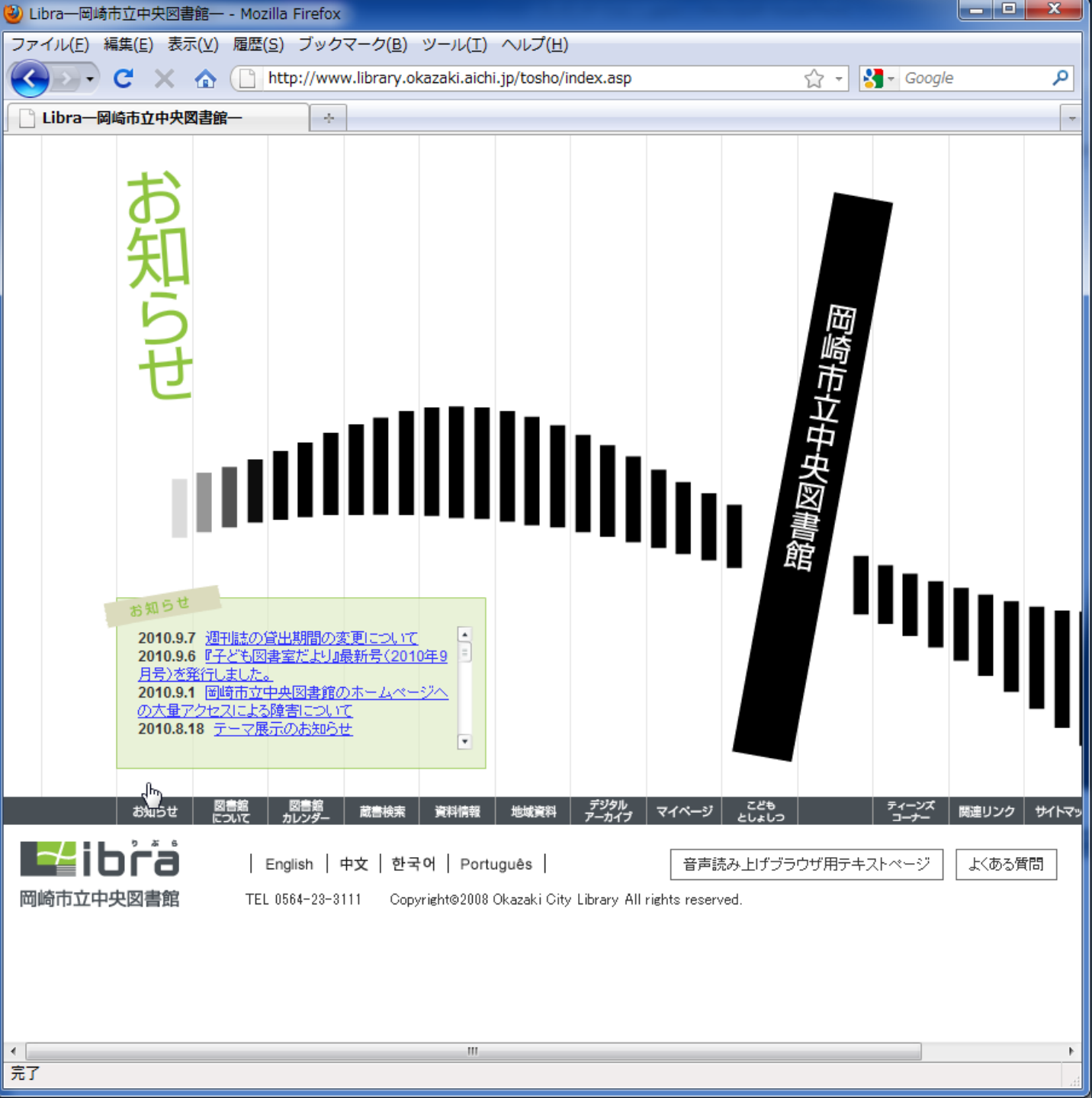
図書館ウェブサイトの公開性

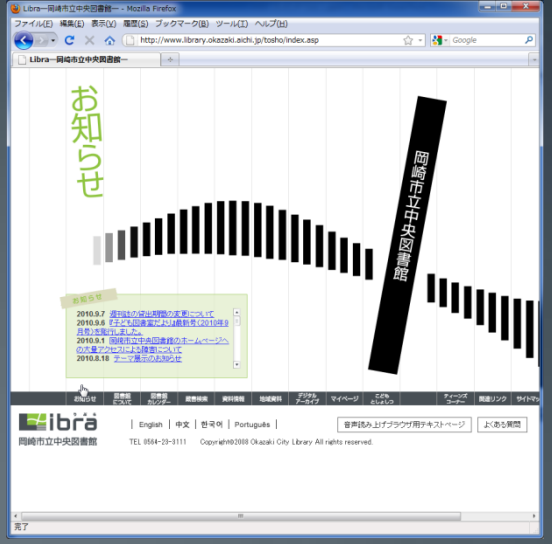
クローラに対するアクセス制御に関する調査

安形輝（亜細亜大学）
agata@asia-u.ac.jp

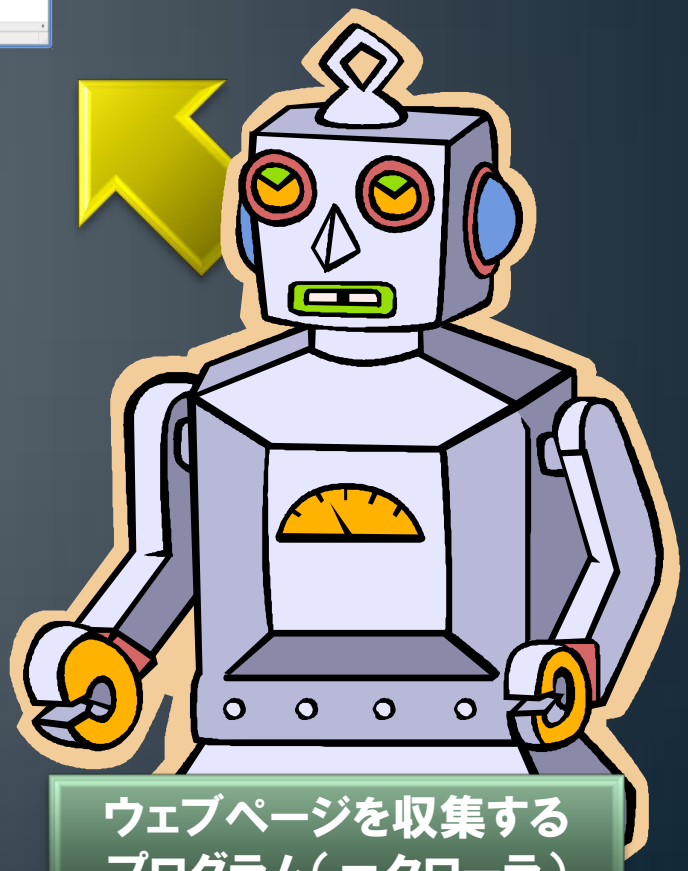


サービスアクセス手段

重要な広報手段



利用者

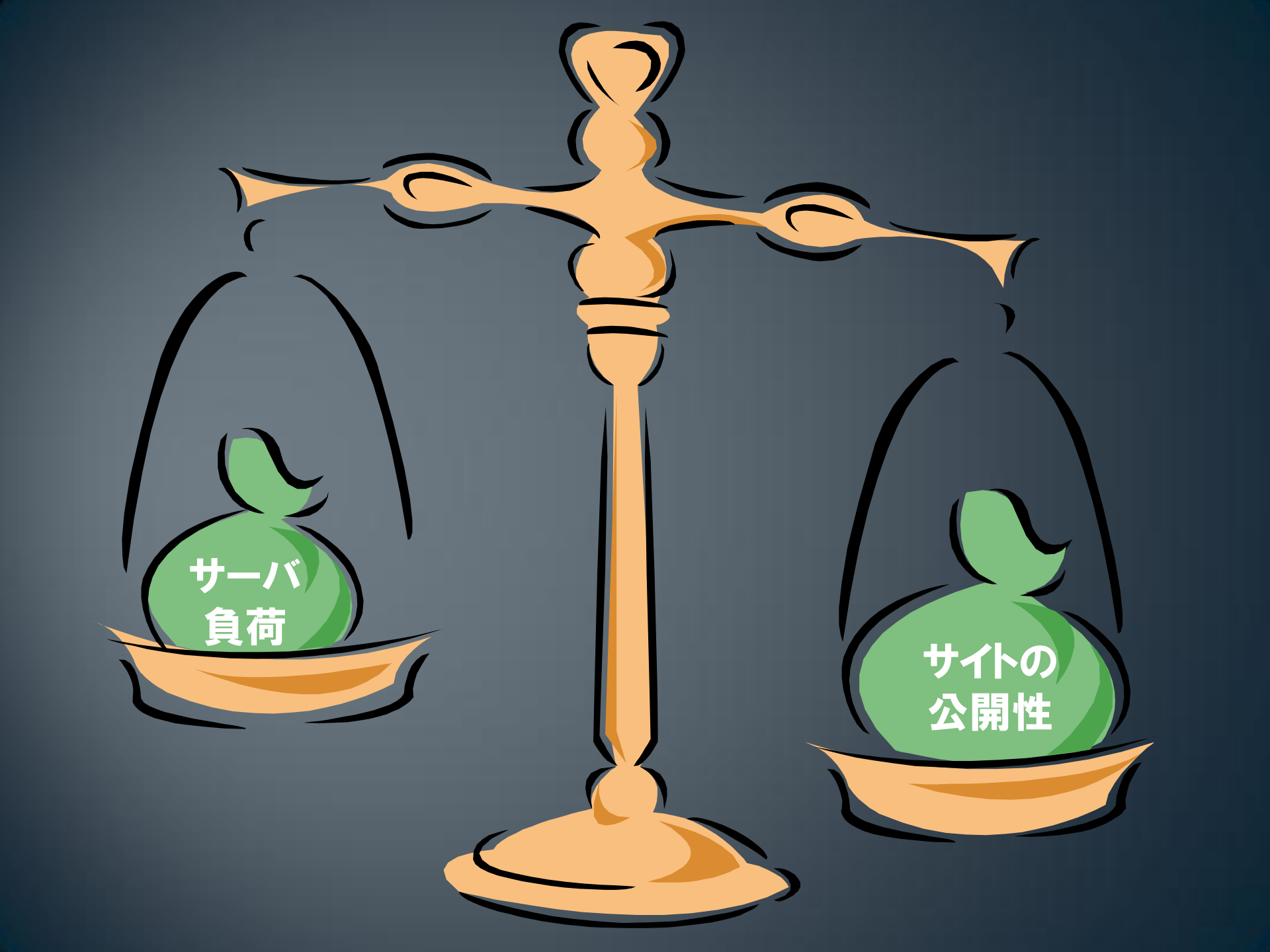


ウェブページを収集する
プログラム（＝クローラ）

2010年6月

Libra（りぶら）

岡崎市図書館交流プラザ


サーバ
負荷
サイトの
公開性

A Large-Scale Study of Robots.txt

Yang Sun, Ziming Zhuang, and C. Lee Giles
The Pennsylvania State University
University Park, PA, USA
{ysun, zzhuang, giles}@ist.psu.edu

ABSTRACT

Search engines largely rely on Web robots to collect information from the Web. Due to the unregulated open-access nature of the Web, robot activities are extremely diverse. Such crawling activities can be regulated from the server side
by deploying the Robots E
robots.txt. Although it is n
ical robots (and many comm
fied in robots.txt. With ou
7,593 websites from educati
ness domains. Five crawls h
to study the temporal chan
of the data, we present a st
rules at the Web scale. The
of robots.txt has increased

General Terms

Experimentation, Measurement.

Keywords

crawler, robots exclusion protocol, robots.txt, search engine.

1. INTRODUCTION

Without robots, there would probably be no search engines. Web search engines, digital libraries, and many other web applications such as offline browsers, internet marketing software and intelligent searching agents heavily depend on robots to acquire documents. Robots, also called "spiders", "crawlers", or "bots", are self-acting agents that navigate around-the-clock through the hyperlinks of the Web, harvesting topical resources at zero costs of human management [4]. Because of the highly automated nature of the robots, rules must be made to regulate such crawling activities in order to prevent undesired impact to the server workload or access to non-public information.

