日本の機関リポジトリに収録された学術情報のアクセス可能性

宮田洋輔(慶應義塾大学)†　　安形輝(亜細亜大学)
池内淳(筑波大学)　　上田修一(慶應義塾大学)
†miyayo@slis.keio.ac.jp

抄録:機関リポジトリに収録された文献の少なくない数が，深層ウェブ化していることが明らかになっている。そこで本研究では，その原因を明らかにするために，日本の機関リポジトリとリポジトリに収録された学術情報のアクセス可能性に関する調査をおこなった。本調査の結果から，robots.txt によって，検索エンジンからのアクセスを排除している事例があること。また，pdf ファイルのテキスト抽出の可否，全文 URL の長さなどの要因が，学術情報へのアクセスの可能性を低めていることが示唆された。

1. はじめに

機関リポジトリは大学や研究所による学術情報資源の公開・蓄積のために設置されている。Lynch は，機関リポジトリを「機関とそのコミュニティの構成員によって作成された電子資源の管理と発信のために，大学がそのコミュニティの構成員に提供する一連のサービス」と定義し，「大学がコミュニティの構成員と大衆とに向けてその責務を示す手段」であり，「より広い世界へと大学の貢献を構成する新しいチャネル」であると述べた[1)]。

現在，日本では，国立情報学研究所の学術機関リポジトリ構築連携支援事業の支援などによって，大学や研究所などの 111 機関によって 115 のリポジトリが構築・運用されている[2)]。

機関リポジトリに収録された学術情報へのアクセス手段として， 1)学術情報に直接，2)機関リポジトリ経由，3)検索エンジン経由，4)横断検索システムなど経由，の 4 つの方法が考えられる。

学術情報資源に直接アクセスする場合，利用者は，何らかのの形で情報資源のウェブ上での識別子 URL を知っている必要がある。

機関リポジトリ経由でアクセスする場合，利用者はそのリポジトリの存在を知っており，URL をアドレスバーに直接入力するか，ブックマークなどによってアクセスし，リポジトリ内の検索システムを用いて，利用者の望む情報資源にアクセスする。

検索エンジンを経由して，リポジトリの学術資源にアクセスする際には，利用者は日常的なウェブ上での情報検索と同様に，各種の検索エンジンでキーワードを入力し，検索結果のなかで機関リポジトリに収録された情報資源と遭遇する可能性がある。その場合には，リポジトリ内の情報が，各種の検索エンジンによって，登録されている必要がある。

横断検索システムなどによるアクセスの場合，機関リポジトリが，たとえば OPEN DOAR[3)]や OAIster[4)]，日本の国立情報学研究所の JAIRO のようなサービス・プロバイダーに登録し，メタデータのハーベスティング(刈り取り)を可能な状態にしておく必要がある。

佐藤らによる機関リポジトリのアクセスログ研究によると，機関リポジトリに収録された文献の半数近くへのアクセスが検索エンジン経由でおこなわれていることが明らかになっている[5)]。

一方，機関リポジトリに収録された学術情報の深層ウェブ(インターネット上に存在していても検索エンジンによってアクセスできないウェブ)化も指摘されている。OAIster からの標本データを用いて検索エンジンでの登録率を調査した McCown ら[6)]，Hagedorn ら[7)]による調査では，機関リポジトリのメタデータ中の半数程度しか，検索エンジンによって，カバーされていないことが明らかになっている。また安形ら[8)]がおこなった，日本の機関リポジトリのメタデータから抽出した全文ファイルの登録状況の調査でも，Google，Yahoo!，Bing のいずれかの検索エンジンに登録されたものが 7 割，佐藤らの調査[5)]でもっともアクセスの割合が多かった Google だけでは 53.2%と，検索エンジン経由でのアクセスが十分に機能していないことが明らかになっている。

そこで，本研究では，機関リポジトリに収録された学術情報へのアクセスの問題の要因を明らかにするために，機関リポジトリ自体とそこに収録された学術情報ファイルの調査をおこなった。

