カーリルおよび
スクレイピングについて

講義（5）「モデルサービスの企画意図と技術設計」

2011.08.24 学術ポータル担当者研修 NII会場

神原啓介（お茶の水女子大学）

http://sappari.org/

自己紹介：神原啓介

- お茶の水女子大学
  - お茶大アカデミックプロダクション
  - 特任リサーチフェロー
- **ユーザーインタフェース**（UI）の研究
  - 直感的に使えるコンピュータのデザイン

Webアプリの企画・開発

- **Nota Inc.**
  - カーリル
  - TwitPaint
- **株式会社はてな**
  - Rimo
  - はてなRSS
  - その他のサービスのUI
- 個人運営
  - Willustrator.org

カーリル：図書館横断検索


twitPaint
ホーム 自分 みんなの作品 お題 @返信 ログアウト
Twitterでお絵かきしよう
133
Width
Color
Opacity
Brush Preview
Undo
Clear All
コメント （140 - URL = 100文字以内）
TwitPaint
Twitterでお絵かきコミュニケーション


willustrator.org/image/edit/a22dc0fdc973c64f72a71744310c124f
Willustrator
kambara | Logout
« ほっしー
Save
Save & Publish (test)
100%
Open
Close Path
Fill :
0
Stroke :
1px
100
Text:
18pt
_sans
Layer :
6
/ 7
Edit :
Path
Size
Willustrator
Web上で協同利用するドローツール

私の語れること

- **Webアプリの作り方**
  - プログラマーやデザイナー任せにしない
  - 誰もがデザインに参加することを重視

- **UIデザイン手法**
  - 直感的に使えるようにするには
  - 使って楽しくするには

私の語れないこと

- 図書館システム
- 機関リポジトリ


日本最大の図書館検索サイト
カーリル™
図書館から探す（例：村上春樹、1Q84、絵本）
民主党
さがす
スタンプラリー開催中!!!
レシピ
今話題の本
図書館マップ
読みたいリスト
kambara (Google) | ログアウト | 図書館の設定
みんなの新着レシピ
断捨離（だんしゃり）
ユウウツな梅雨にぴったりの・・・
気分はネパール満喫中
もっと見る
今日のいいね！レシピ
銀河SFで星の世界に♪
ADHDの入門書
最近読んだ本やこれから読みたい本
もっと見る
話題のキーワード
Dragon Ash
今日、誕生日の作家
5人います。
借りたい一冊、見つかる。
全国5000館以上の図書館/図書室から現在の貸し出し状況が検索できます
1 図書館を選ぼう 2 本を探そう 3 図書館へ行こう
twitter でフォローしてください！
全国の図書館横断検索


日本最大の図書館検索サイト
カーリル
スタンプラリー開催中!!
レシピ
今話題の本
岩手県盛岡市
盛岡市
市立図書館 (詳細)
渋民図書館 (詳細)
都南BM (詳細)
農林水産関係試験研究機関総合目録
岩手医科大学
岩手大学
冷麺（れいめん、ネンミョン / レンミ
する。2.については冷やし中華を参
1.について説明
出典:Wikipedia
盛岡市の図書館を探しています
▼図書館を絞り込み
▼図書館を絞り込み
岩手県盛岡市
市立図書館
渋民図書館
都南図書館
市立BM
都南BM
⇒図書館の設定
盛岡冷麺物語
盛岡冷麺物語[繋新書]
小西 正人
蔵書あり(貸出可)
読みたい 読んだ
パスタマシンで麺道楽
パスタマシンで麺道楽 ーうどん、中華麺、韓国風冷麺、もちろんパスタ!
大森 大和
人気店が公開する調理技術 ラーメンつけ麺冷し麺
人気店が公開する調理技術 ラーメンつけ麺冷し麺
行くぞ!冷麺探険隊
行くぞ!冷麺探険隊 (文春文庫)
東海林さだお
初の全旅行記集
ラーメン・湯麺(たんめん)・冷麺(れいめん)―中国料理のコツ (新潮文
自分の住んでいる/働いている
場所近辺の図書館を一括検索


蔵書あり
盛岡冷麺物語
盛岡冷麺物語[繋新書]
小西 正人
蔵書あり(貸出可)
読みたい　読んだ
パスタマシンで麺道楽
パスタマシンで麺道楽
―うどん、中華麺、韓国風冷麺、もちろんパスタ!
大森 大和
蔵書なし
読みたい　読んだ
蔵書なし
盛岡市の蔵書：蔵書あり
市立図書館
貸出可
渋民図書館
貸出可
都南図書館
貸出可
予約する
amazon.co.jp 詳細ページへ
在庫あり ¥1,575 中古あり ¥1,000より