The Robots Exclusion Protocol has been proposed [3] to provide advisory regulations for robots to follow. A file called robots.txt, which contains robot access policies, is deployed at the root directory of a website and accessible to all robots. Ethical robots read this file and obey the rules during their visit to the website. Despite the criticality of the robots.txt convention for both content providers and harvesters, little work has been done to investigate its usage in

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

detail, especially at the scale of the Web. A study
usage of robots.txt in UK universities and colleges
tigated 163 websites and 53 robots.txt [2]. Robots.t
were examined in terms of file size and the use of Robo
clusion Protocol within the UK university domains.
[1] studied the usage of robots.txt as an aid for index
samples from Fortune Glob

ent the first large-scale st
e domains of education, g
We present our observati
data than previous studie

TION

llect the initial URLs to fe
tory Project (DMOZ). O
three domains: education,
and government. The university domain is further
down into the American, European, and Asian uni
domains. We use the Fortune Top 1000 Company I
our data source in the business domain. Our crawl
performed five crawls for the same set of websites b
Dec. 2005 and Oct. 2006.

3. RESULTS

Statistics: We crawled and investigated 7,593
websites including 600 government websites, 2,047 ne
per websites, 1,487 USA university websites, 1,420
pean university websites, 1,039 Asian university we
and 1,000 company websites.

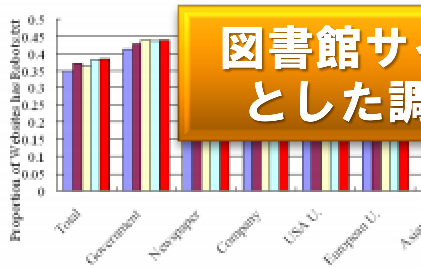


Figure 1: Probability of a website that
robots.txt in each domain.

