図書館ウェブサイトの公開性
－クローラに対するアクセス制御に関する調査－

安形輝（亜細亜大学）　　agata@asia-u.ac.jp

1. 図書館サイトの公開性

現在、多くの図書館がウェブサイト上で蔵書目録検索を始めとする様々なサービスを提供している。さらに、検索 API などの機械的なアクセスを前提としたサービスを提供している館もある。

一方で、研究者や一般利用者がクローラ（ウェブページを収集するためのプログラム）を用いて、ウェブサイトのコンテンツを機械的に収集することは技術的には容易になりつつある。例えば、図書館・情報学においては所蔵調査にウェブサービスを応用する研究[1)]などを挙げることができる。一般利用者がクローラあるいはそれに類するプログラムを用いる事例も増えている。2010年5月25日には図書館の積極的利用者が自作のクローラによって公共図書館のウェブサイトへアクセスし、偽計業務妨害容疑で逮捕されるという事件、（一般的には岡崎市立中央図書館事件と呼ばれている）が発生している[2)]。

一般的なウェブサイトのアクセス制御に関しては多くの大規模な調査が行われてきた[3)]。しかし、図書館がどのような形でアクセス制御を行っているかの調査は行われていない。図書館サービスの公開性とサーバに対する負荷の点からは、単にクローラを排除するだけではなくバランスの取れたアクセス制御を行うことが好ましい。本研究では日本の公共図書館や大学図書館のウェブサイトを対象として、クローラのアクセス制御に関する調査を行った。

クローラのアクセス制御を行う中で、検索エンジンのクローラをも排除すると、そのウェブサイトが検索エンジンの検索結果に含まれなくなる、あるいは、適切に表示できなくなることがある。一般的な利用者が図書館ウェブサイトにアクセスする中心的な手段が検索エンジンであることを考えると、検索エンジンからアクセスしにくい図書館ウェブサイトは、知る自由を保証するという図書館の使命からは望ましくない（図1）。そこで併せて、主要な検索エンジンで図書館ウェブサイトがどのように検索されるかについても調査を行った。

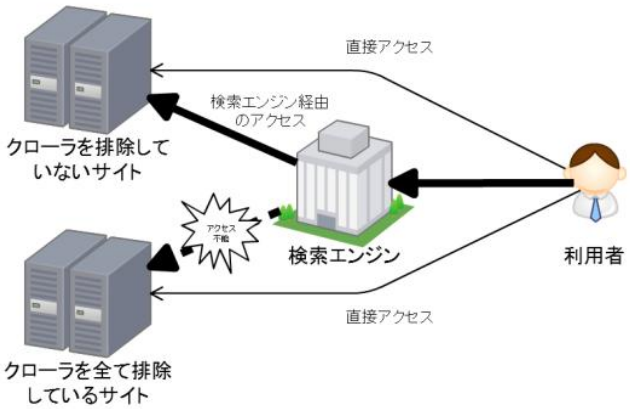

図 1 アクセスしやすさの差異

2. クローラのアクセス制御に関する調査

2.1 調査方法

2.1.1 調査対象館

調査対象とした図書館ウェブサイトは、日本図書館協会の図書館リンク集[4)]に掲載された公共図書館と大学図書館である。このリンク集から図書館名とウェブサイトの URL を取得した。図書館名については括弧等を除去したものを正式名称とした。全部で 2,450 サイトの URL を取得できた。ただし、図書館の異動、サイトの移転等によって必ずしも全ての URL にアクセスできるとは限らないため、ウェブサイトにアクセス可能かを 2010 年 9 月 5 日に調査し、アクセスできた 2,065 館（公共図書館 1,277 館、大学図書館 788 館）を対象とした。

2.1.2 クローラのアクセス制御の調査方法

ウェブサーバに対するクローラのアクセスを制御するにはロボット排除プロトコル(Robots Exclusion Protocol あるいは Robots Exclusion

Standard : 以下、REP)[5]を用いることになる。REP によるクローラのアクセス制御にはいくつかの方法が用意されているが、最も一般的な手法はウェブサイトのトップに robots.txt というファイルを置くことである。なお、robots.txt がないウェブサイトはクローラに対して公開されていることになる。例えば、全てのクローラを排除する robots.txt の内容は図2のようになる。

```
User-agent: *
Disallow: /
```