カーリルの裏側

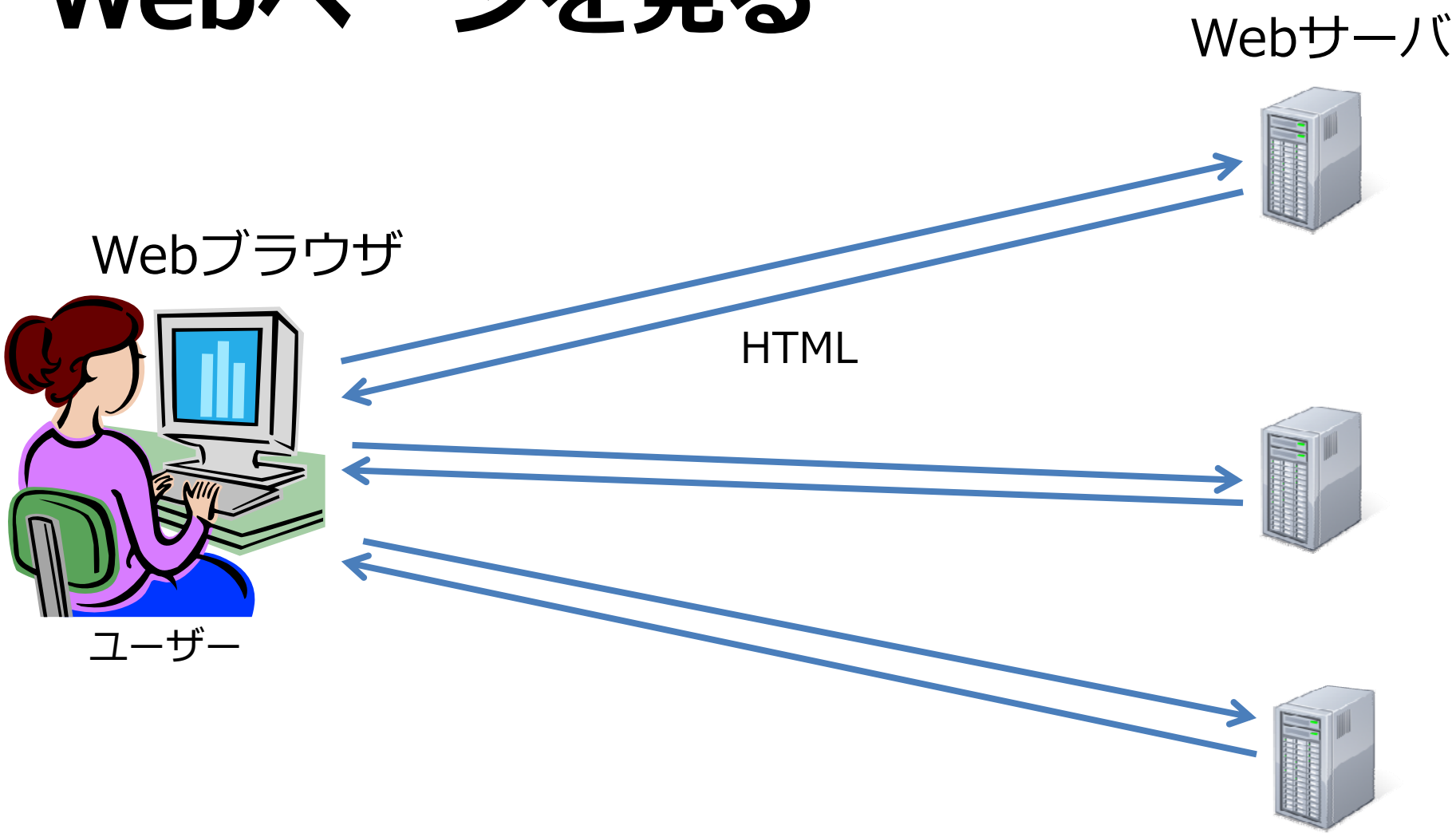
スクレイピング

スクレイピング

- プログラムでWebページを取得し、HTMLから必要な情報を取り出す

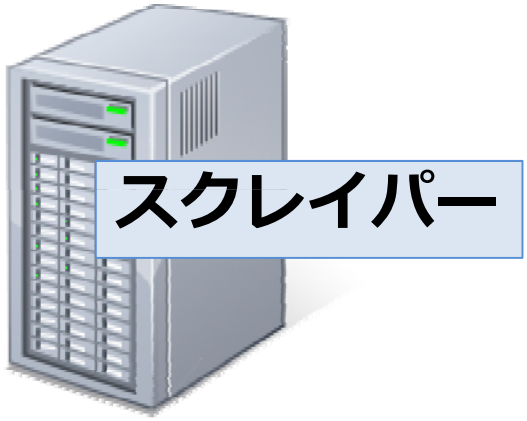
- Web APIを提供していないサイトが対象
  - 古いサイトなど

Webページを見る

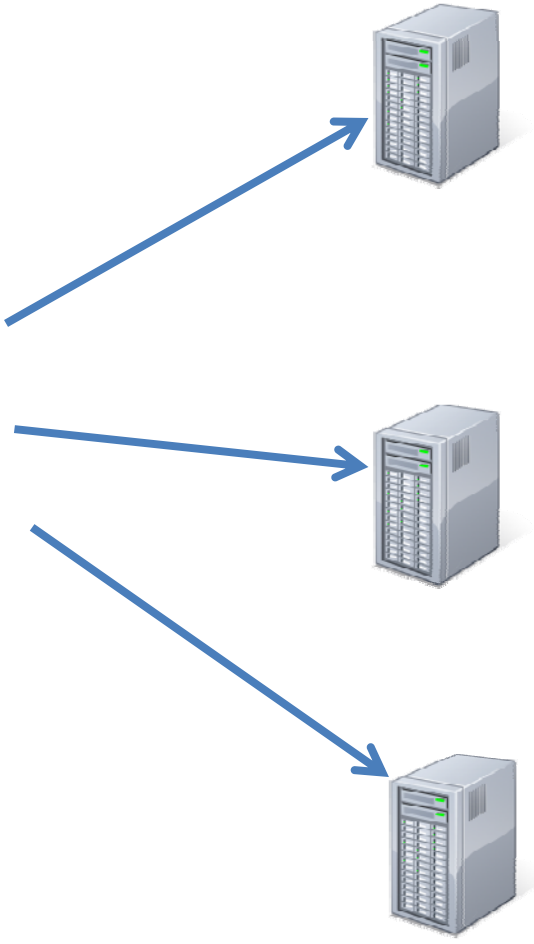