2. 調査の概要

表 1 に，1,000 件以上の全文データを持った機関リポジトリでの，Google，Yahoo!，Bing のいずれかの検索エンジンからのアクセス可能な割合の上位 10 機関を，表 2 に下位 10 機関を示した。表から，検索エンジンでの登録率が 100%のリポジトリも存在するものの，登録率が高いリポジトリでも必ずしも 100%

表 1 検索エンジンからのアクセス可能性の高い 10 機関

機関名	全文	全エンジン	%	いずれか	%
1 早稲田大学	13,808	1,408	10.2%	13,808	100.0%
2 岐阜大学	4,030	359	8.9%	4,027	99.9%
3 お茶の水女子大学	17,638	9,658	54.8%	17,581	99.7%
4 NAIST	3,632	12	0.3%	3,609	99.4%
5 岡山大学	10,342	1,059	10.2%	10,249	99.1%
6 横浜国立大学	3,068	1,638	53.4%	3,033	98.9%
7 小樽商科大学	2,007	298	14.8%	1,980	98.7%
8 鹿児島大学	4,773	278	5.8%	4,697	98.4%
9 静岡大学	2,740	876	32.0%	2,688	98.1%
10 秋田大学	1,210	267	22.1%	1,185	97.9%

表 2 検索エンジンからのアクセス可能性の低い 10 機関

機関名	全文	全エンジン	%	いずれか	%
1 同志社大学	9,630	0	0.0%	0	0.0%
2 明治大学	1,718	0	0.0%	15	0.9%
3 三重大学	4,562	0	0.0%	129	2.8%
4 高知大学	1,081	0	0.0%	77	7.1%
5 島根大学	4,404	1	0.0%	424	9.6%
6 筑波大学	20,225	39	0.2%	2,738	13.5%
7 大分大学	8,540	52	0.6%	1,304	15.3%
8 北海道大学	28,025	65	0.2%	7,992	28.5%
9 群馬大学	3,485	2	0.1%	1,403	40.3%
10 大阪教育大学	3,000	7	0.2%	1,342	44.7%

ではない事例があること，また，極端に登録率の低いポジトリも存在することが分かる。

これらのリポジトリにおける学術情報資源へのアクセスの障害は，2 つの問題に分割して考えることができる。つまり，機関リポジトリの構築・運用の方針が，外部からのアクセスに関する問題を抱えている場合，そしてもう 1 つが，アクセスの対象となる学術情報資源それぞれの特性が持つアクセスに関する問題である。そこで，本研究では，上記の 2 つの観点から，日本の機関リポジトリに収録された学術情報のアクセス可能性についての調査をおこなった。

機関リポジトリの構築・運用の問題点としては，NII 作成の日本の機関リポジトリのリスト [2]に掲載された機関リポジトリに対して，a) 各リポジトリのロボット排除プロトコル robots.txt の設定から検索エンジンのクローリングへの対処法を調査するとともに，b)対応するメタデータ形式の調査をおこなった。robots.txt の分析から，各機関リポジトリの検索エンジンなどのロボットへの対応を明らかにする。また提供するメタデータ形式の調査によって，日本の機関リポジトリ横断検索システム JAIRO への登録状況を調査する。

一方，ほとんどの学術情報が検索エンジンに登録されているにもかかわらず，一部のデータが登録されていなかった事例においては，登録されていなかった学術情報資源ファイルそのものが持つ特性によることも考えられる。そこで，a)リポジトリに収録された文献の pdf ファイルの形式の分析と b)全文 URL の特性の分析をおこなった。pdf ファイルの分析によって，機関リポジトリに登録された学術情報にセキュリティが設定されていないか、テキストではなく画像データではないかを，URL の分析によって，機関リポジトリによる公開の方法(URL の長さなど)がクローリングを阻害していないか，を調査した。

3. 調査結果

3.1. 構築・運用の問題

構築・運用の問題に関する調査結果を以下に示す。調査対象は，NII 作成の機関リポジトリ一覧に掲載された 115 の機関リポジトリである [2]。

3.1.1. robots.txt の調査

robots.txt は，サイトのトップディレクトリに配置される。115 の機関リポジトリのトップディレクトリから robots.txt を収集した。

robots.txt の有無を表 3 に示した。robots.txt を適切な位置に設置し，ロボットへのなんらかの対応を設定しているリポジトリは 39.1%であった。

