の25%（67文字以下のブロック）では84.8%がいずれかの検索エンジンからアクセスできるようになっている。次に，25%から50%までの68～73文字，50%から75%までの74～80文字でも，70.9%と79.4%でかなり高いカバー率を占めている。しかし，最も長い上位25%（81文字以上のブロック）では，カバー率は56.0%と他の数値に比べて顕著にカバー率が低下していることが明らかになった。

次にURLが静的か動的か，と検索エンジンからのアクセス可能性との関係を見る。表から静的なURLは77.0%がいずれかの検索エンジンからカバーされているのに対して，なんらかのパラメータを含んだ動的なURLの場合は26.3%しかカバーされておらず，動的なURLは検索エンジンからのアクセス可能性が明らかに低いことが分かった。近年の検索エンジンは，不得手とされてきた動的なURLの索引化に対しての取り組みを示してきていた[38]。しかし，調査の結果から，検索エンジンは未だに静的なURLと同じくらい網羅的には動的なURLを収集できていないことが分かった。

最後にディレクトリの深さとアクセス可能性については，それぞれの階層ごとに集計し比較した。ディレクトリの深さでのカバー率を比較すると，階層数が10を超えたURLに関しては1件も検索エンジンからアクセスできなかったものの，文献からの予想とは異なり，ディレクトリ階層が深く

表5 URLの特徴と検索可能性

		検索不可		検索可	
		件数	%	件数	%
URLの長さ	67文字以下	13,803	15.2%	77,209	84.8%
	68～73文字	32,012	29.1%	78,037	70.9%
	74～80文字	19,245	20.6%	74,179	79.4%
	81文字以上	48,347	44.0%	61,599	56.0%
URLのタイプ	静的	83,955	23.0%	280,513	77.0%
	動的	29,452	73.7%	10,511	26.3%
ディレクトリの深さ	1	4	4.3%	88	95.7%
	2	834	2.7%	29,928	97.3%
	3	26,722	57.1%	20,116	42.9%
	4	14,171	23.0%	47,502	77.0%
	5	1,066	3.0%	35,000	97.0%
	6	69,669	30.8%	156,456	69.2%
	7	935	32.8%	1,913	67.2%
	10	4	100.0%	0	0.0%
	16	1	100.0%	0	0.0%
	不正値*	1	4.5%	21	95.5%
合計		113,407	28.0%	291,024	72.0%

*「不正値」はURLとして不正な値が入力されていたもの

なるとカバー率が下がるという傾向はなかった。階層数が10以下のものに関しては，ディレクトリ階層数が3の場合，検索不可が57.2%で，より階層数の多い6（30.8%），7（32.8%）の場合に比べても，検索できない割合が高かった。

3.3 PDFファイルの調査

検索エンジンがPDFファイルを索引化する際に，PDFファイルの特性によって，索引化が阻害されている可能性が考えられる。そこで，PDFファイルの特性とアクセス可能性の調査を行った。

PDFファイルの調査では，各リポジトリに収録された学術情報資源の全文URLから2万件を無作為抽出した。2万件のURLから，ファイルがダウンロードできなかったものとPDF以外のファイルを除去し，調査対象のPDFファイルは，14,196件であった。

検索エンジンの索引化に影響を与えうるPDFファイルの特性として，「テキスト抽出の可否」と，「暗号化の有無」について調査した。

「テキスト抽出」はPDFファイルからの，記録されたテキストの情報を取得できるかどうかを調査した。PDFファイルは，作成に際して，文書のファイルから作成する場合と，1枚以上の画像を結合することでファイルを作成することがある。前者の場合は，文書のテキストを利用することができるが，後者の場合はテキストを利用することはできない。たとえば，CiNiiで提供されているPDFファイルのような画像を結合することで生成されたPDFファイルでは，テキストの抽出はできなくなる。