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，「
大衆とに向けてその貢
広い世界へと大学の貢
であると述べた[1]。
現在，日本では，国
ポジトリ構築連携支援
学や研究所などの 111
が構築・運用されてい

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
システムを用いて，利
スする。
リポジトリの学術資源に
は日常的なウェブ上で
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
ン経由でおこなわれてい
[5]。
収録された学術情報の深
に存在していても検索エ
きないウェブ）化も指摘さ
標本データを用いて検索
査した McCown ら[6]，
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%

インターネット全体が対象のクローラのアクセス排除に関する調査は多い

機関リポジトリについては、我々の研究グループで行った例がある

図書館サイトを対象とした調査はない

図書館サイトは
クローラの

アクセスを
排除して
いるか

その影響は
あるか

図書館サイトはクローラのアクセスを排除しているか

ロボット排除プロトコル

複数の方法／robots.txtが事実上の標準

例えば、robots.txtの記述例

全コンテンツ排除

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

一部コンテンツ排除

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

調査対象館

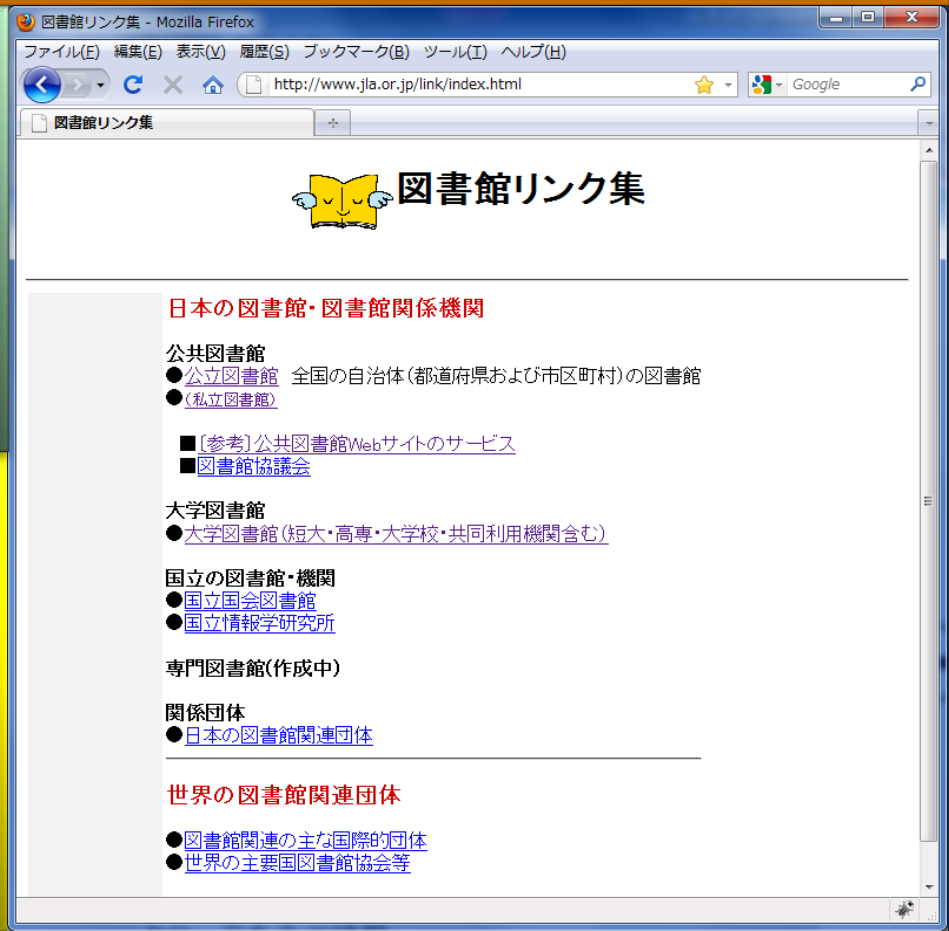
日本図書館協会図書館リンク集
http://www.jla.or.jp/link/