図2 すべてのクローラを排除する robots.txt

REP 調査は、図書館のウェブサイトのトップにおかれた robots.txt を収集することで行った(2010 年 9 月 5 日)。調査対象とした大学図書館の中にはウェブサイトが他の図書館と一緒に大学サーバ内に置かれていることもある。そのため、収集対象となった robots.txt は 2,065 館ではなく 2,014 件である。収集できた robots.txt について記述内容を調査した。

調査の観点としては、robots.txt に誤りがあるか、クローラを全て排除しているか、である。

2.2 調査結果

2.2.1 robots.txt の有無

表1は調査した図書館ウェブサイトでの robots.txt の割合である。全体で8割弱の図書館はクローラに対する robots.txt を返さない。返さないことは、クローラに対してアクセス制限を行わないことを意味するため、公開性の点からは必ずしも悪いことではない。館種別にみると、公共図書館が 16%であるのに対して、大学図書館は 33%とほぼ 1/3 の図書館のウェブサイトが robots.txt を持っていることがわかる。

表1 robots.txt の有無

	公共図書館		大学図書館		全体	
	館数	割合	館数	割合	館数	割合
なし	1072	83.9%	526	66.8%	1598	77.4%
あり	205	16.1%	262	33.2%	467	22.6%
合計	1277	100.0%	788	100.0%	2065	100.0%

2.2.2 robots.txt の誤り

(1)REP ではなく HTML ページを返す

取得できた robots.txt の中には REP ではなくエラーメッセージ等の書かれた HTML 形式のページが 59 件あった(表2)。存在しないページがリクエストされたさいに、HTTP のレスポンスコードとしてエラーを返さず、エラーメッセージを記載した HTML ページを戻すような設定だと推測される。存在しない robots.txt がリクエストされた時にも同様に HTML ページを戻すことになる。

表2 robots.txt の内容

	公共図書館		大学図書館		全体	
	館数	割合	館数	割合	館数	割合
アクセス制御	185	90.2%	223	85.1%	408	87.4%
エラーページ	20	9.8%	39	14.9%	59	12.6%
合計	205	100.0%	262	100.0%	467	100.0%

クローラは解釈できない robots.txt があったときにアクセス制限されていないと判断すると予想される。そのため、ここでは HTML ページであった場合には、アクセス制限がかけられていないウェブサイトと解釈した。ただし、エラーページに REP に関連する文字列が入る場合には予想外のアクセス制限がされてしまう恐れもあるため、あまり望ましい状態ではない。

(2) 文法的な誤り(致命的な間違い)

岡崎市立図書館を始めとする公共図書館 17 館のウェブサイトは全てのクローラからの全てのアクセスを排除していた("Disallow:/")。しかし、それに関わらず、サイト内の他の部分に関しても排除するための Disallow 文を詳細に記述していた。これはサイト一部へのアクセスは許可しようとしたが、REP への理解不足により、失敗している事例だと考えられる。17 館のうち、特に岡崎市立図書館は223行と最も長く詳細な robots.txt を記述している。

17 館の robots.txt の記述はほぼ同様の誤りを示している。図書館ウェブサイトの構築を同じ業者(三菱電機インフォメーションシステムズ)に委託していることから、その業者が robots.txt に誤った記述を行ったものと推測される。なお、この 17 館以外に、元々は同じ業者によると推測される robots.txt であるが、誤りが修正されている図書館が5館(下呂市はぎわら図書館、石狩市民図書館、藤枝市立図書館、長岡京市図書館、東村山市図書館)あった。

