

2013/03/07

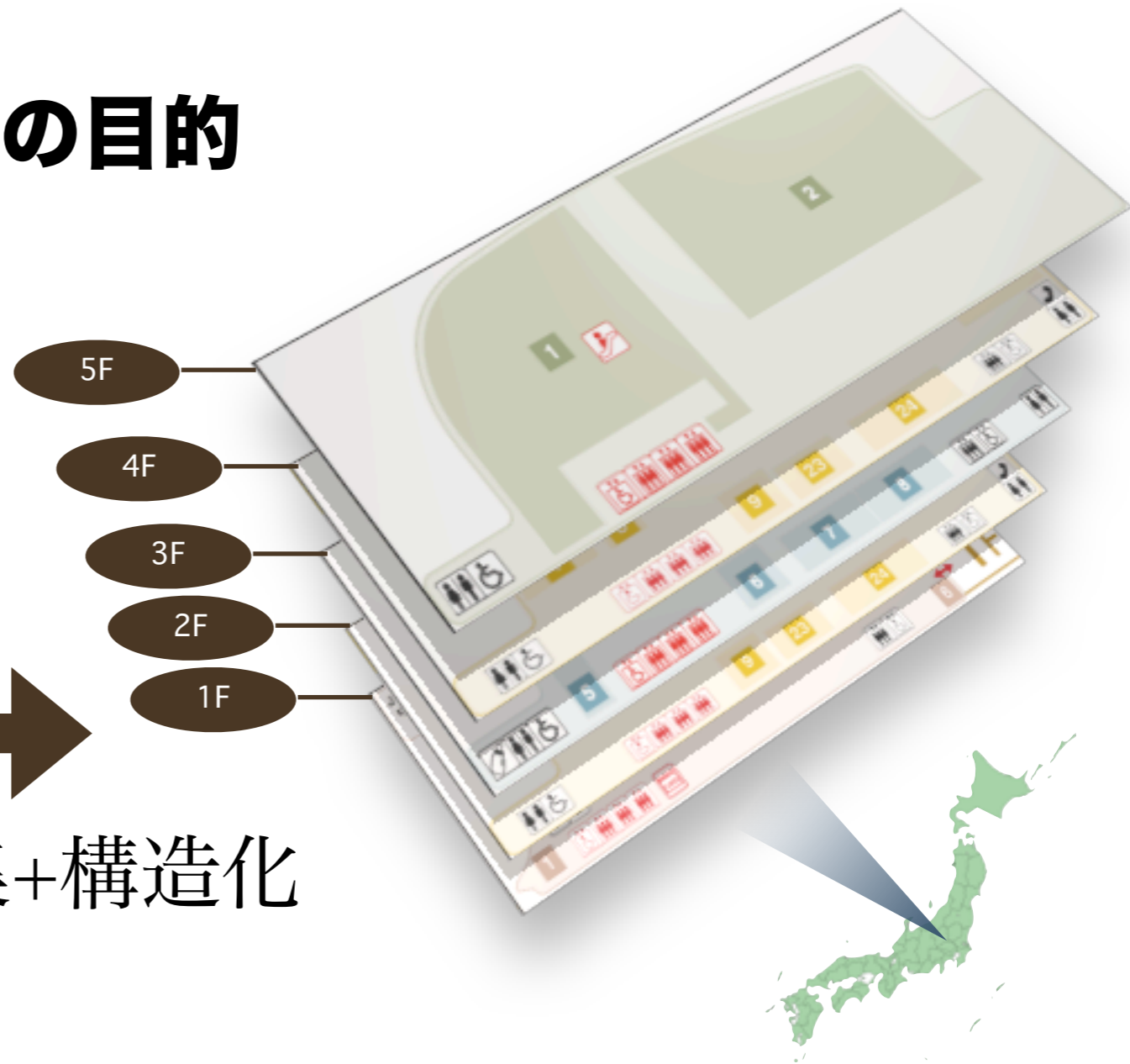
# マイクロジオデータ研究会

屋内地図情報収集のためのWebクローリング

# フロアマップの収集システムの目的



収集+構造化



- 屋外の地図のように、航空写真やGPSログからの生成はできない
- 「フロアマップ」は公開されている
- ただし、Webページの「図」として貼り付けてあるだけ。空間的に整理されているわけではない。
- 収集して構造化できないか？