表 3 各 IR における robots.txt の有無

内容	IRの数	比率
robots.txtあり	45	39.1%
404エラー	66	57.4%
アクセス不可	4	3.5%
合計	115	100%

つぎに 45 件の robots.txt における内容の集計を表 4 に示した。あらゆる User-agent に対して、クローリングを完全に排除しているリポジトリが存在した。排除の指定がされていたロボットとしては，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 (フリーのクローラ) があった。robots.txt を作成しているものの, 何も記述がない事例も 2 件存在した。

表 4 robots.txt の内容

内容	IRの数	%
User-agent:*のみ	40	88.9%
特定のUser-agentを指定	3	6.7%
何も記述していない	2	4.4%
合計	45	100%

3.1.2. メタデータ形式の調査

つぎに, メタデータ形式の調査の結果を示す。OAI-PMH のベース URL が不明であった 6 リポジを除いて, 109 の機関リポジトリに対して, OAI-PMH の ListMetadataFormats 要求によって, 各リポジトリが提供するメタデータ形式に関する情報を得た。表 5 に日本の機関リポジトリが提供するメタデータ形式を示した。

表 5 日本の機関リポジトリがサポートするメタデータ形式

メタデータ形式	件数	%
oai_dc	108	99.1%
junii2	100	91.7%
junii	45	41.3%
context_object	5	4.6%
uketd_dc	5	4.6%
didl	5	4.6%
mets	5	4.6%
okayama	4	3.7%
mods	3	2.8%
akf	2	1.8%
rdf	1	0.9%
dsrrs	1	0.9%
oai_ir	1	0.9%

1 件を除いて, すべてのリポジトリが oai_dc をサポートしていた。100 リポジトリ (91.7%) が, NII の機関リポジトリ横断検索システム JAIRO のメタデータ形式 junii2 に対応しており, 横断検索システムからのアクセス可能性は確保できている。

3.2. 個々の学術情報資源の問題

個々の学術資源の特性を調査するために, 2009 年 4 月 11 日に junii2 形式でメタデータを取得可能な 92 の機関リポジトリから, メタデータを収集した。

3.2.1. pdf 形式の調査

各リポジトリに収録された学術情報資源 pdf ファイルから無作為抽出した 2 万件の標本を用いて, ファイルの特性の調査をおこなった。ファイルを実際にダウンロードできたのは, 19,729 件で 271 件はダウンロードできなかった。pdf ファイルの分析には, iText2.1.7 と, テキスト抽出のために PDFBox0.7.4 を用いた。

全文 URL から実際にファイルをダウンロードできた 19,729 件のファイルのうち, 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, 13,806 件で, 残りの 5,923 はいずれの検索エンジンにも登録されていなかった。

pdf ファイルからのテキストの抽出の可否と, 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もののほうが 71.0% と検索エンジンからアクセス可能性が高かった。

表 6 テキスト抽出の可否とアクセス可能性との関係

テキスト抽出	登録なし		いずれか	
	件数	%	件数	%
可	3,961	28.5%	9,951	71.5%
不可	1,962	33.7%	3,855	66.3%
合計	5,923	30.0%	13,806	70.0%

つぎに, ファイルの暗号化と検索エンジンからのアクセス可能性について, 表 7 に示した。pdf 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った。

表 7 暗号化の有無とアクセス可能性との関係

暗号化	登録なし		いずれか	
	件数	%	件数	%
なし	3485	25.2%	10321	74.8%
あり	2438	41.2%	3485	58.8%
合計	5923	30.0%	13806	70.0%

両特性に対してカイ二乗検定によって, 検索エンジンへの登録との関係の分析をおこなったところ, テキスト抽出の可否と暗号化の有無とによって、99% 水準で標本間に有意な差が見られた。

3.2.2. 全文 URL の調査

学術情報の全文 URL の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。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。「URL が動的であるか」は URL 中にパラメータを付加するための「?」が存在するかどうかで判断した。「ディレクトリの深さ」は URL 中に含まれる「/」の数によって計測した。

