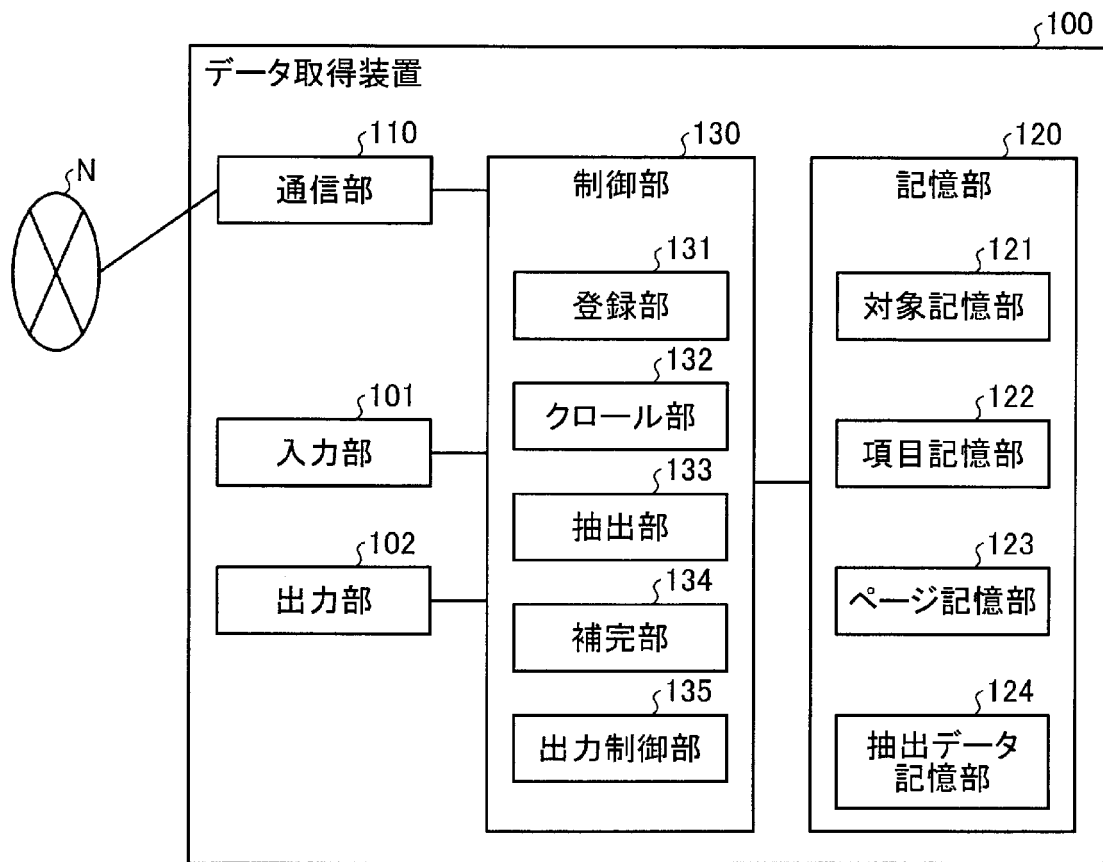


[図1]



[図2]

121

URLID	対象URL	抽出対象部分の位置特定情報			...
		タイトル	住所	...	
1	http://aaaa.bbb.ccc/ddd/eee/001.html	<DIV class="title"> </DIV> 順番:1, /title/	<DIV class="address"> </DIV> 順番:1, /info/address/	...	...
2	http://aaaa.bbb.ccc/ddd/eee/002.html	<DIV class="title"> </DIV> 順番:2, /tps/table/	/<DIV.*>(<+></DIV>/ /住所:(.+)\$/ 順番:3, /info/address/	...	...
3	http://aaaa.bbb.ccc/ddd/eee/003.html	div#left h2, 順番:3, /tps/table/	#infoContent @ 所在地 ¥s+? (<+?> @is, 順番:5, /info/address/	...	...
...	...	...	...	...	...

[図3]

122
↘

項目ID	データ名	データ型	切り出し手法	関連	...
1	タイトル	文字	CSSセレクタ	住所、電話番号	...
2	住所	文字	CSSセレクタ、正規表現	施設名称、電話番号、位置情報	...
3	施設名称	文字	CSSセレクタ、正規表現	タイトル、住所、電話番号	...
4	電話番号	数字	CSSセレクタ、正規表現	タイトル、住所	...
5	更新日	日付	正規表現	タイトル、説明文	...
6	位置情報	緯度経度	CSSセレクタ、正規表現	住所、施設名称、電話番号	...
7	説明文	文字	CSSセレクタ	タイトル、更新日	...
...	...	...	...	...	...