# トップページ



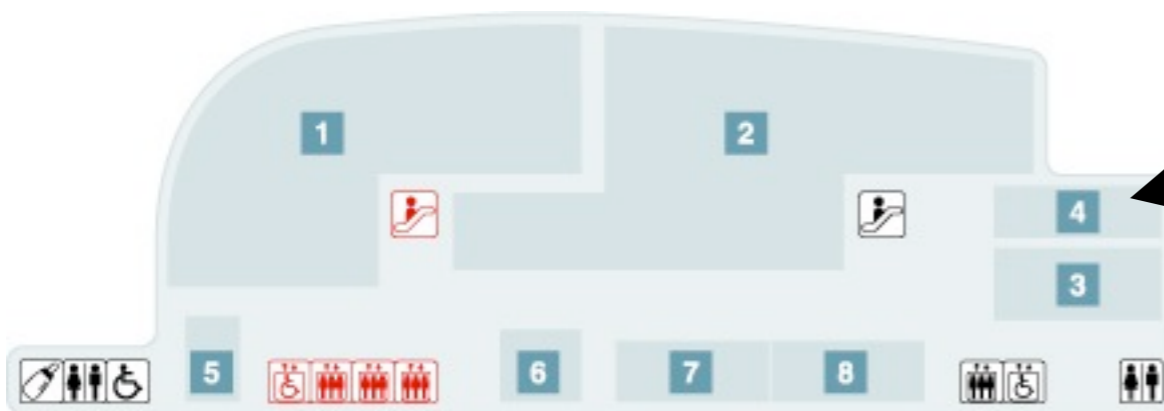
個別フロアのページ



フロアのインデックスページ



フロアマップ画像



# トップページ



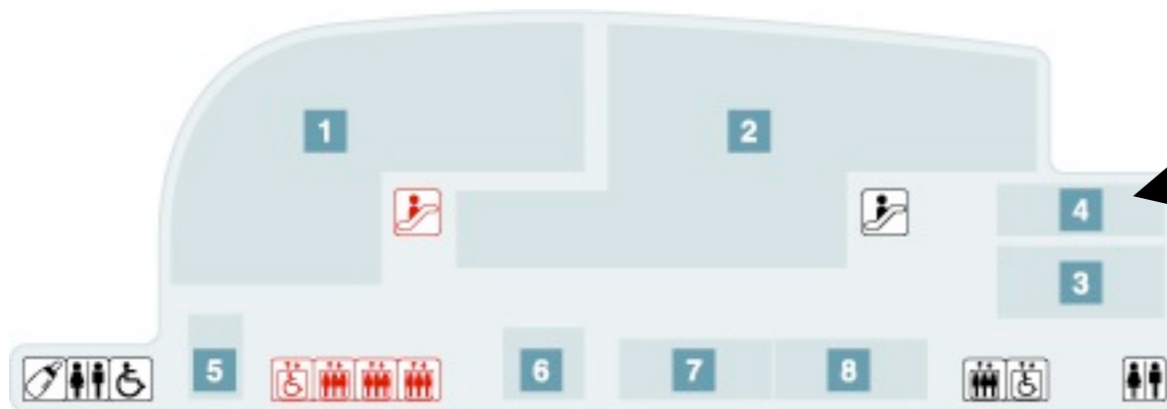
個別フロアのページ



フロアのインデックスページ

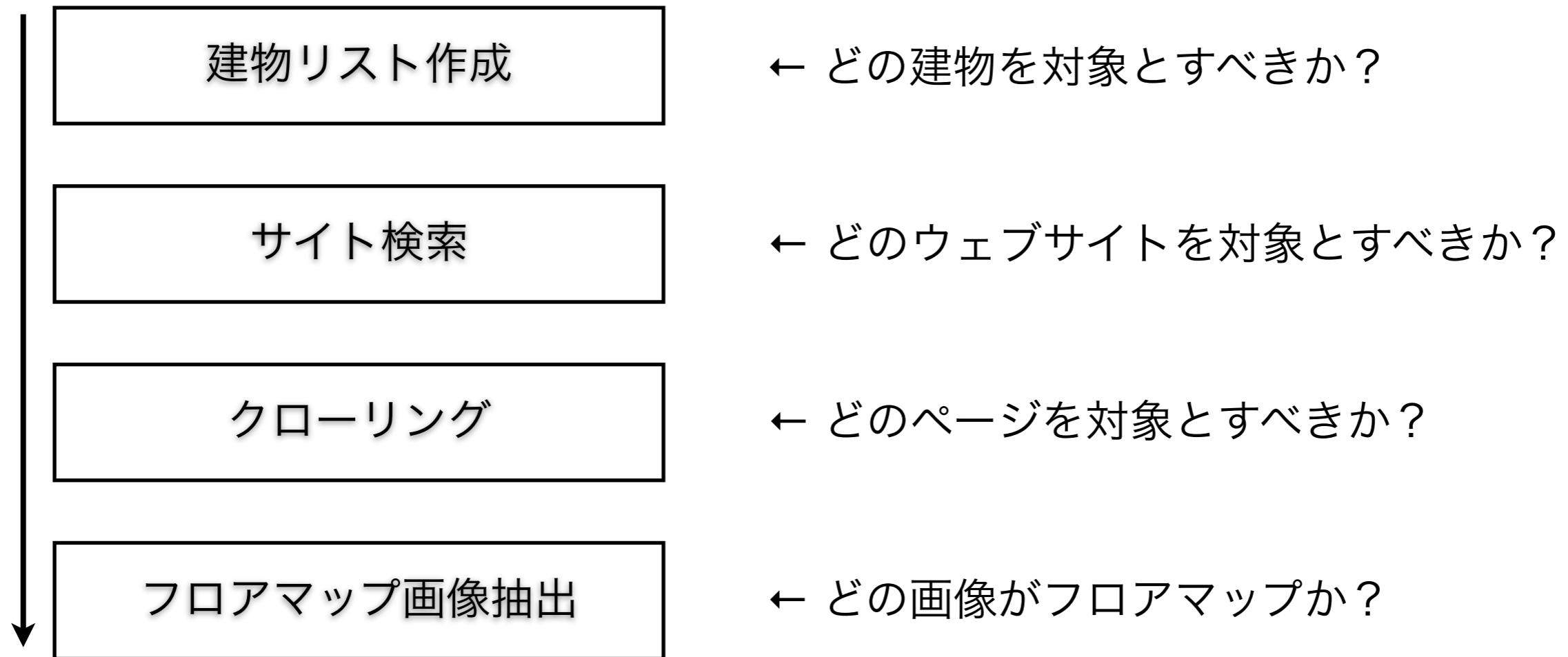


フロアマップ画像



# フロアマップ収集の流れ

---



# 建物リスト作成

元データ

建物リスト

+

別記（入居テナント一覧）

← ゼンリンの住宅地図データから  
対象となる建物をピックアップする

入居者（商業カテゴリ）ひとつにつき1点として数える

店舗らしいパターンの名前はもう1点追加する(計2点)