公共図書館
大学図書館

合計2,450館

2010年9月5日にアクセス可能だった2,065館

公共図書館	1,277館
大学図書館	788館



調査の観点

robots.txtの有無

・ない場合、クローラのアクセス排除をしていない

⇒ クローラOK

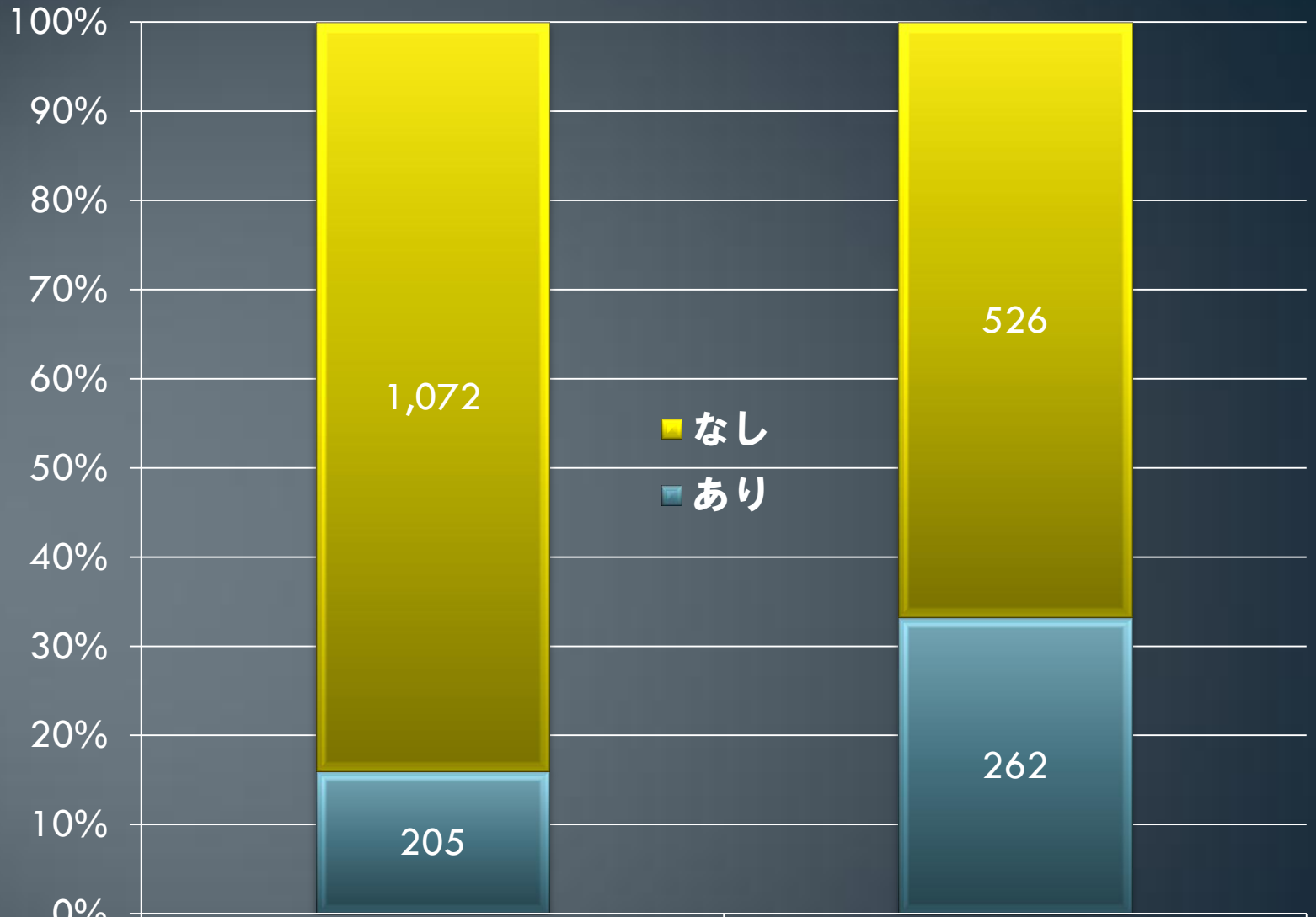
robots.txtの内容

・誤りがあるか

アクセス制限の対象

・クローラのアクセスを全コンテンツで排除しているか

Robots.txtの有無



	公共図書館	大学図書館
なし	1,072	526
あり	205	262

誤りがあるか

Robots.txtの内容

ウェブページを返す

記述の誤り

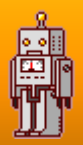
ウェブページを返す

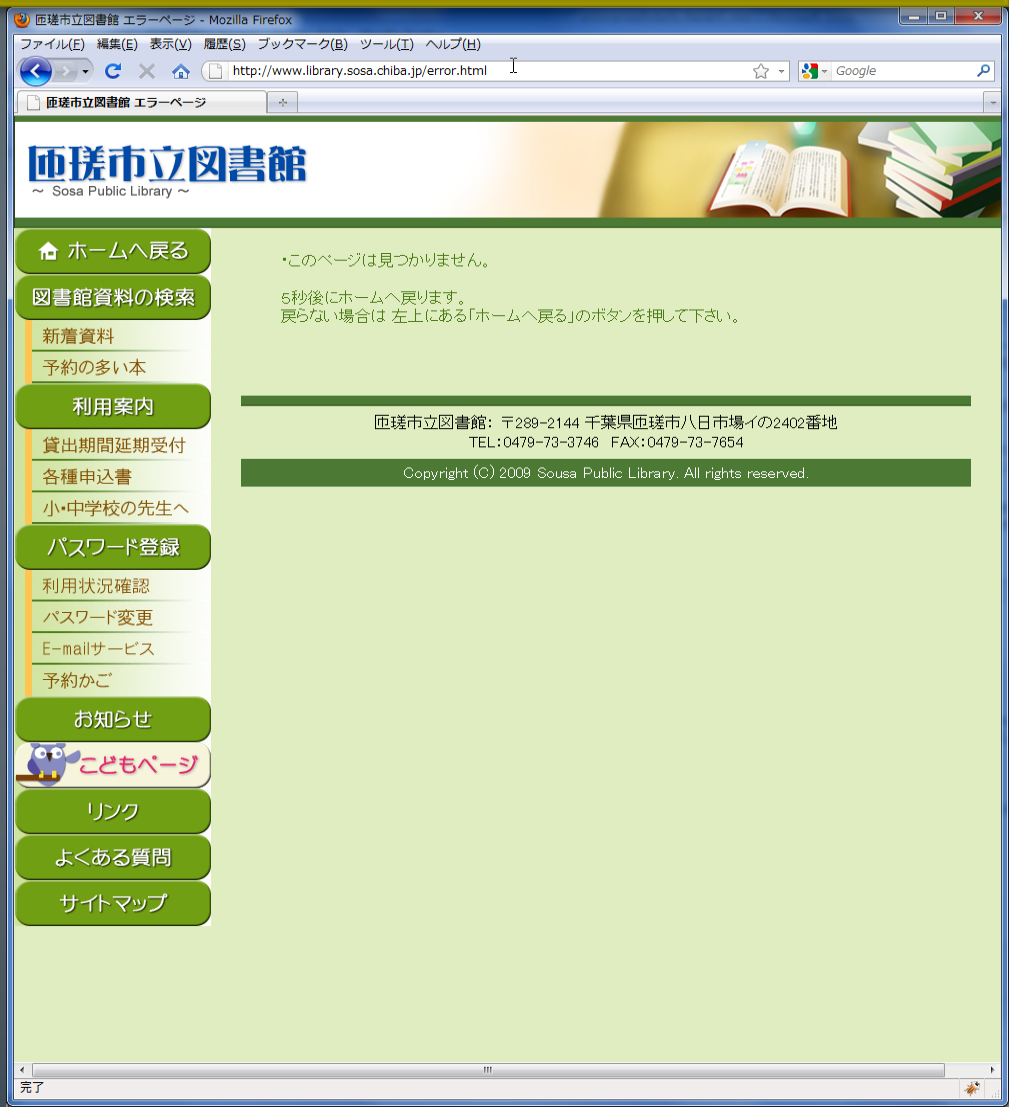
Robots.txtの内容

robots.txtをリクエストするとウェブページを返す

ないのと同じ

⇒クローラOK