スクレイピングする場合

ユーザー

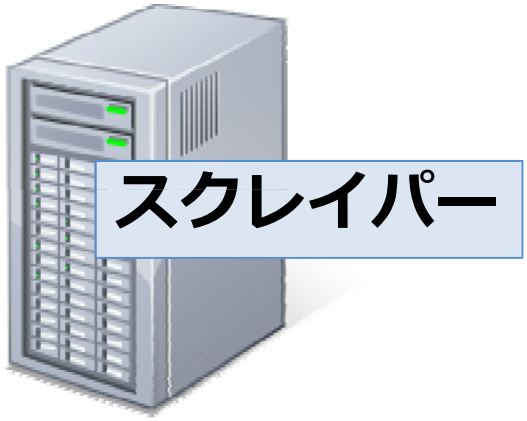


サーバ/PC



プログラムでページ取得

ユーザー



サーバ/PC

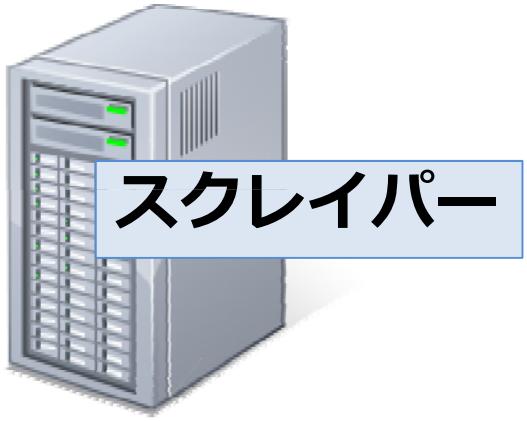
HTML

```
<body>
ゴチャゴチャ
<div>
  すごい情報
</div>
…
```

```
<body>
…
<div>
  役立つ情報
</div>
アレコレ
…
```