事務所らしい名前は加点しない



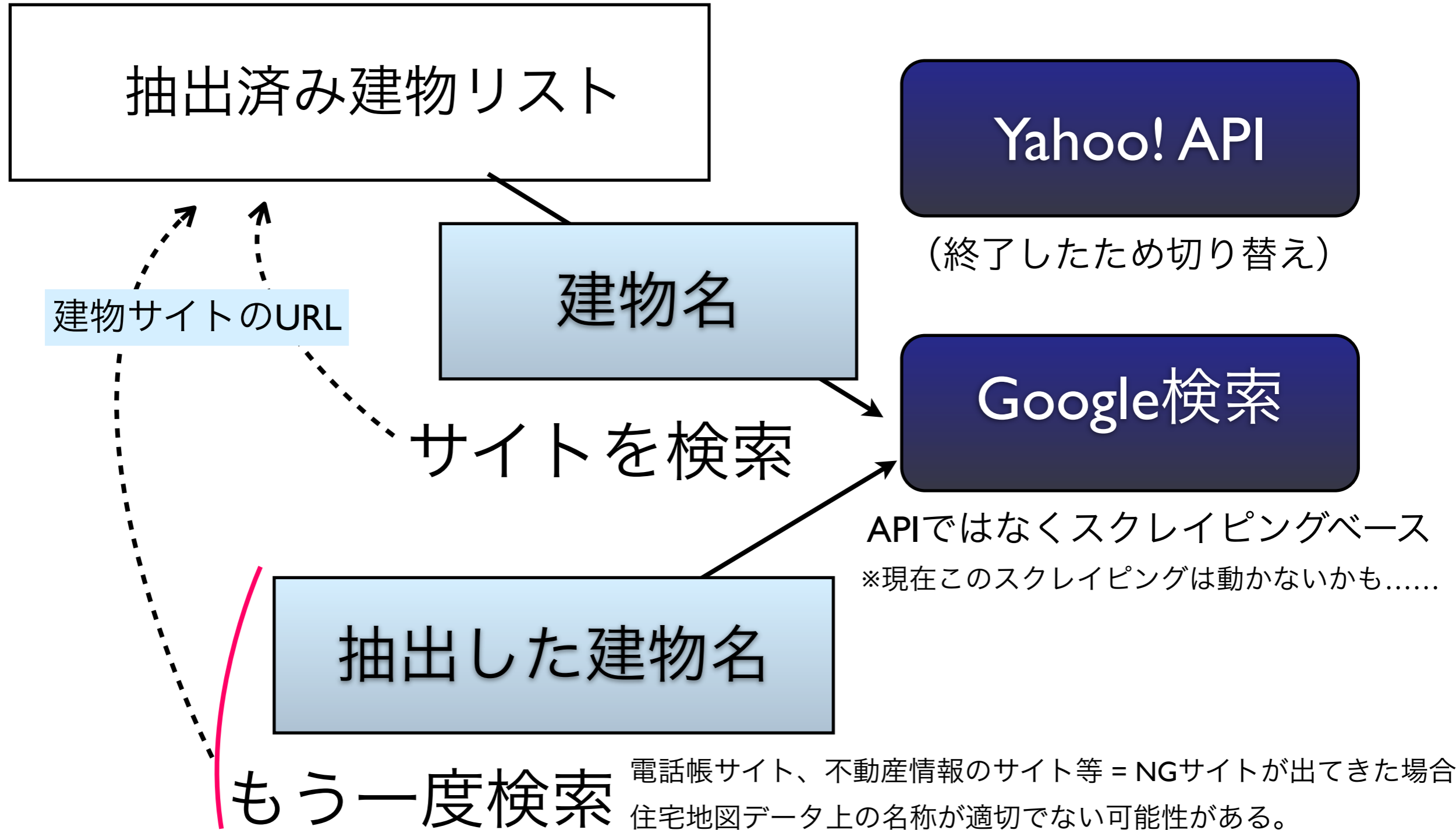
抽出済み建物リスト

# 建物リスト例

#	aID	処理無し	事務所名フィルタ	店舗名加算	両方	抽出名
1	0042410900EA024500000039	新宿センタービル 227	ルミネエスト新宿店 222	新宿センタービル 275	■ ルミネエスト新宿店 251	ネエスト エスト
2	0042410900EB0245000000750	ルミネエスト新宿店 222	マルイメン新宿玄海第二ビル 170	ルミネエスト新宿店 251	■ 新宿NSビル 172	新宿NSビル 宿NSビル
3	0042410900EA0244000000557	新宿パークタワー 217	新宿NSビル 144	新宿パークタワー 240	■ マルイメン新宿玄海第二ビル 172	
4	0042410900EB02490000000FB	高田馬場ダイカンプラザ 215	新宿ミロード 125	高田馬場ダイカンプラザ 217	■ 新宿ミロード 140	ミロード 新宿ミロード
5	0042410900EA0246000000974	新宿アイランドタワー 175	新宿パークタワー 117	新宿アイランドタワー 187	■ 新宿パークタワー 138	パークタワー ークタワー
6	0042410900EC0246000000F77	マルイメン新宿玄海第二ビル 170	ルミネ1 108	新宿NSビル 185	■ ルミネ1 136	新宿 宿
7	0042410900E90243000000A72	東京オペラシティビル東京オペラシティタワー 158	新宿三越アルコット 107	東京オペラシティビル東京オペラシティタワー 179	■ 新宿三越アルコット 120	新宿三越アルコット 宿三越アルコット
8	0042410900EA024500000002D	新宿NSビル 154	新宿エルタワー 94	マルイメン新宿玄海第二ビル 172	■ 新宿エルタワー 120	支 エルタワー
9	0042410900EA0245000000013	新宿住友ビルディング 152	高田馬場ダイカンプラザ 92	新宿住友ビルディング 170	■ 小田急百貨店本館 113	新宿 宿
10	0042410900EB02460000000A2	新宿ダイカンプラザA館 151	ルミネ2 92	新宿ダイカンプラザA館 158	■ 新宿センタービル 111	ル 新宿センタービル
11	0042410900EA0246000000410	新宿野村ビル 137	新宿センタービル 83	新宿野村ビル 156	■ ルミネ2 110	新宿 宿
12	0042410900EA0245000000016	新宿三井ビル 131	新宿野村ビル 82	新宿三井ビル 148	■ 新宿野村ビル 99	野村ビル 新宿野村ビル
13	0042410900EB02450000007AA	新宿ミロード 125	新宿プリンスホテル西武新宿ベベ 81	新宿ミロード 140	■ 高田馬場ダイカンプラザ 93	