[図4]

123

URLID	対象URL	記憶領域	...
1	http://aaaa.bbb.ccc/ddd/eee/001.html	D:¥page¥aaaa.bbb.ccc¥ddd¥eee¥001	...
2	http://aaaa.bbb.ccc/ddd/eee/002.html	D:¥page¥aaaa.bbb.ccc¥ddd¥eee¥002	...
3	http://aaaa.bbb.ccc/ddd/eee/003.html	D:¥page¥aaaa.bbb.ccc¥ddd¥eee¥003	...
...	...	...	...

124

[illegible]

[図6]

抽出対象部分の設定

抽出対象部分の選択	優先順位	抽出定義	変換処理	結果
タイトル 重要項目 ☐ addinfoに追加 ☐ ファイル保存 削除	0	Div#left h2		チエック
カテゴリ 重要項目 ☐ addinfoに追加 ☐ ファイル保存 削除	0	U#thead ((.+?)@s)		チエック
説明文 重要項目 ☐ addinfoに追加 ☐ ファイル保存 削除	0	#photoContent ((.+?) @hs)		チエック
住所 重要項目 ☐ addinfoに追加 ☐ ファイル保存 削除	0	<#infoContent (所在地 *>? (.+?) @hs)	replace / 1[0-3-1]/ / s replace / TEL: / / s	チエック 約100%