HTMLから情報を抜き出す

ユーザー



サーバ/PC

スクレイパー

HTML

<body>
ゴチャゴチャ

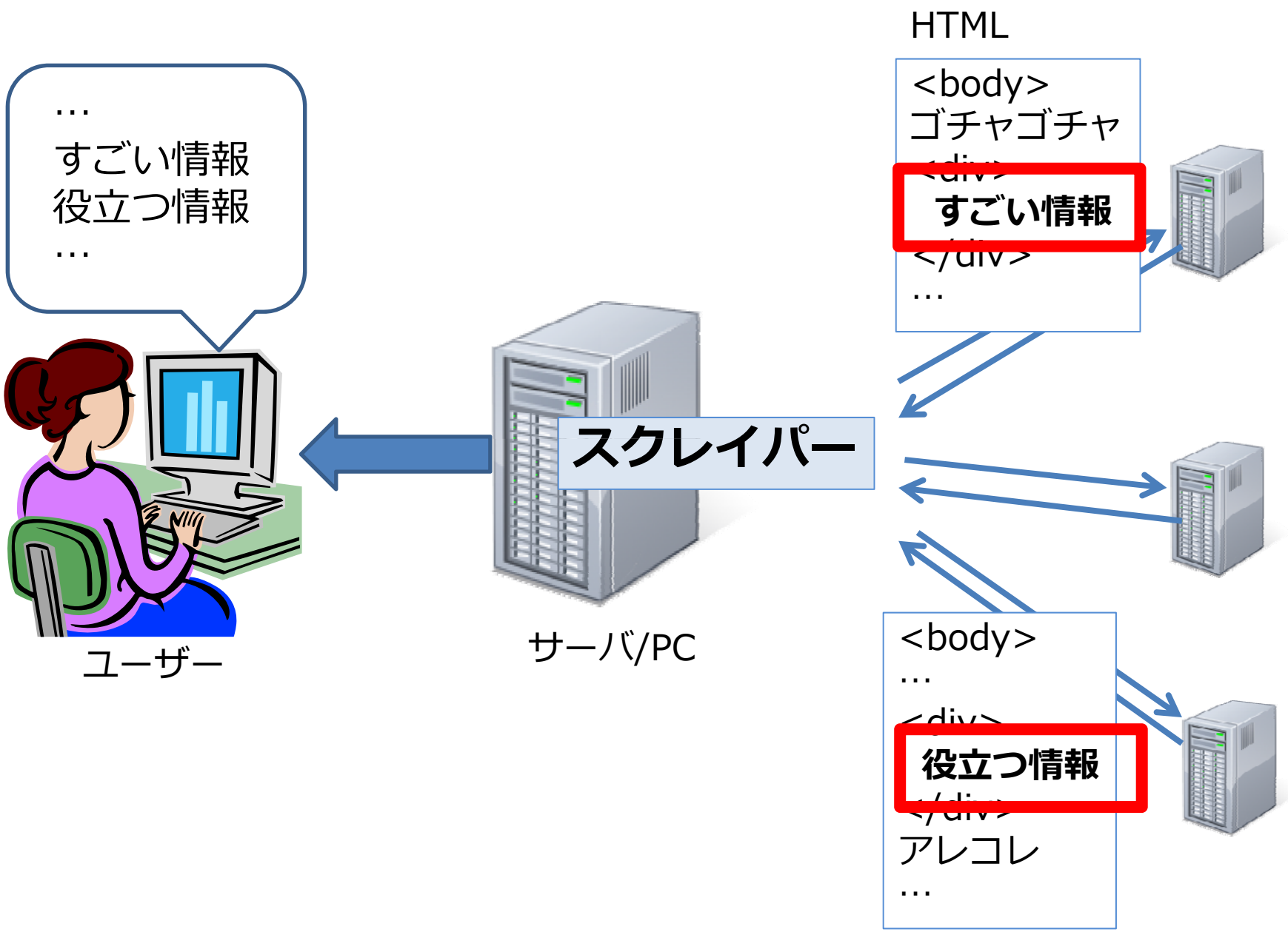
すごい情報

…

<body>
…

役立つ情報

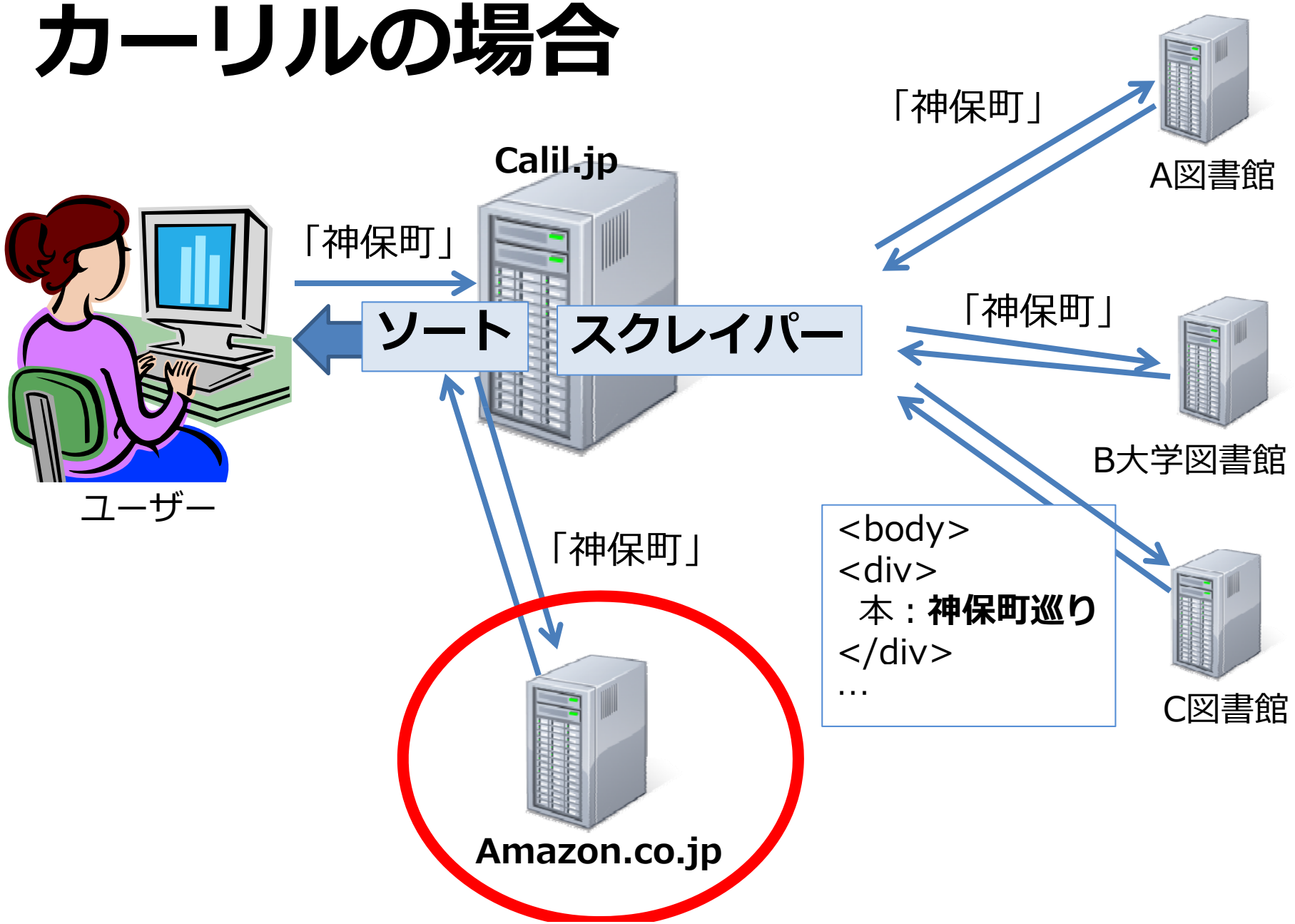
アレコレ
…


…
すごい情報
役立つ情報
…
ユーザー
スクレイパー
サーバ/PC
HTML
<body>
ゴチャゴチャ
すごい情報

…
<body>
…
役立つ情報

アレコレ
…

カーリルの場合


Calil.jp
「神保町」
ソート
スクレイパー
ユーザー
「神保町」
A図書館
「神保町」
B大学図書館
「神保町」
<body>

本：神保町巡り

…
C図書館
「神保町」
Amazon.co.jp

情報を抜き出す技術

- 問題：サイトやページによって
  HTMLの構造が違う
- XMLやHTMLの特定のノード（タグ）を
  指定する構文
- XPath (XML Path Language)
  - 例：
    `ul[@id="booklist"]/li[@class="book"]//span.title`
- CSS Selector
  - 例：`ul#booklist > li.book span.title`
  - 書きやすい

様々なスクレイピングツール

- 最近では楽にスクレイピングするための様々なツールが登場している

- 各プログラミング言語用ライブラリ
  - Mechanize, Nokogiri, htmlSQL
- 大手サイト用ライブラリ
  - ニコニコ動画用
  - Google＋用
- ScraperWiki
  - スクレイピング専用Webサービス
  - スクレイピング用のプログラムやデータを共有

スクレイピングの問題点

- HTML構造が変わると
  正しく情報を取り出せなくなる
  - その度にプログラムを直す必要がある
- やり過ぎると相手のサーバに負荷
  - Librahack岡崎図書館事件
- 権利関係がグレー

スクレイピングの問題点

- HTML構造が変わると
  正しく情報を取り出せなくなる
  - その度にプログラムを直す必要がある
- 相手のサーバに負荷
  - Librahack岡崎図書館事件
- 権利関係がグレー

図書館関係者と協力して
図書館Web API環境を整備したい

よろしくお願いします。