14	0042410900EB02450000003B4	ルミネ1 108	新宿ダイビル 76	ルミネ1 136	■ 新宿プリンスホテル西武新宿ベベ 92	新宿ベベ 西武新宿ベベ
15	0042410900EB0245000000736	新宿三越アルコット 107	新宿モノリス 75	新宿三越アルコット 120	■ 新宿ダイカンプラザA館 82	大ガード 新宿
16	0042410900EA024500000000F	新宿第一生命ビルディング 101	新宿ダイカンプラザA館 75	新宿エルタワー 120	■ 新宿モノリス 82	新宿モノリス 新宿モノリス
17	0042410900EB0245000000217	レイフラット新宿 95	小田急百貨店本館 72	小田急百貨店本館 113	■ 東京オペラシティビル東京オペラシティタワ ー 81	オペラシティ ペラシティ
18	0042410900EA024600000072A	新宿エルタワー 94	新宿アイランドタワー 65	ルミネ2 110	■ 新宿ダイビル 81	新宿アルタ 宿アルタ
19	0042410900EB0245000000073	ルミネ2 92	歌舞伎町ダイカンプラザ星座館 64	新宿第一生命ビルディング 109	■ 新宿住友ビルディング 79	住友ビル 新宿住友ビル
20	0042410900EA0247000000711	フコレー新宿第一ビル 86	新宿住友ビルディング 63	レイフラット新宿 96	■ 飯田橋セントラルプラザ東京都飯田橋庁舎 76	飯田橋 田橋
21	0042410900EC0245000000034	新宿Qフラットビル 86	東京オペラシティビル東京オペラシティタワー 60	新宿プリンスホテル西武新宿ベベ 92	■ 新宿アイランドタワー 73	アイランドタワー 新宿アイランド
22	0042410900EB024600000002C	新宿ダイカンプラザ756 85	Lee 3 57	エステック情報ビル 88	■ 新宿三井ビル	新宿三井ビル 宿三井ビル

“外れ” が下位に押し下げられる

# サイト検索



NGサイトリストに含まれている場合

建物名が一般的ではない可能性

→ 適切でない名称の例：道玄坂共同ビル渋谷109 (渋谷109)

→ テナント名に含まれる建物名を抽出

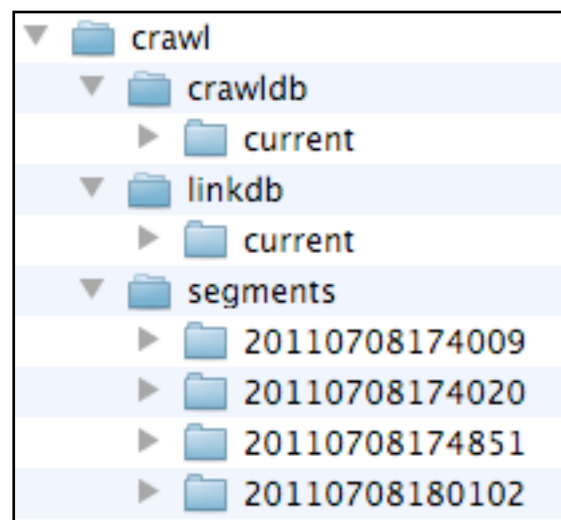


# クローリング (STEP1 : 単純クローリング)

nutch を使用してクローリングを行う。

ここではフロアマップが掲載されているか判断せず、あらゆるページをクローリングする

- ・クローリングは（目的自体は）単純だが、実装は意外と大変。Web 上にはおかしな（汚い）ページがたくさんある。それらにぶつかっても、クラッシュしないよう処理を続ける必要がある。
- ・今のところ nutch でクローリングして後からデータベースを抽出するという実装