(3) 些細な誤りあるいは特殊なアクセス制御

robots.txt を調査する中で、些細な間違いあるいは、特殊なアクセス制御を行っている事例が見つかった。いくつか特徴的な事例について紹介しておく。

・日本文理大学図書館は多くのクローラからのアクセスを許可しているが、検索エンジンBaidu と国会図書館のクローラ(ndl-japan)からのアクセスは全て拒否している。
・国立国会図書館からのクローラのアクセスを許可した上で、他のすべてのクローラからのアクセスを排除している図書館2館（高専、大学図書館）があった。これは国会図書館のガイドライン[6]通りである。ただし2館の親組織である学校は近年、他に吸収合併されており、ウェブサイトも休止状態である。
・奈良先端科学技術大学院大学図書館のウェブサイトは各検索エンジンからのアクセスは許可した上で、/index.html へのアクセスを禁止している。実際に index.html にアクセスすると"403 Not Found"が返ってくるため、その対策と推測される。しかし、例えば Google では検索できてしまうため、あまり意味をなしていない(index.html へのアクセスをリダイレクトする方が有効な対策である)。また、国会図書館のクローラ限定で/library/以下へのアクセスを許可しているが、2010 年 9 月 5 日現在/library/へのアクセスはできない。

2.2.3 クローラを全て排除している図書館

クローラによるアクセスを認めている図書館、すべて排除している館をまとめたものが表3になる。ここで、ウェブサイトトップに robots.txt がなかった館、エラーページを返す館はクローラに対して公開されていると考え、「完全・一部公開」に含めた。また、、何らかのアクセス制御がされていても主要な検索エンジンのクローラからのアクセスは受け入れている図書館も「完全・一部公開」に含めた。それ以外の全てのクローラからのアクセスを排除している図書館を「全て排除」に含めた。

全てのクローラの全てのアクセスを排除している図書館は全体で 71 館（3.5%）存在する。そのうち、本来、開かれた図書館であるはずの公共図書館において、ウェブサイトがクローラを全て排除している館が 61 館（公共図書館で 4.8%）であった。特に、前述の robots.txt の記述に誤りがある17館を除く、44館はクローラからのアクセスを積極的に排除する設定にしている。クローラからの大量のアクセスに対して負荷の点からウェブサーバを守るという考えだと思われる。しかし、公立機関にも関わらず国会図書館のクローラをも排除していることは問題がある。さらには、利用者からの検索エンジン経由のアクセスをも阻害している可能性が高い。そこで、検索エンジンによる検索可能性調査を行った。

表3 クローラを全て排除している図書館

	公共図書館		大学図書館		全体	
	館数	割合	館数	割合	館数	割合
公開・一部公開	1216	95.2%	778	98.7%	1994	96.6%
全て排除	61	4.8%	10	1.3%	71	3.4%
合計	1277	100.0%	788	100.0%	2065	100.0%

3. 検索エンジンによる検索可能性調査

3.1 調査手法

REP によって検索エンジンのクローラによるアクセス制限がされている場合、検索エンジンはそのウェブサイト内のページ収集ができなくなる。しかし、その場合でも検索エンジンでまったく検索できなくなることは少ない。なぜなら、該当ページへのリンク（アンカータグ内のテキスト等）を用いて、ページの情報をある程度収集することが可能だからである。しかし、そのようにして収集された情報は不十分なことが多い。例えば、検索結果においてタイトル情報が誤っていたり、要約（サマリー）の表示がされなくなる。

ここでは図書館のウェブサイトが検索されやすいか、きちんと検索されるかを調べるために検索エンジンの調査も行った。

調査対象とした検索エンジンは日本における主要な検索エンジンとした。インターネット調査会社であるネットレイティングスによれば、日本の検索回数における検索エンジンシェアは 2010 年 5 月時点で Yahoo!（53.2%）、Google（37.2%）、Bing（2.5%）となっている[7]。そこで上位2つの Yahoo!と Google を用いた。

図書館名称を検索式とした場合と URL を検索式とした場合に出力された検索結果上位 8 位までを調査した。各検索結果についてはタイトル、URL、要約のデータを取得した。Yahoo!の検索は Yahoo!デベロッパーズネットワークのウェブ検索[8]、Google の検索は Google AJAX Search API[9]を通じて行った。