URL の長さの四分位範囲と検索エンジンからのアクセス可能性との関係を表 8 に示した。合計件数に対する割合をみると，URL の長さが第 4 四分位範囲の URL に関しては，いずれの検索エンジンにも登録のないものが多いことが分かる。

表 8 URL 長とアクセス可能性との関係

URL長	登録なし		いずれか	
	件数	%	件数	%
第1	13,803	15.2%	77,209	84.8%
第2	32,012	29.1%	78,037	70.9%
第3	19,245	20.6%	74,179	79.4%
第4	48,347	44.0%	61,599	56.0%
合計	113,407	28.0%	291,024	72.0%

URL が動的かいなか，と検索エンジンからのアクセス可能性との関係を表 9 に示した。表から動的な URL のほうが静的な URL に比べて，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る。

表 9 URL のタイプとアクセス可能性との関係

URLのタイプ	登録なし		いずれか	
	件数	%	件数	%
静的	83,955	23.0%	280,513	77.0%
動的	29,452	73.7%	10,511	26.3%
合計	113,407	28.0%	291,024	72.0%

ディレクトリの深さとアクセス可能性との関係を表 10 に示した。ディレクトリ階層数の中央値 6 以下のものを「浅い」，中央値より大きいものを「深い」とした。わずかではあるが，ディレクトリ階層の深いもののほうが，登録なしの割合が上昇していることが分かる。

表 10 ディレクトリの深さとアクセス可能性との関係

ディレクトリの深さ	登録なし		いずれか	
	件数	%	件数	%
浅い	112,467	28.0%	289,111	72.0%
深い	940	32.9%	1,913	67.1%
合計	113,407	28.0%	291,024	72.0%

4. 議論と考察

機関リポジトリの構築・運用に関する問題と学術情報資源それぞれの特性に関する問題との 2 つの観点から，日本の機関リポジトリに収録された学術情報のアクセス可能性に関する調査をおこなった。調査からは，組織の所属員の研究成果を公表すべきである機関リポジトリは，横断検索システムからのアクセス経路は確保されているものの，検索エンジン経由でアクセス可能であった学術情報はいくつかの要因によってアクセスが阻害された状況にあり，機関リポジトリのなかにはその目的を果たしきれていないものが存在することが明らかとなった。

機関リポジトリに収録された学術情報のアクセス可能性を高めるためには、1)robots.txt による不要な排除をおこなわない，2)pdf ファイルの形式をテキスト抽出可能な形にする，3)pdf ファイルの暗号化を設定しない，4)URL を短くする，5)静的な URL を用いる，などの方策が有効である。

引用文献

1) Lynch, Clifford A. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL: A Bimonthly Report. 2003, no. 226, p. 1-7. http:// www. arl. org/ resources/ pubs/ br/ br226/ br226ir.shtml , (accessed 2009-10-01)

2)学術機関リポジトリ構築連携支援事業. 機関リポジトリ一覧. http://www.nii.ac.jp/irp/list/, (参照 2009-10-01).

3) The Directory of Open Access Repositories - OpenDOAR. http://www.opendoar.org/, (accessed 2009-10-01).

4) OAIster. http://www.oaister.org/, (accessed 2009-10-01).

5) 佐藤翔. 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 年度. 慶應義塾大学, 2009-9-26. 三田図書館・情報学会, 2009, p. 9-12.

6) McCown Frank; Liu, Xiaoming; Nelson, Michael M; Zubair, Mohammed. Search engine coverage of the OAI-PMH corpus. IEEE Internet Computing. 2006, Vol. 10, No. 2, p. 66-73.

7) Hagedorn, Kat; Santelli, Joshua. Google still not indexing hidden web URLs. D-Lib Magazine. 2008, vol. 14, no. 7/8, http://www.dlib.org/dlib/july08/hagedorn/07hagedorn.html, (accessed 2009-10-01).

8) 安形輝, 宮田洋輔, 池内淳, 上田修一. 学術情報流通における深層ウェブの実態: 機関リポジトリに収録された文献を用いた調査. 三田図書館・情報学会研究大会発表論文集 2009 年度. 慶應義塾大学, 2009-9-26. 三田図書館・情報学会, 2009, p. 37-40.