nutch のクローリングデータベース（多数のファイル）を一つの DBファイルに変換し、自前のシステムではここからデータを取り込んで処理をする

# クローリング (STEP2: ページのスコア計算)

---

ここからは、自前のシステムで処理を行う。

## リンク追跡・評価

起点ページ（検索で見つかったページ）から順番にリンクを辿りながらリンクの評価をする。リンクに含まれるキーワードに応じて得点を付ける。

## ページ単体評価

リンク同様、ページのタイトルやURLにもキーワードが含まれているかどうか調べ、得点を付ける。

→ページ単体の得点に、流入リンクの得点を合計してページ最終得点とする

→但し、ページの深さ（起点ページからのホップ数）により流入リンクを加算するか判断

# リンク評価用スコア表

- keyword: フロア

score: 3.0

- keyword: floor

filter: downcase

score: 3.0

- keyword: マップ

score: 3.0

- keyword: map

filter: downcase

score: 3.0

- keyword: ガイド

score: 1.0

- keyword: (b|B|地下)?[1-9]+[fF f F 階]

filter: regexp

score: 1.0

← ここは正規表現によるパターンマッチ  
(○階 というパターンにマッチさせる )

# ページ評価用スコア表

- keyword: フロア

score: 3.0

- keyword: floor

filter: downcase

score: 3.0

- keyword: マップ

score: 3.0

- keyword: map

filter: downcase

score: 3.0

- keyword: ガイド

score: 1.0

- keyword: (b|B|地下)?[1-9]+[fF f F 階]

filter: regexp

score: 1.0

(現在のところ同じ)