PDFファイルからのテキスト抽出では，テキストのレイアウトに関する情報は除去されることが多い。本研究では，PDFファイルから機械的にテキストを抽出し扱うことができるかどうかを検討しているため，抽出されたテキストがPDFファイルに記録されたそのものを再現できているかどうかに関しては考慮していない。PDFファイルからのテキスト抽出にはPDFBox0.7.4[39]を用いた。

PDFファイルは，内容に対する暗号化によって，テキストや画像のコピー・編集や，文書の印刷，注釈の作成などに対してセキュリティを設定することができる[40]。これらのセキュリティの設定が，PDFファイルに対する索引化に影響を与えていることが考えられる。そこで，「暗号化」では，PDFファイルへの操作に対する何らかのセキュリティが設定されているかどうかと検索エン

ジンからのアクセス可能性を調査した。PDFファイルの解析にはiText2.7.1[41]を用いた。上記したようにPDFファイルに対するセキュリティにはテキストのコピーや印刷など様々なセキュリティの設定が可能であるが，この調査ではPDFファイルに対してかけられているセキュリティの違いについては考慮せず，なんらかのセキュリティが設定されている場合は，暗号化されているとした。

無作為抽出した2万件のURLのうち，実際にファイルをダウンロードすることができたPDFファイル14,196件のうち，Google，Yahoo!，Bingのいずれかの検索エンジンに登録されていたのは，10,197件（71.8%）で，残りの3,999件（28.2%）はいずれの検索エンジンにも登録されていなかった。

PDFファイルの特徴と，検索エンジンからアクセスのアクセス可能性との関係を表6に示した。

テキスト抽出可能なPDFファイルでは，13,912件中の9,951件（71.5%）が検索エンジンからアクセス可能であったのに対して，テキスト抽出ができなかったPDFファイルでは284件中246件（86.6%）が検索エンジンによってカバーされていた。この結果から，テキスト抽出ができないPDFファイルであっても検索エンジンに登録される可能性の高いことが分かった。近年は，検索エンジンは索引化対象となるコンテンツの情報だけではなく，コンテンツに対しするリンクの情報なども利用して索引化している。そのような技術によって，テキストが抽出できないPDFファイルの場合でも，検索エンジンがカバーできていると考えられる。

ファイルの暗号化の有無と検索エンジンからのアクセス可能性について比較すると，暗号化がされていないPDFファイルの場合は，8,273件中6,712件（81.1%）が検索エンジンからアクセス可能であった。一方なんらかの暗号化がなされているPDFファイルでは，5,923件中の3,466件（58.8%）が検索エンジンからアクセス可能であった。この結果から，PDFファイルに対する暗号化がないほうが検索エンジンからのアクセス可能が高まる傾向にあることが分かった。

PDFファイルからの「テキスト抽出」と，PDFファイルに対する「暗号化」によるセキュリティのそれぞれの特性と，検索エンジンへの登録との関係をカイ二乗検定によって分析した。その結果，テキスト抽出の可否（$\chi^2(1)=30.6$，$p<0.01$）と暗号化の有無（$\chi^2(1)=846.7$，$p<0.01$）とは，検索エンジンからのアクセス可能性との間に，99%水準で有意な関連が見られた。

4. 深層ウェブの実態とその要因

本稿では，機関リポジトリに登録された文献のURLを用いて，深層ウェブの実態の調査を行った。その結果，最もカバー率の高かったGoogleで53.2%，もっとも低かったBingでは28.6%で，検索エンジンは必ずしもすべてのコンテンツを索引化できておらず，かなりの割合の深層ウェブが生じていることが明らかになった。

Google，Yahoo! Japan，Bingのそれぞれで索引化している情報源の範囲は異なり，3つの検索エンジンいずれかが索引化している比率は72.0%で，最もカバー率が高かったGoogleで53.2%であった。