ソース

Http://aaa.bbb.cc/ddd/eee/00/.html

</p>
夜明け

<p>
<h3>開催地</h3>
〇〇神社

</p>
<h3>所在地</h3>
<p>
〒XXXX-XXXX
東京都豊島区〇〇〇〇〇〇〇〇
豊島区豊島市〇〇〇〇〇〇〇〇
</p>

21

22

23

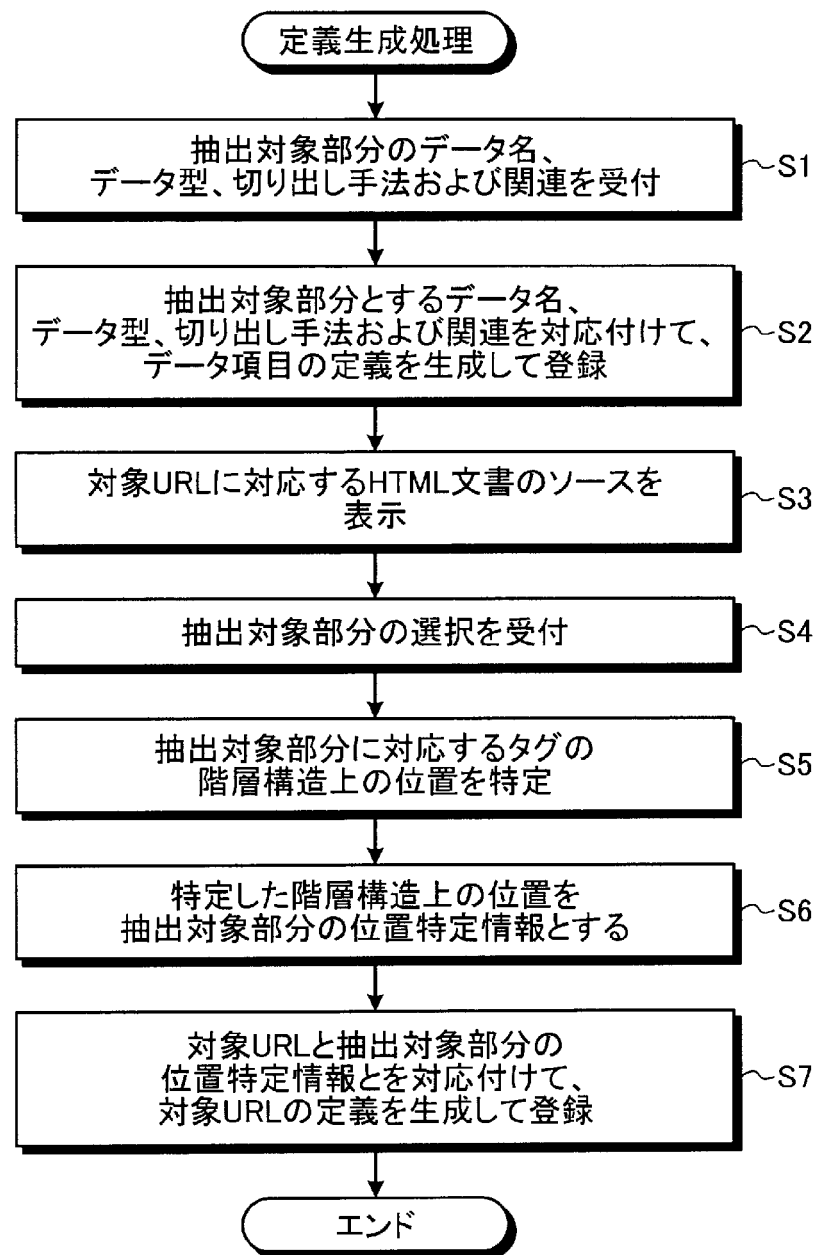
24

25

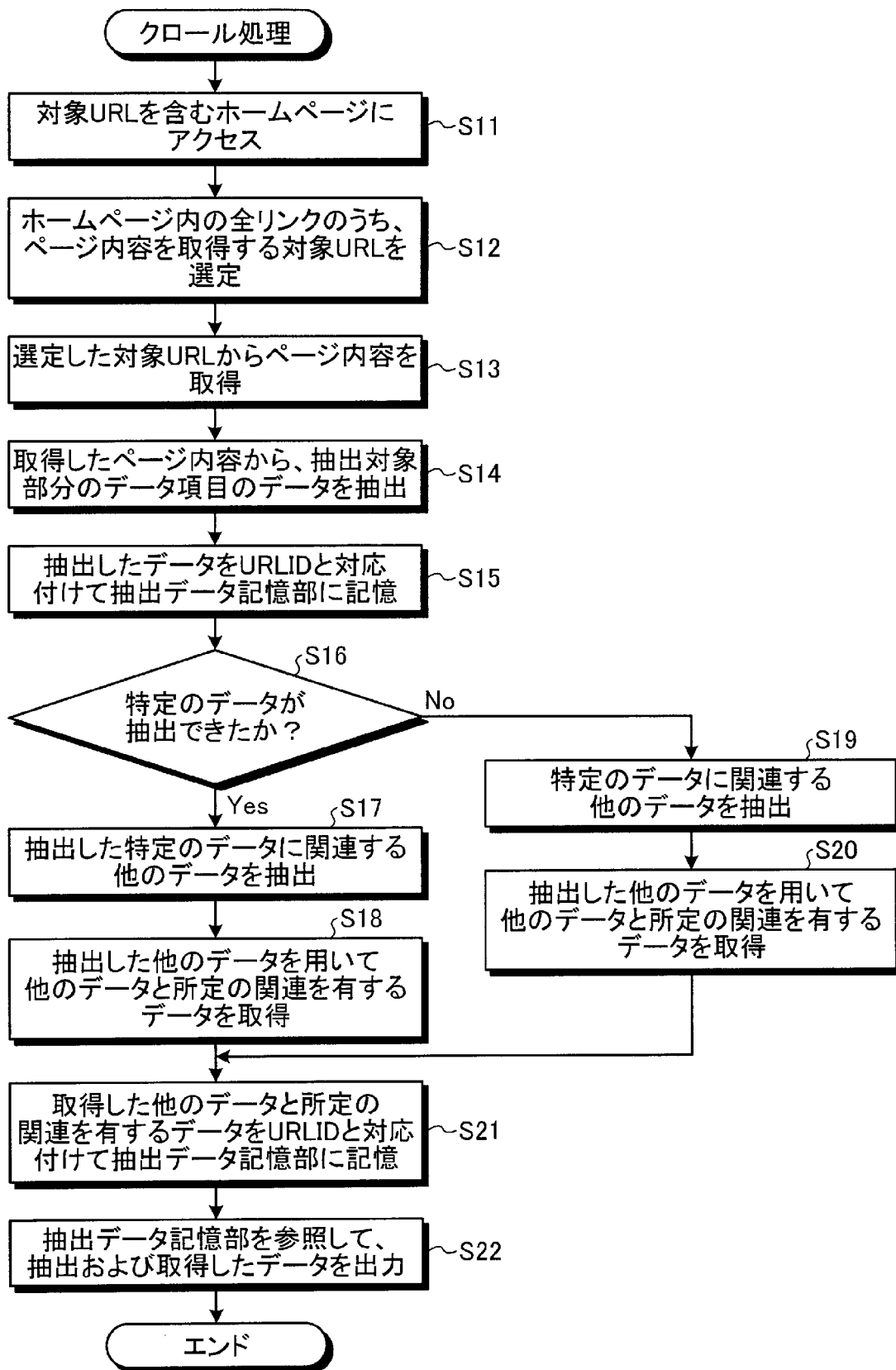
26

27

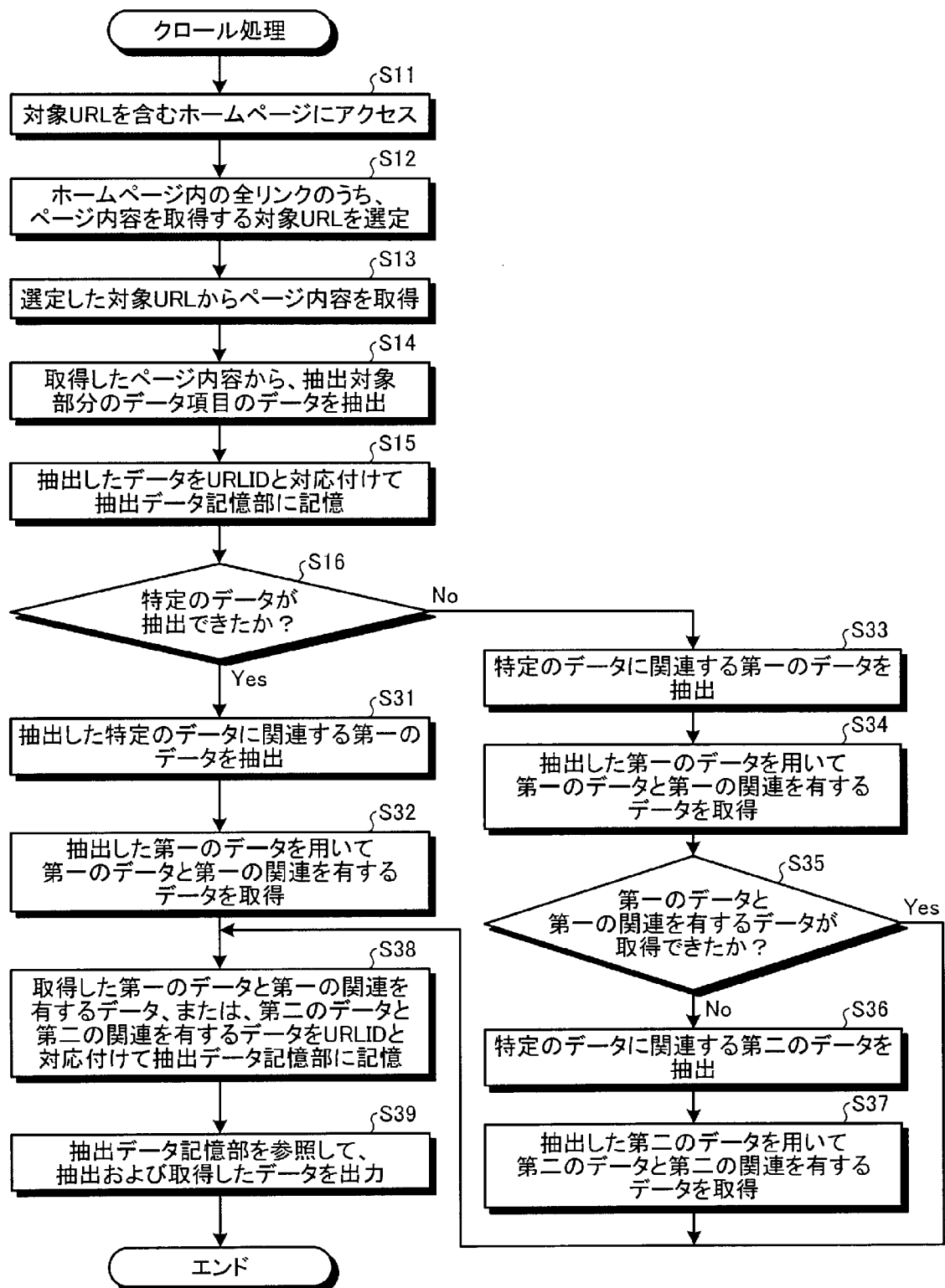
[図7]



[図8]



[図9]



[図10]