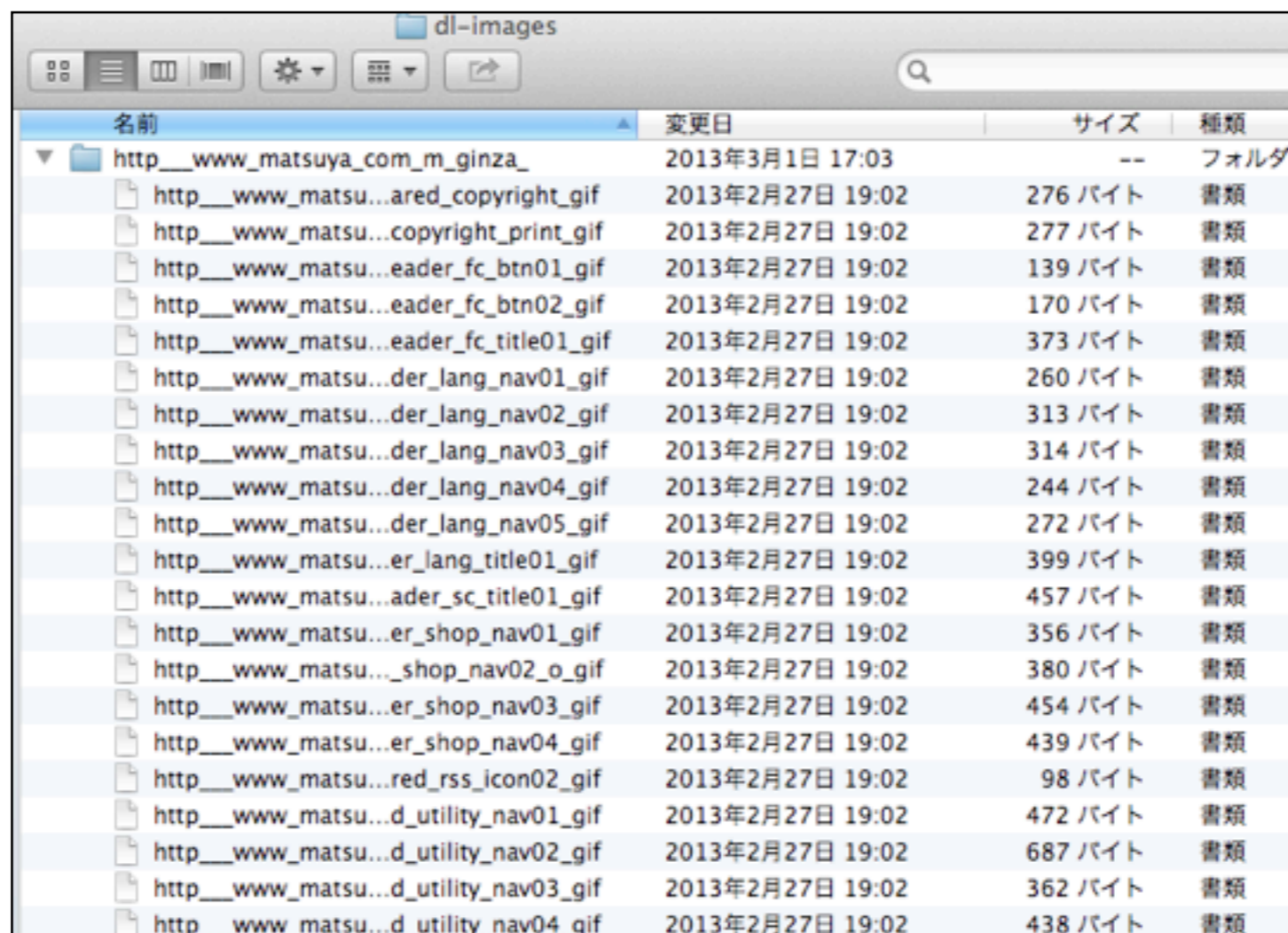
# ページスコア+リンクスコア例

D	タイトル/URL	KW	Pr	Cmb
2	ヨドバシAkiba   フロアガイド2-6F 19761B <a href="http://www.yodobashi-akiba.com/information/information_2_6.html">http://www.yodobashi-akiba.com/information/information_2_6.html</a>	5.0 U 0.0 / T 5.0	40.0	45.0
7 Incoming Link(s)				
[1] ヨドバシAkiba   フロアガイド : K 7.0 - ヨドバシAkiba 6F : 1.0				
[1] ヨドバシAkiba   フロアガイド : K 7.0 - ヨドバシAkiba 5F : 1.0				
[1] ヨドバシAkiba   フロアガイド : K 7.0 - ヨドバシAkiba 4F : 1.0				
[1] ヨドバシAkiba   フロアガイド : K 7.0 - ヨドバシAkiba 3F : 1.0				
[1] ヨドバシAkiba   フロアガイド : K 7.0 - ヨドバシカメラ2F : 1.0				
[2] ヨドバシAkiba   フロアガイド2-6F : K 5.0 - ページTOP : 0.0				
[2] ヨドバシAkiba   フロアガイドB1-1F : K 5.0 - 2-6F : 1.0				
2	ヨドバシAkiba   フロアガイドB1-1F 22864B <a href="http://www.yodobashi-akiba.com/information/information_b1_f1.html">http://www.yodobashi-akiba.com/information/information_b1_f1.html</a>	5.0 U 0.0 / T 5.0	16.0	21.0
4 Incoming Link(s)				
1	ヨドバシAkiba   フロアガイド 21820B <a href="http://www.yodobashi-akiba.com/information/information_floor.html">http://www.yodobashi-akiba.com/information/information_floor.html</a>	7.0 U 3.0 / T 4.0	7.0	14.0
67 Incoming Link(s)				
1	ヨドバシAkiba   アクセスガイド 17722B <a href="http://www.yodobashi-akiba.com/information/access_map.html">http://www.yodobashi-akiba.com/information/access_map.html</a>	4.0 U 3.0 / T 1.0	6.0	10.0
244 Incoming Link(s)				
1	ヨドバシAkiba   カフェ   ピアードババ	0.0	1.0	1.0

# 画像ダウンロード

候補のページに含まれる画像をダウンロードする

ただし、ページの得点が0であればダウンロードを行わない



名前	変更日	サイズ	種類
▼ http__www_matsuya_com_m_ginza_	2013年3月1日 17:03	--	フォルダ
http__www_matsu...ared_copyright_gif	2013年2月27日 19:02	276 バイト	書類
http__www_matsu...copyright_print_gif	2013年2月27日 19:02	277 バイト	書類
http__www_matsu...eader_fc_btn01_gif	2013年2月27日 19:02	139 バイト	書類
http__www_matsu...eader_fc_btn02_gif	2013年2月27日 19:02	170 バイト	書類
http__www_matsu...eader_fc_title01_gif	2013年2月27日 19:02	373 バイト	書類
http__www_matsu...der_lang_nav01_gif	2013年2月27日 19:02	260 バイト	書類
http__www_matsu...der_lang_nav02_gif	2013年2月27日 19:02	313 バイト	書類
http__www_matsu...der_lang_nav03_gif	2013年2月27日 19:02	314 バイト	書類
http__www_matsu...der_lang_nav04_gif	2013年2月27日 19:02	244 バイト	書類
http__www_matsu...der_lang_nav05_gif	2013年2月27日 19:02	272 バイト	書類
http__www_matsu...er_lang_title01_gif	2013年2月27日 19:02	399 バイト	書類
http__www_matsu...ader_sc_title01_gif	2013年2月27日 19:02	457 バイト	書類
http__www_matsu...er_shop_nav01_gif	2013年2月27日 19:02	356 バイト	書類
http__www_matsu...shop_nav02_o_gif	2013年2月27日 19:02	380 バイト	書類
http__www_matsu...er_shop_nav03_gif	2013年2月27日 19:02	454 バイト	書類
http__www_matsu...er_shop_nav04_gif	2013年2月27日 19:02	439 バイト	書類
http__www_matsu...red_rss_icon02_gif	2013年2月27日 19:02	98 バイト	書類
http__www_matsu...d_utility_nav01_gif	2013年2月27日 19:02	472 バイト	書類
http__www_matsu...d_utility_nav02_gif	2013年2月27日 19:02	687 バイト	書類
http__www_matsu...d_utility_nav03_gif	2013年2月27日 19:02	362 バイト	書類
http__www_matsu...d_utility_nav04_gif	2013年2月27日 19:02	438 バイト	書類