先行研究での結果と比較すると，本研究で最もカバー率が高かったGoogleに関しては，2006年のマッカウンらの調査，2008年のハーゲドーン

表6 PDFファイルの特徴と検索可能性

		検索不可		検索可	
		件数	%	件数	%
テキスト抽出	可	3,961	28.5%	9,951	71.5%
	不可	38	13.4%	246	86.6%
暗号化	なし	1,561	18.9%	6,712	81.1%
	あり	2,438	41.2%	3,485	58.8%
合計		3,999	28.2%	10,197	71.8%

とサンテッリの調査ともに，深層ウェブの比率は56%（カバー率が44%）であったのに対して，2009年の本調査では46.8%（カバー率が53.2%）と大幅に減少していた。

一方，3つの検索エンジンを使った場合を比較すると，マッカウンらの調査では21%であったのに対して，本調査では28.0%で深層ウェブの比率は増加している傾向があった。

本研究では，先行研究ではなされてこなかった，コンテンツが深層ウェブとなる原因の調査も実施した。その結果，ロボット排除プロトコル（robots.txt）で全ての検索エンジンからのクローリングを排除している事例が見られた。ロボット排除プロトコルによる，検索エンジンのロボットの排除には，ウェブコンテンツ全般を考えた場合はコンテンツの公開範囲の制御のような理由が考えられ，ロボットを排除することが望ましくないというわけではない。しかし，本研究が対象とした学術情報の場合，公開の意義について考えると，公開したコンテンツを容易にアクセスできるようにすることは重要な要素であり，ロボット排除プロトコルによって検索エンジンからのクローリングを除外することは望ましいこととは言えないだろう。

全文ファイルのURLからは，URLの長さと動的なURLが検索エンジンからのカバー率に影響を与えていることが明らかになった。URLの構造は，利用している機関リポジトリソフトウェアによって規定されている場合が多いことが伺われ，即座に改善することは難しいかもしれないが，動的なURLを静的なURLに置き換えることなどの効果は指摘されており，深層ウェブとなっているファイルを検索可能にするためには有効であるだろう。また，URLのディレクトリ階層数が3の場合に，他に比べて検索できない比率が高かった。その要因と対策は本調査からは明らかではない。今後，他のコンテンツでのアクセス可能性を調査することによって，これが機関リポジトリに特有の傾向かどうかを明らかにできるだろう。

またPDFファイルの調査から，PDFに対する暗号化処理がなされている場合，暗号化をしていないPDFファイルと比べて，検索エンジンからのカバー率が落ちることが分かった。PDFファイルに対するセキュリティの設定は，著作権への配慮や，論文ファイル中に記載されたメールアドレスを収集するロボットへの対策などが考えられる。一方で，暗号化していなければカバー率が向上し検索エンジンからのアクセスを提供できていた可能性も考えられる。PDFファイルへのセキュリティの設定に関しては，これら両面でのバランスを検討した上で設定する必要があるだろう。

以上，本稿では，機関リポジトリに収録された文献のURLを用いて，深層ウェブの実態を明らかにし，さらに文献のURLとファイルの特徴を分析することで深層ウェブとなっている要因を明らかにした。

本研究の結果から，深層ウェブはかなりの割合で存在していることが明らかになった。また，要因の分析から，様々な要因によって，深層ウェブとなることが分かった。

今後，情報検索における検索エンジンの存在はますます大きくなるであろうし，またウェブでの情報公開に対する要請も益々上がってくるだろう。そのような状況において，ウェブ上に情報を公開する際に深層ウェブになることを避けるためには，より細かな配慮が必要になってくるであろう。

謝　辞

この研究は，日本学術振興会科学研究費補助金基盤研究（B）「ウェブ上の文書から学術論文を自動判定し，検索するシステムの設計開発」（平成21年度～平成23年度，研究代表者：上田修一）によって行った。