Robots.txtの内容

記述の誤り

岡崎市立図書館を始めとする17館

誤り（記述に矛盾）で全アクセス排除

同一の業者によるサイト

誤りが修正されている館が5館

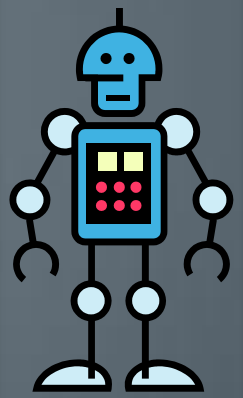
アクセス制限の対象

	公共図書館		大学図書館		全体	
	館数	割合	館数	割合	館数	割合
全公開・大半公開	1,216	95.2%	778	98.7%	1,994	96.6%
全排除	61	4.8%	10	1.3%	71	3.4%
合計	1,277	100.0%	788	100.0%	2,065	100.0%

図書館はクローラのアクセス制御をしているか

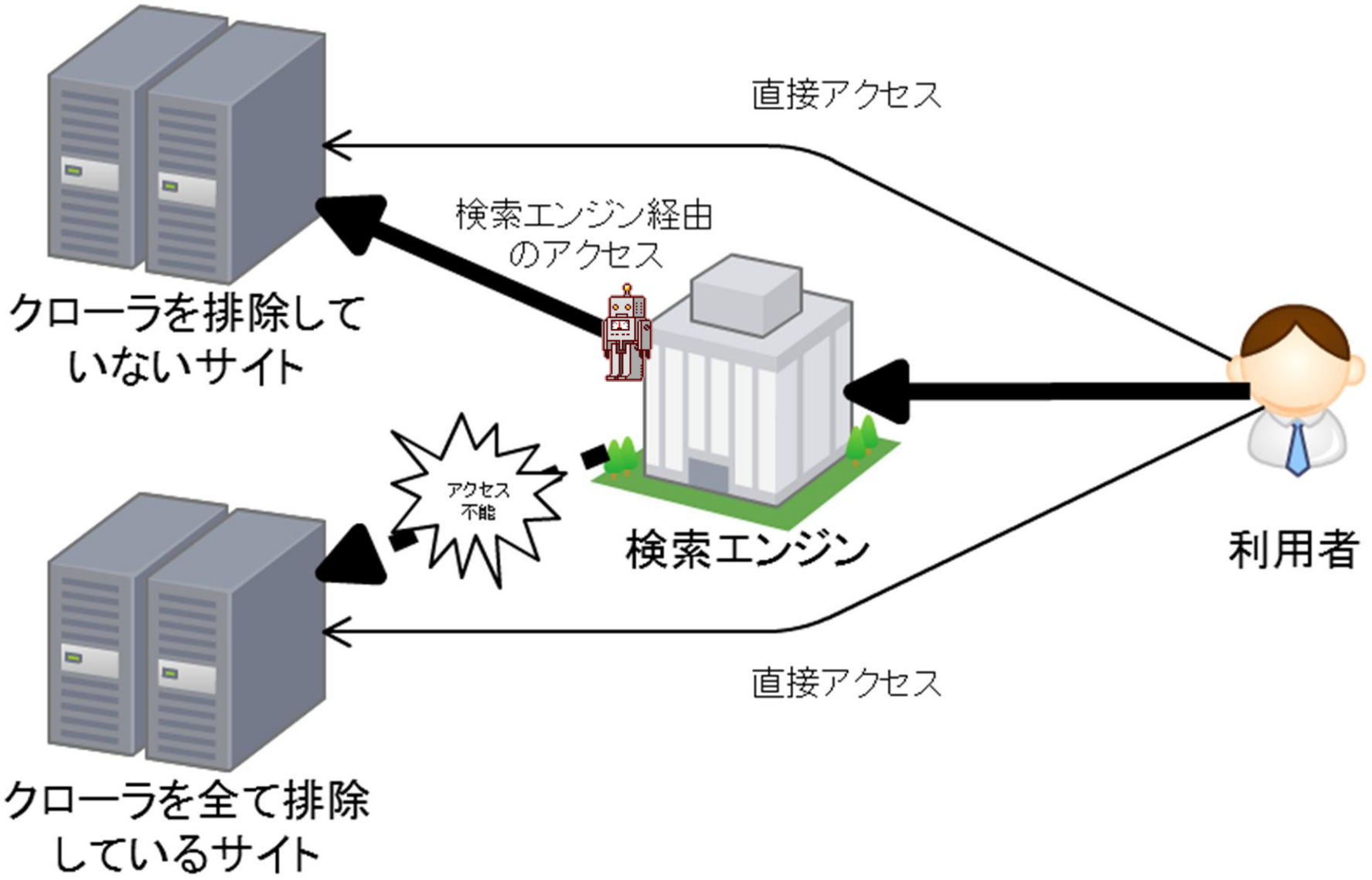
クローラのアクセスを

ほとんどの図書館は全て・大半の部分で認めている



一部の館は全排除

クローラのアクセス排除の影響


直接アクセス
検索エンジン経由
のアクセス
クローラを排除して
いないサイト
アクセス
不能
検索エンジン
利用者
直接アクセス
クローラを全て排除
しているサイト