ダウンロードしたディレクトリにJSONファイルを作り、スコア等を記録する

# 画像の抽出

---

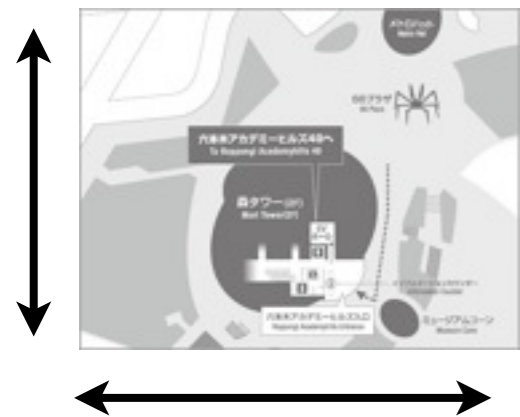
いくつかのアルゴリズムを組み合わせて画像の得点を計算する

画像寸法スコア	$S_g = \sqrt{w * h}$
関連テキスト キーワード	$S_k$
参照数	$S_r = 1/n$
SIFT特徴	$S_s$

$$\text{Score} = \{g\_normalize(S_g) + S_k + S_s\} S_r$$

\*  $g\_normalize(x) = x / S_{gmax}$  ページ内で一番大きい画像のスコアで割る

## 画像評価-1 画像寸法



画像の幅、高さ

- 画像の重要度に加味する
- 極端に小さい画像を除外する（ボタンなど）

## 画像評価-2 関連テキスト キーワード

- 代替テキスト、タイトルテキスト、URLに含まれるファイル名の3つを対象にキーワードをマッチさせる
- キーワードはページ評価用のものと同じ

## 画像評価-3 参照数

いろいろなページで参照されている画像は重要でない可能性が高い

## 画像評価-4 SIFT特徴

SIFT特徴により画像の分類を行う。分類結果のクラスごとに得点を割り当てる。



# 例

1st



寸法 キーワード SIFT 参照数

$$(1.0 + 10.0 + 2.0) / 1.0 = 13.0$$

2nd



$$(0.76 + 4.0 + 2.0) / 1.0 = 6.76$$

3rd



$$(0.43 + 8.0 + 0.0) / 3.0 = 2.81$$

# フロア名抽出

---

何階のフロアマップであるか決定

- 画像の関連テキスト
- 画像のURL
- 画像を含むページのタイトル
- 画像を含むページのURL

(上にあるものが優先)

ページのスコア修正（次スライド）にも利用する

# ページのスコア修正

---

画像の評価結果によりページのスコアを修正する

$$\text{page\_score}' = \text{page\_score} * f$$

$$f = 1 \quad \text{その他}$$



$$f = 0.5 \quad \text{画像のクラスが'other' またはフロア名が抽出されなかった場合}$$

→ 例えば、ショッパ一覧とフロアマップを掲載しているページが別々にある場合、フロアマップを掲載しているページが上位に来るようにする。ショッパ一覧の場合は店内の写真などが掲載されていると期待する。

# 収集例

各ページの最高スコア画像

タイトル: ヨドバシAkiba | フロアガイド2-6F  
URL: [http://www.yodobashi-akiba.com/information/information\\_2\\_6.html](http://www.yodobashi-akiba.com/information/information_2_6.html)  
スコア: 45.0, 45.0