注・引用文献

1) Bergman, Michael K. “The deep Web: surfacing hidden value,” Journal of Electronic Publishing. Vol. 7, No. 1, 2001, http://dx.doi.org/10.3998/3336451.0007.104, (参照 2011-11-02)
2) 前掲 1)
3) BrightPlanet. “BrightPlanet unveils the ‘Deep’ Web: 500 times larger than the existing Web,” http://web.archive.org/web/ 20080915150729/http://www.brightplanet.com/news/prs/deep-web-500-times-larger.html, (参照 2011-11-02)
4) Madhavan, Jayant, *et al.* “Google’s Deep-Web crawl,” Proceedings of the VLDB Endowment. Vol. 1, Issue 2, 2008, 1241-1252.
5) クローラー，スパイダーなどとも呼ばれるが，

本稿では，検索エンジンがコンテンツの収集に用いているプログラムを「ロボット」とした。

6) O'Neill, Edward T., *et al.* "Trends in the Evolution of the Public Web 1998 - 2002," D-Lib Magazine. Vol. 10, No. 6, 2004, http://www.dlib.org/dlib/april03/lavoie/04lavoie.html, (参照 2011-11-02)

7) Sherman, Chris and Price, Gary. The Invisible Web: uncovering information sources: search engines can't see. Medford, N.J., Information Today, 1998, xxix, 439 p.

8) 警察庁．不正アクセス行為の禁止等に関する法律．http://www.npa.go.jp/cyber/legislation /gaiyou/houann.htm, (参照 2011-11-02)

9) Lawrence, Steve and Giles, C. Lee. "Accessibility of information on the web," Nature, Vol. 400, No. 8, p. 107-109.

10) WorldWideWebSize.com. The size of the World Wide Web (The Internet). http:// www.worldwidewebsize.com/, (参照 2011-11-02)

11) Alpert, Jesse and Hajaj, Nissanj. "We knew the web was big...," The Official Google Blog. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, (参照 2011-11-02)

12) 前掲 7)

13) Senellart, Pierre, *et al.* "Understanding the hidden Web," ERCIM NEWS, No. 72, p. 32-33.

14) Wikipedia ではその他にも Deepnet, DarkNet, Undernet とも記されている。http://en.wikipedia.org/w/index.php?title=Invisible_Web&oldid=457606182, (参照 2011-11-02)

15) Connaway, Lynn Silipigni and Dickey, Timothy J. The Digital Information Seeker: Report of the Findings from Selected. OCLC, RIN, and JISC User Behaviour Projects. Bristol, JISC, 2010, 61p. http://www.jisc.ac .uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf, (参照 2011-11-02)

16) Research Information Network. Researchers and discovery services: Behaviour, perceptions and needs. 2006.11, 113p. http://www.rin. ac.uk/system/files/attachments/Researchers-discovery-services-report.pdf, (参照 2011-11-02)

17) Niu, Xi, *et al.* "National study of information seeking behavior of academic researchers in the United States," Journal of the American Society for Information Science and Technology. Vol. 61, No. 5, 2010, p. 869-890.

18) 一般的に，「メタサーチ」は「複数のサーチエンジンを横断的に検索するサーチエンジン」と捉えられるがヘミンガーらはこの「多数の情報源を一括して検索する機能」に対して，「メタサーチ」という用語を用いている。

19) Hemminger, Bradley M., *et al.* "Information seeking behavior of academic scientists," Journal of the American Society for Information Science and Technology. Vol. 58, No. 14, 2007, p. 2205-2225.

20) Davis, Philip M. "Open access, readership, citations: a randomized controlled trial of scientific journal publishing," FASEB Journal. Vol. 25, No. 7, 2011, p. 2129-2134.

21) Organ, Michael. "Download statistics - What do they tell us?: the example of research online, the open access institutional repository at the University of Wollongong, Australia," D-Lib Magazine. Vol. 12, No. 11, 2006, http://www.dlib.org/dlib/november06/organ/11organ.html, (参照 2012-02-05)

22) 佐藤翔，逸村裕．"機関リポジトリ収録コンテンツにおける利用数とアクセス元，アクセス方法，コンテンツ属性の関係，" 三田図書館・情報学会研究大会発表論文集，2009，p. 9-12.