調査方法

検索対象

- YAHOO! JAPAN
- Google 日本

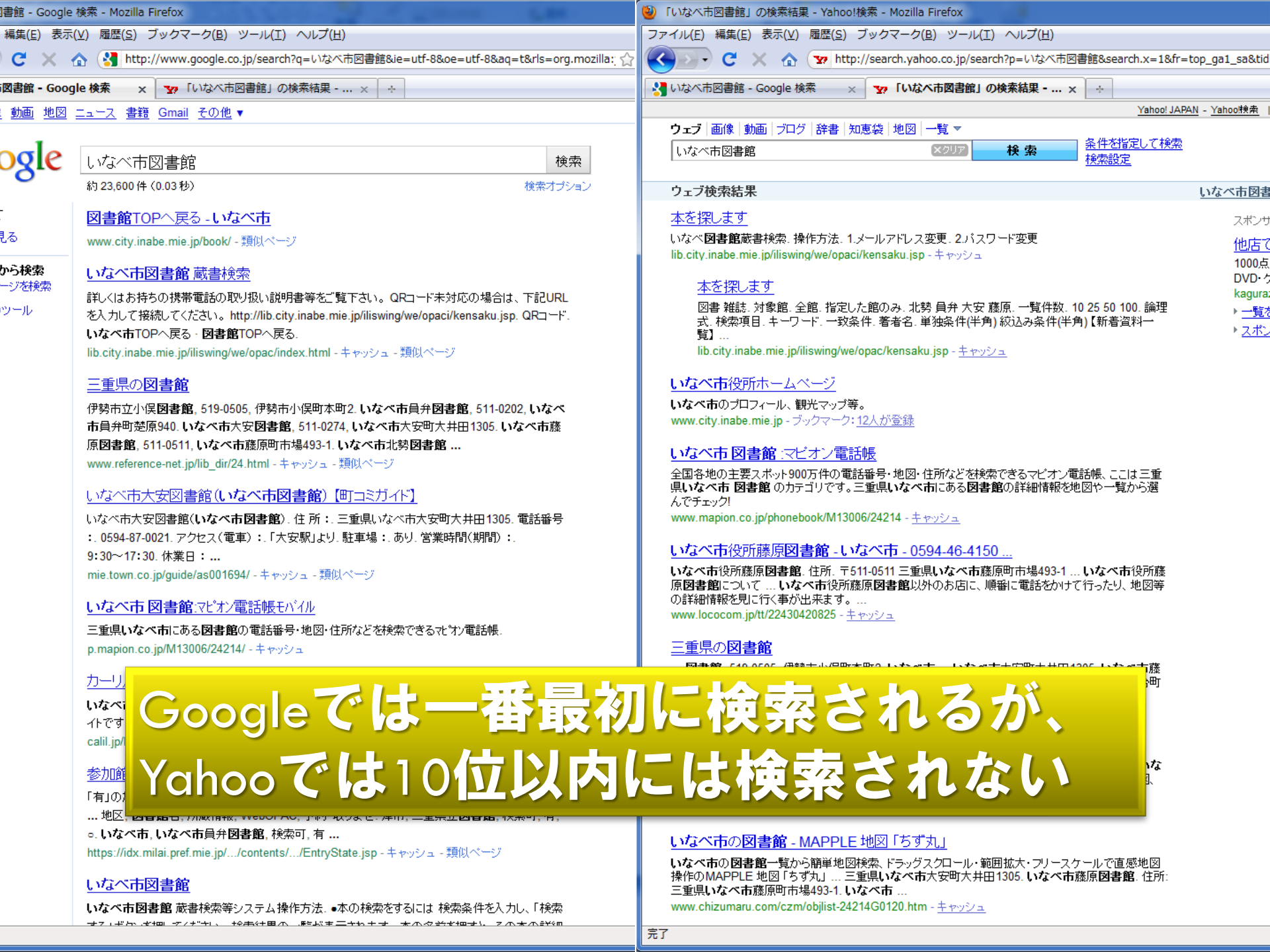
検索式

- 図書館の名称

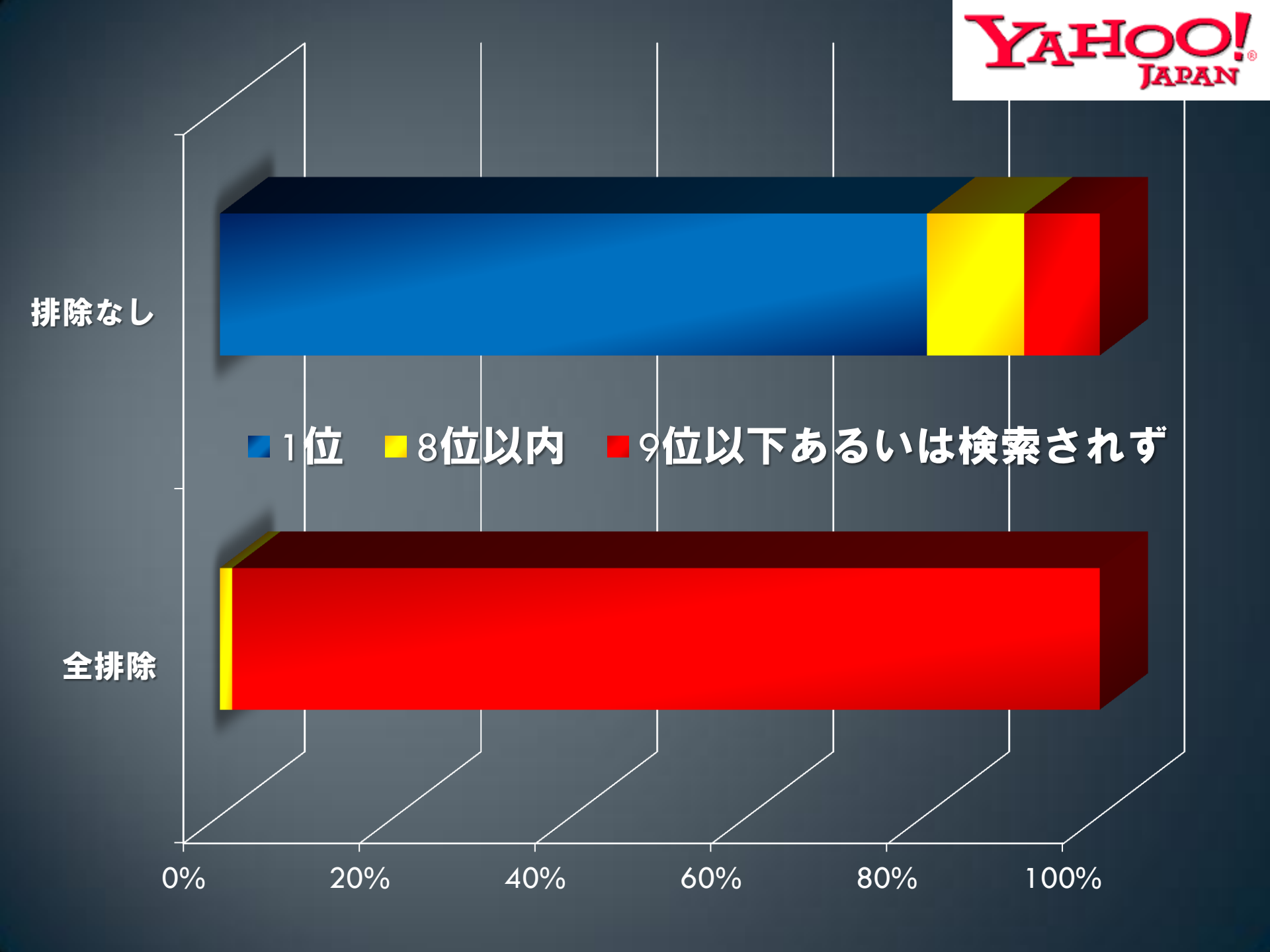
検索結果

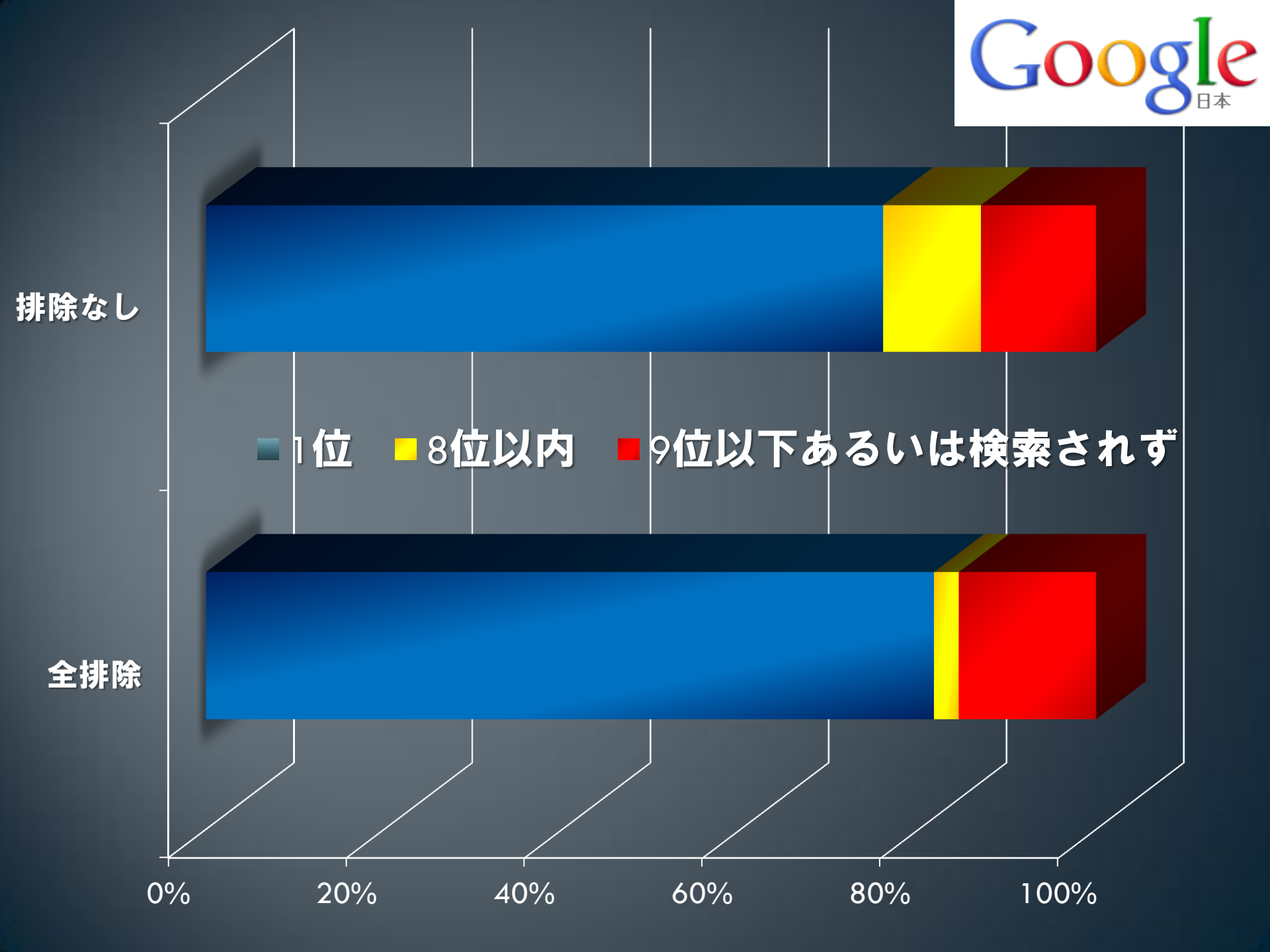
- タイトル、URL、要約
- 上位8位まで


クローラのアクセスを全て排除している図書館をGoogleとYahooで検索すると・・・

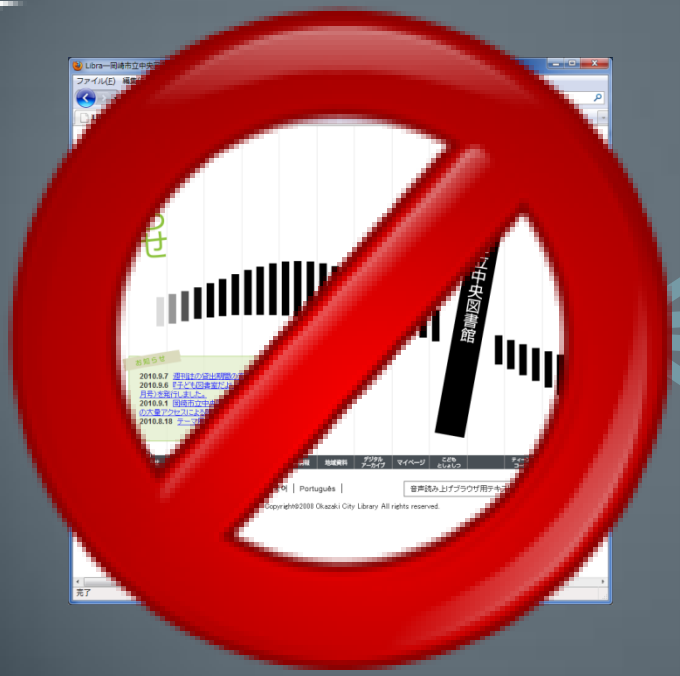

書館 - Google 検索 - Mozilla Firefox
編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.google.co.jp/search?q=いなべ市図書館&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:
図書館 - Google 検索
「いなべ市図書館」の検索結果 - ...
動画 地図 ニュース 書籍 Gmail その他
いなべ市図書館
検索
約 23,600 件 (0.03 秒)
検索オプション
図書館TOPへ戻る - いなべ市
www.city.inabe.mie.jp/book/ - 類似ページ
いなべ市図書館 蔵書検索
詳しくはお持ちの携帯電話の取り扱い説明書等をご覧下さい。QRコード未対応の場合は、下記URLを入力して接続してください。http://lib.city.inabe.mie.jp/iliswing/we/opaci/kensaku.jsp. QRコード. いなべ市TOPへ戻る - 図書館TOPへ戻る.
lib.city.inabe.mie.jp/iliswing/we/opac/index.html - キャッシュ - 類似ページ
三重県の図書館
伊勢市立小俣図書館, 519-0505, 伊勢市小俣町本町2. いなべ市員弁図書館, 511-0202, いなべ市員弁町楚原940. いなべ市大安図書館, 511-0274, いなべ市大安町大井田1305. いなべ市藤原図書館, 511-0511, いなべ市藤原町市場493-1. いなべ市北勢図書館 ...
www.reference-net.jp/lib_dir/24.html - キャッシュ - 類似ページ
いなべ市大安図書館（いなべ市図書館）【町コミガイド】
いなべ市大安図書館(いなべ市図書館). 住 所：. 三重県いなべ市大安町大井田1305. 電話番号：. 0594-87-0021. アクセス(電車)：.「大安駅」より. 駐車場：. あり. 営業時間(期間)：. 9:30～17:30. 休業日：...
mie.town.co.jp/guide/as001694/ - キャッシュ - 類似ページ
いなべ市 図書館:マピオン電話帳モバイル
三重県いなべ市にある図書館の電話番号・地図・住所などを検索できるマピオン電話帳.
p.mapion.co.jp/M13006/24214/ - キャッシュ
参加館
... いなべ市, いなべ市員弁図書館, 検索可, 有 ...
https://idx.milai.pref.mie.jp/.../contents/.../EntryState.jsp - キャッシュ - 類似ページ
いなべ市図書館
いなべ市図書館 蔵書検索等システム操作方法. ●本の検索をするには 検索条件を入力し、「検索