1	2	3	4																																																								
 <table border="1"> <tr><td>テキスト</td><td>2F-6F フロアマップ </td></tr> <tr><td>KWスコア</td><td>7.0</td></tr> <tr><td>URL KWスコア</td><td>3.0</td></tr> <tr><td>vfeatクラス</td><td>figure</td></tr> <tr><td>最終スコア</td><td>12.0</td></tr> <tr><td>抽出フロア</td><td>:2</td></tr> <tr><td>スコア詳細</td><td>Kw:10.0 Vt:1.0 G:364.4516428828384 nG:1.0 Gmax:364.4516428828384 nRef:1</td></tr> </table>	テキスト	2F-6F フロアマップ	KWスコア	7.0	URL KWスコア	3.0	vfeatクラス	figure	最終スコア	12.0	抽出フロア	:2	スコア詳細	Kw:10.0 Vt:1.0 G:364.4516428828384 nG:1.0 Gmax:364.4516428828384 nRef:1	 <table border="1"> <tr><td>テキスト</td><td>2F-6F ヨドバシカメラ・サービス・カフェ </td></tr> <tr><td>KWスコア</td><td>1.0</td></tr> <tr><td>URL KWスコア</td><td>3.0</td></tr> <tr><td>vfeatクラス</td><td>figure</td></tr> <tr><td>最終スコア</td><td>5.609078688871741</td></tr> <tr><td>抽出フロア</td><td>:2</td></tr> <tr><td>スコア詳細</td><td>Kw:4.0 Vt:1.0 G:221.97972880423114 nG:0.6090786888717409 Gmax:364.4516428828384 nRef:1</td></tr> </table>	テキスト	2F-6F ヨドバシカメラ・サービス・カフェ	KWスコア	1.0	URL KWスコア	3.0	vfeatクラス	figure	最終スコア	5.609078688871741	抽出フロア	:2	スコア詳細	Kw:4.0 Vt:1.0 G:221.97972880423114 nG:0.6090786888717409 Gmax:364.4516428828384 nRef:1	<p>1F~6F <b>ヨドバシカメラ</b> のフロアガイドはコチラ!!</p> <table border="1"> <tr><td>テキスト</td><td>1F~6Fヨドバシカメラのフロアガイドはコチラ!! </td></tr> <tr><td>KWスコア</td><td>5.0</td></tr> <tr><td>URL KWスコア</td><td>3.0</td></tr> <tr><td>vfeatクラス</td><td>other</td></tr> <tr><td>最終スコア</td><td>2.795624290729697</td></tr> <tr><td>抽出フロア</td><td>:1</td></tr> <tr><td>スコア詳細</td><td>Kw:8.0 Vt:0 G:140.99645385611655 nG:0.3868728721890909 Gmax:364.4516428828384 nRef:3</td></tr> </table>	テキスト	1F~6Fヨドバシカメラのフロアガイドはコチラ!!	KWスコア	5.0	URL KWスコア	3.0	vfeatクラス	other	最終スコア	2.795624290729697	抽出フロア	:1	スコア詳細	Kw:8.0 Vt:0 G:140.99645385611655 nG:0.3868728721890909 Gmax:364.4516428828384 nRef:3	<p>Floor Guide</p> <table border="1"> <tr><td>テキスト</td><td>Floor Guide </td></tr> <tr><td>KWスコア</td><td>3.0</td></tr> <tr><td>URL KWスコア</td><td>3.0</td></tr> <tr><td>vfeatクラス</td><td>figure</td></tr> <tr><td>最終スコア</td><td>2.525010331230004</td></tr> <tr><td>抽出フロア</td><td>:6</td></tr> <tr><td>スコア詳細</td><td>Kw:6.0 Vt:1.0 G:209.57099035887578 nG:0.5750309936900</td></tr> </table>	テキスト	Floor Guide	KWスコア	3.0	URL KWスコア	3.0	vfeatクラス	figure	最終スコア	2.525010331230004	抽出フロア	:6	スコア詳細	Kw:6.0 Vt:1.0 G:209.57099035887578 nG:0.5750309936900
テキスト	2F-6F フロアマップ																																																										
KWスコア	7.0																																																										
URL KWスコア	3.0																																																										
vfeatクラス	figure																																																										
最終スコア	12.0																																																										
抽出フロア	:2																																																										
スコア詳細	Kw:10.0 Vt:1.0 G:364.4516428828384 nG:1.0 Gmax:364.4516428828384 nRef:1																																																										
テキスト	2F-6F ヨドバシカメラ・サービス・カフェ																																																										
KWスコア	1.0																																																										
URL KWスコア	3.0																																																										
vfeatクラス	figure																																																										
最終スコア	5.609078688871741																																																										
抽出フロア	:2																																																										
スコア詳細	Kw:4.0 Vt:1.0 G:221.97972880423114 nG:0.6090786888717409 Gmax:364.4516428828384 nRef:1																																																										
テキスト	1F~6Fヨドバシカメラのフロアガイドはコチラ!!																																																										
KWスコア	5.0																																																										
URL KWスコア	3.0																																																										
vfeatクラス	other																																																										
最終スコア	2.795624290729697																																																										
抽出フロア	:1																																																										
スコア詳細	Kw:8.0 Vt:0 G:140.99645385611655 nG:0.3868728721890909 Gmax:364.4516428828384 nRef:3																																																										
テキスト	Floor Guide																																																										
KWスコア	3.0																																																										
URL KWスコア	3.0																																																										
vfeatクラス	figure																																																										
最終スコア	2.525010331230004																																																										
抽出フロア	:6																																																										
スコア詳細	Kw:6.0 Vt:1.0 G:209.57099035887578 nG:0.5750309936900																																																										

タイトル: ヨドバシAkiba | フロアガイドB1-1F  
URL: [http://www.yodobashi-akiba.com/information/information\\_b1\\_1.html](http://www.yodobashi-akiba.com/information/information_b1_1.html)  
スコア: 21.0, 21.0

1	2	3	4										
		<p>1F~6F <b>ヨドバシカメラ</b> のフロアガイドはコチラ!!</p> <table border="1"> <tr><td>テキスト</td><td>1F~6Fヨドバシカメラのフロアガイドはコチラ!! </td></tr> <tr><td>KWスコア</td><td>5.0</td></tr> <tr><td>URL KWスコア</td><td>3.0</td></tr> </table>	テキスト	1F~6Fヨドバシカメラのフロアガイドはコチラ!!	KWスコア	5.0	URL KWスコア	3.0	<p>Floor Guide</p> <table border="1"> <tr><td>テキスト</td><td>Floor Guide </td></tr> <tr><td>KWスコア</td><td>3.0</td></tr> </table>	テキスト	Floor Guide	KWスコア	3.0
テキスト	1F~6Fヨドバシカメラのフロアガイドはコチラ!!												
KWスコア	5.0												
URL KWスコア	3.0												
テキスト	Floor Guide												
KWスコア	3.0												

# 主な課題

---

- フロア掲載ページとそれ以外の区別（スコアの足切り）
- 建物の通称（道玄坂共同ビル→渋谷109）の抽出
- 電話帳サイト、Wikipedia 等の排除
- PDF、Flash等（Flashは最近減ってきたので幸運）
- スコア、計算式の調整