23) Lagoze, Carl, *et al.* The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, (参照 2011-11-02)

24) McCown, Frank, *et al.* "Search engine coverage of the OAI-PMH corpus," IEEE Internet Computing. Vol. 10, No. 2, 2006, p. 66-73.

25) Hagedorn, Kat and Santelli, Joshua. "Google still not indexing hidden Web URLs," D-Lib Magazine. Vol. 14, No. 7/8, 2008, http://www.dlib.org/dlib/july08/hagedorn/07hagedorn.html, (参照 2011-11-02)

26) メタデータ・フォーマット junii2. http://www.nii.ac.jp/irp/archive/system/junii2.html, (参照 2011-11-02)

27) NETMARKETSHARE. Top Search Engine Share Trend. http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=5, (参照 2011-11-02)

28) The Web Robots Pages. http://www. robotstxt.org/, (参照 2011-11-02)

29) Kelly, Brian and Peacock, Ian. WebWatching UK Web Communities: Final Report For The WebWatch Project British Library Research and Innovation Centre, 1999. http://www.ukoln.ac.uk/web-focus/webwatch/reports/final/rtf-html/report.html, (参照 2011-11-02)

30) Drott, M. Carl. "Indexing aids at corporate websites: the use of robots.txt and META tags," Information Processing and Management. Vol. 38, No. 2, 2002, p. 209-219.

31) Sun, Yang, *et al.* "Determining bias to search engines from robots.txt," Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007, p. 149-155.

32) 国立情報学研究所．機関リポジトリ一覧．http://

www.nii.ac.jp/irp/list/, (参照 2011-11-02)
33) Network Working Group. Hypertext Transfer Protocol - HTTP/1.1. http://www.ietf.org/rfc/rfc2616.txt, (参照 2011-11-02)
34) Microsoft. [IE] URL に使用可能な文字数は最大 2,083 文 字 . http://support.microsoft.com/kb/208427/ja, (参照 2011-11-02)
35) 前掲 4)
36) Spencer, Stephan. "Underscores are now word separators, proclaims Google," CNET News. http://news.cnet.com/8301-13530_3-9748779-28.html, (参照 2011-11-02)
37) Google. 『検索エンジン最適化スタートガイド』. California, Google Inc. 2010, 32p. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.co.jp/ja/jp/webmasters/docs/search-engine-optimization-starter-guide-ja.pdf, (accessed 2011-10-30)
38) 前掲 4)
39) Apache PDFBox: Java PDF Library. http:// pdfbox.apache.org/, (参照 2011-11-02)
40) Adobe Systems Inc. PDF Reference sixth edition: Adobe Portable Document Format version 1.7. November 2006. San Jose, Adobe Systems Inc. http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/pdf_reference_1-7.pdf, (参照 2011-11-02)
41) iTextR: Free / Open Source PDF Library for Java and C#. http://itextpdf.com/, (参照 2011-11-02)

The Extent of the Deep Web in Japanese Institutional Repositories

Yosuke MIYATA
Keio University
Teru AGATA
Asia University
Atsushi IKEUCHI
University of Tsukuba
Emi ISHITA
Kyushu University
Shuichi UEDA
Keio University

The more the size of Web increases, the more serious the problem of the deep Web (the Web not accessible to search engines) becomes. McCown *et al.* (2006) and Hagedorn & Santelli (2008) surveyed extent of deep Web using metadata contained in institutional repositories. In this research, applying the method used in that previous work, we measured the extent of the deep Web on a larger scale using PDF file URLs contained in institutional repositories in Japan in September 2009. The results show that the coverage rate of major search engines (Google, Yahoo! and Bing) is 72%, leaving 28 % as the maximum extent of the deep Web. And examination of the characteristics of the files revealed that dynamic URLs and longer URLs are associated with decreased coverage rates for search engines.