「いなべ市図書館」の検索結果 - Yahoo!検索 - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://search.yahoo.co.jp/search?p=いなべ市図書館&search.x=1&fr=top_ga1_sa&tid
いなべ市図書館 - Google 検索
「いなべ市図書館」の検索結果 - ...
Yahoo! JAPAN - Yahoo!検索
ウェブ 画像 動画 ブログ 辞書 知恵袋 地図 一覧
いなべ市図書館
クリア
検索
条件を指定して検索
検索設定
ウェブ検索結果
いなべ市図書
本を探します
いなべ図書館蔵書検索. 操作方法. 1.メールアドレス変更. 2.パスワード変更
lib.city.inabe.mie.jp/iliswing/we/opaci/kensaku.jsp - キャッシュ
本を探します
図書 雑誌. 対象館. 全館. 指定した館のみ. 北勢 員弁 大安 藤原. 一覧件数. 10 25 50 100. 論理式. 検索項目. キーワード. 一致条件. 著者名. 単独条件(半角) 絞込み条件(半角)【新着資料一覧】...
lib.city.inabe.mie.jp/iliswing/we/opac/kensaku.jsp - キャッシュ
いなべ市役所ホームページ
いなべ市のプロフィール、観光マップ等。
www.city.inabe.mie.jp - ブックマーク:12人が登録
いなべ市 図書館 :マピオン電話帳
全国各地の主要スポット900万件の電話番号・地図・住所などを検索できるマピオン電話帳、ここは三重県いなべ市 図書館 のカテゴリです。三重県いなべ市にある図書館の詳細情報を地図や一覧から選んでチェック!
www.mapion.co.jp/phonebook/M13006/24214 - キャッシュ
いなべ市役所藤原図書館 - いなべ市 - 0594-46-4150 ...
いなべ市役所藤原図書館. 住所. 〒511-0511 三重県いなべ市藤原町市場493-1 ... いなべ市役所藤原図書館について ... いなべ市役所藤原図書館以外のお店に、順番に電話をかけて行ったり、地図等の詳細情報を見に行く事が出来ます。...
www.lococom.jp/tt/22430420825 - キャッシュ
三重県の図書館
いなべ市の図書館 - MAPPLE 地図「ちず丸」
いなべ市の図書館一覧から簡単地図検索、ドラッグスクロール・範囲拡大・フリースケールで直感地図操作のMAPPLE 地図「ちず丸」... 三重県いなべ市大安町大井田1305. いなべ市藤原図書館. 住所: 三重県いなべ市藤原町市場493-1. いなべ市 ...
www.chizumaru.com/czm/objlist-24214G0120.htm - キャッシュ
完了
Googleでは一番最初に検索されるが、
Yahooでは10位以内には検索されない

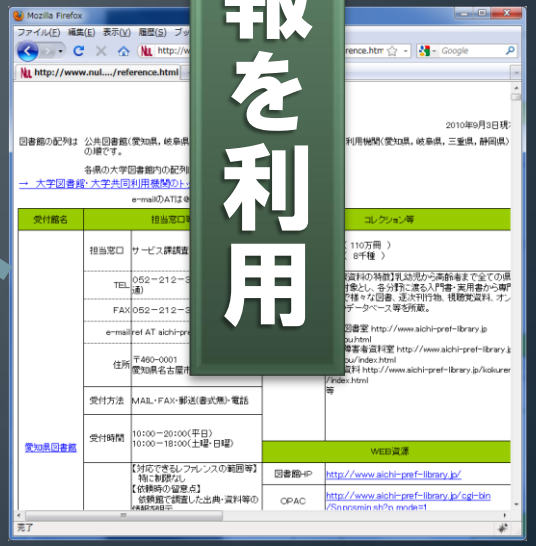
		クローラによるアクセス			
		排除なし		全排除	
		結果数	割合	結果数	割合
検索結果順位	1	1601	80.3%	0	0.0%
	2	167	8.4%	1	1.4%
	3	34	1.7%	0	0.0%
	4	11	0.6%	0	0.0%
	5	2	0.1%	0	0.0%
	6	5	0.3%	0	0.0%
	7	2	0.1%	0	0.0%
	8	0	0.0%	0	0.0%
9位以下		172	8.6%	70	98.6%
合計		1994	100.0%	71	100.0%