検索エンジンに対する調査は、該当する図書館ウェブサイトの URL が上位 8 位までの検索結果として出力されたか、出力されたさいに、

タイトル、URL、要約が正しい形で出力されたかといった点から集計した。

3.2 調査結果

表4と表5はYahoo! JapanとGoogleを用いて図書館ウェブサイトを検索した結果をまとめたものである。Yahoo! Japan ではクローラを全て排除している図書館サイトを検索しても、98.6%と1館を除く全ての館について8位以内に検索できない。クローラの排除によって図書館ウェブサイトの情報を取得できなかったことが大きく影響している。 一方で Google では排除されている図書館でも上位に検索される。これは、Google がリンクを活用しているためだと推測される。

表4 Yahoo! Japan の検索順位

Yahoo		クローラによるアクセス			
		排除なし		全排除	
		結果数	割合	結果数	割合
検索結果順位	1	1601	80.3%	0	0.0%
	2	167	8.4%	1	1.4%
	3	34	1.7%	0	0.0%
	4	11	0.6%	0	0.0%
	5	2	0.1%	0	0.0%
	6	5	0.3%	0	0.0%
	7	2	0.1%	0	0.0%
	8	0	0.0%	0	0.0%
9位以下		172	8.6%	70	98.6%
合計		1994	100.0%	71	100.0%

表5 Google の検索順位

Google		クローラによるアクセス			
		排除なし		全排除	
		結果数	割合	結果数	割合
検索結果順位	1	1515	76.0%	58	81.7%
	2	127	6.4%	0	0.0%
	3	72	3.6%	2	2.8%
	4	8	0.4%	0	0.0%
	5	3	0.2%	0	0.0%
	6	4	0.2%	0	0.0%
	7	4	0.2%	0	0.0%
	8	2	0.1%	0	0.0%
9位以下		259	13.0%	11	15.5%
合計		1994	100.0%	71	100.0%

Google ではクローラによるアクセスを全て排除している図書館であっても検索可能である。しかし、Google も必ずしも正しい検索結果が得られているわけではない。順位 1 位の検索結果について、要約がきちんと出力されたかを示したのが表6である。ここで、クローラの排除を行っていない図書館では 99.3%とほぼ全ての図書館サイトの説明がきちんと出力された。一方、排除を行った図書館では 81.0%と多くのサイトの要約が出力されなくなった。

表6 Google の検索順位1位での要約出力

	クローラによるアクセス			
	排除なし		全排除	
	結果数	割合	結果数	割合
要約あり	1505	99.3%	11	19.0%
要約なし	10	0.7%	47	81.0%
計	1515	100.0%	58	100.0%

4. まとめ

日本の図書館のほとんど(96.6%)はクローラによるアクセス制限は行っていないか、行っていても、主要な検索エンジンのクローラからのアクセスは認めていることが調査から明らかとなった。ただし、一部の図書館はクローラからのアクセスを全て排除している。排除している図書館については、検索エンジンの調査を行い、検索エンジン経由でのアクセスしやすさに問題が生じていることが明らかとなった。

【注・引用文献】

1) 大場博幸ほか. "図書館はどのような本を所蔵しているか". 日本図書館情報学会第 58 回研究大会.

2) "図書館HPにアクセス3万3千回業務妨害容疑、38歳を逮捕 愛知県警". 朝日新聞. 2010年5月26日朝刊など大手新聞他、"岡崎市立中央図書館 検索システムが過負荷でダウン 利用者が逮捕される". 日経コンピュータ 2010/8/4 号 p.78-80 など一般雑誌にも掲載された

3) 例えば、Santanu Kolay , Paolo D'Alberto , Ali Dasdan , Arnab Bhattacharjee, A larger scale study of robots.txt, Proceeding of the 17th international conference on World Wide Web, April 21-25, 2008, Beijing, China

4) "図書館リンク集". http://www.jla.or.jp/link/index.html

5) "The web robots pages" http://www.robotstxt.org/

6) "改正国立国会図書館法によるインターネット資料の収集について". http://warp.da.ndl.go.jp/bulk_info.pdf

7) "月間検索クエリ数のトップは Yahoo! Search で 23 億 6903 万回". http://www.netratings.co.jp/New_news/News05272010.htm

8) "Yahoo!デベロッパーネットワーク ウェブ検索". http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

9) "Google AJAX Search API". http://code.google.com/intl/ja/apis/ajaxsearch/