YAHOO! JAPAN
排除なし
全排除
1位
8位以内
9位以下あるいは検索されず
0%
20%
40%
60%
80%
100%


Google 日本
排除なし
全排除
1位
8位以内
9位以下あるいは検索されず
0%
20%
40%
60%
80%
100%

全排除のページをなぜGoogleは検索できる？





リンク情報を利用


港区立図書館 - Google 検索 - Mozilla Firefox
港区立図書館
約 261,000 件 (0.10 秒)
港区立図書館サイト・トップページ
港区立図書館 - 国民読書年 - お問い合わせ - English. 最終更新日：2010年9月7日. みなと図書館 - 三田図書館 - 赤坂図書館 - 高輪図書館 - 港南図書館 - 麻布図書サービスセンター - 郷土資料館. みなと図書館 ...
www.lib.city.minato.tokyo.jp/ - キャッシュ - 類似ページ
資料を探す
赤坂図書館
みなと図書館
高輪図書館
三田図書館
港南図書館
所在地・休館日・開館時間
図書館の利用案内
minato.tokyo.jp からの検索結果 »
港区立図書館 | みなと図書館
港区立図書館 - 国民読書年 - お問い合わせ - English - 図書館TOP - みなと図書館 - 三田図書館 - 赤坂図書館 - 高輪図書館 - 港南図書館 - 麻布図書サービスセンター - 郷土資料館 - 図書館TOP > みなと図書館 ...
www.lib.city.minato.tokyo.jp/j/minato.cgi - キャッシュ - 類似ページ
【東京図書館制覇！】港区立図書館
港区立図書館の一覧、地図、データをまとめています。... 東京タワーの北東にある、港区立図書館の中央館。地域資料が充実しています。地域資料室への通路にある展示コーナーでは、図書館が所蔵する貴重な資料を使った展示が行われています。...
tokyo-toshokan.net/00000024.htm - キャッシュ
全国各地の主要スポット900万件の電話番号・地図・住所などを検索できるマピオン電話帳、ここは東京都港区 図書館 のカテゴリです。東京都港区にある図書館の詳細情報を地図や一覧から選んでチェック！
www.mapion.co.jp › ... › 公共施設 › 図書館 › 東京都 - キャッシュ - 類似ページ
港区立図書館でも月曜開館！：日本創新党 荒川区議会議員小坂英二の ...
タイトル
要約
いなべ市図書館 - Google 検索 - Mozilla Firefox
いなべ市図書館
検索
23,600 件 (0.03 秒)
検索オプション
図書館TOPへ戻る - いなべ市
www.city.inabe.mie.jp/book/ - 類似ページ
いなべ市図書館 蔵書検索
詳しくはお持ちの携帯電話の取り扱い説明書等をご覧下さい。QRコード未対応の場合は、下記URLを入力して接続してください。http://lib.city.inabe.mie.jp/iliswing/we/opaci/kensaku.jsp. QRコード.
いなべ市TOPへ戻る - 図書館TOPへ戻る.
lib.city.inabe.mie.jp/iliswing/we/opac/index.html - キャッシュ - 類似ページ
三重県の図書館
伊勢市立小俣図書館, 519-0505, 伊勢市小俣町本町2. いなべ市員弁図書館, 511-0202, いなべ市員弁町楚原940. いなべ市大安図書館, 511-0274, いなべ市大安町大井田1305. いなべ市藤原図書館, 511-0511, いなべ市藤原町市場493-1. いなべ市北勢図書館 ...
www.reference-net.jp/lib_dir/24.html - キャッシュ - 類似ページ
いなべ市大安図書館（いなべ市図書館）【町コミガイド】
いなべ市大安図書館（いなべ市図書館）. 住 所：. 三重県いなべ市大安町大井田1305. 電話番号：. 0594-87-0021. アクセス（電車）：.「大安駅」より. 駐車場：. あり. 営業時間（期間）：. 9:30～17:30. 休業日：...
mie.town.co.jp/guide/as001694/ - キャッシュ - 類似ページ
いなべ市 図書館:マピオン電話帳モバイル
三重県いなべ市にある図書館の電話番号・地図・住所などを検索できるマピオン電話帳.
p.mapion.co.jp/M13006/24214/ - キャッシュ
いなべ市図書館
いなべ市図書館 蔵書検索等システム操作方法. ●本の検索をするには 検索条件を入力し、「検索
排除していない港区立図書館の結果と
全てを排除している図書館の結果

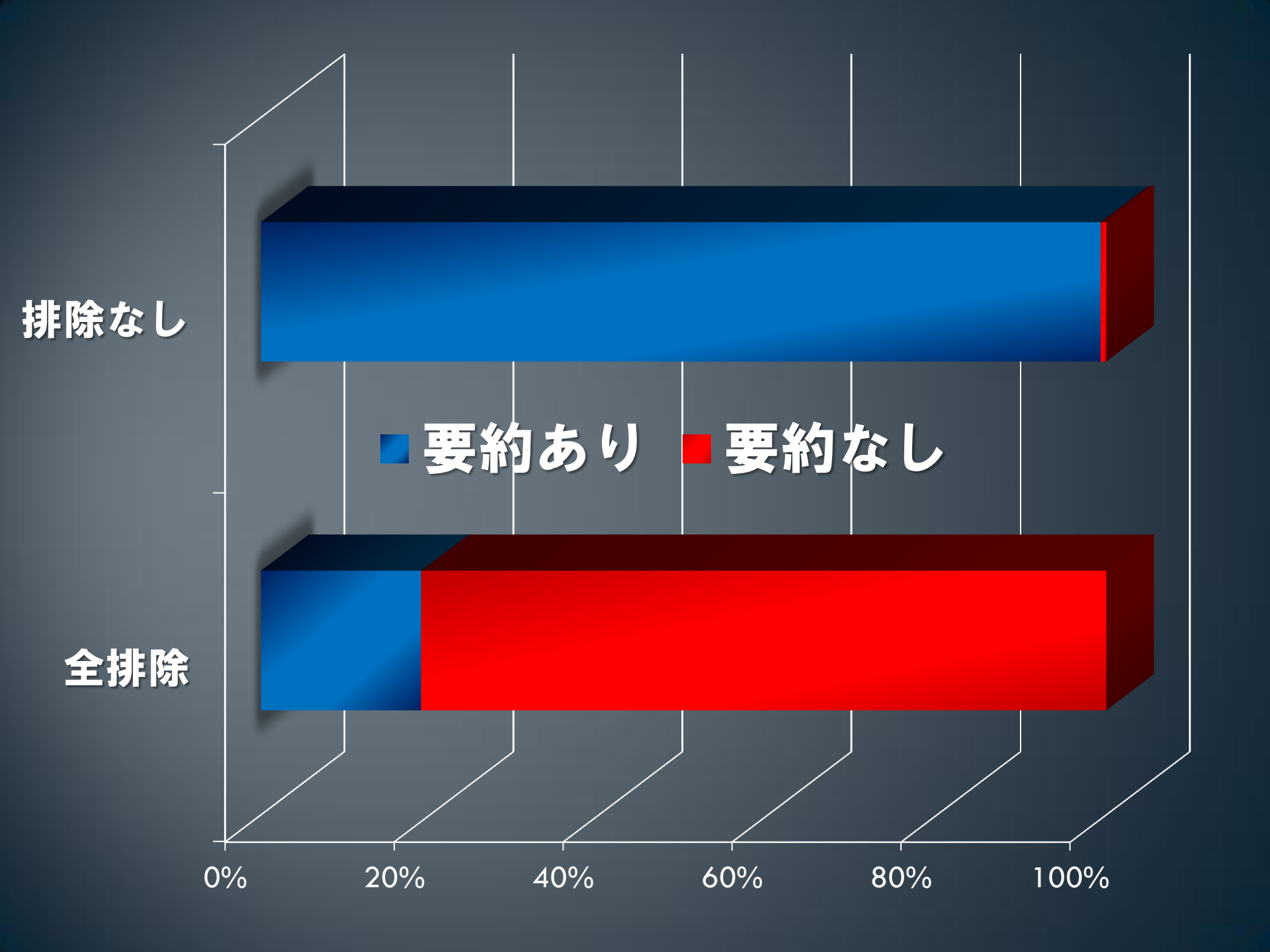
図書館サイトなのに危険なサイトに似ている？

正式名称でないタイトル

要約がない

Google検索結果中の要約

	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	1,505	99.3%	11	19.0%
要約なし	10	0.7%	47	81.0%
計	1,515	100.0%	58	100.0%

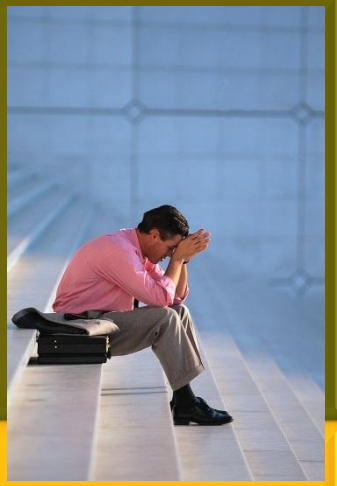

排除なし
全排除
要約あり
要約なし
0%
20%
40%
60%
80%
100%

クローラを排除したときの影響はあるか

検索結果に表示されなくなる

表示されても危険なサイト風になる

まとめ



一部を除き多くの館はクローラを排除していない

全コンテンツで排除している館は検索しにくい状態に

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