# ネット空間上の「交通の要衝」に関する情報の収集方法について

波床　正敏†

# Research notes on how to collect information on “transportation hubs” on the Internet

HATOKO Masatoshi†

**概要**

「交通の要衝」という概念を明らかにするため，基礎情報の収集をネット空間上に存在するサイトから今後実施する分析に用いるための多量のデータを収集した．本稿はその収集方法および工夫について説明したものである．通常は手作業でする事柄を自動化するとともに，サーバやネットに負荷をかけずに時間をかけてデータ収集した方法について説明している．

キーワード：交通の要衝，ネット検索，ウェブスクレイピング，自動化

**Abstract**

In order to clarify the concept of "transportation hub", a large amount of data was collected from sites existing in the net space. The data will be used for future analysis. This research note explains the method to collect the information and ingenuity of the collection. It describes how to automate things that are normally done manually and collect data over time without imposing a load on the server or the Internet.

Key Words: Transportation Hub, Internet Search, Web Scraping, Automation

† 大阪産業大学　工学部　都市創造工学科　教授

草 稿 提 出 日　10月31日
最終原稿提出日　10月31日

## 1　はじめに

### 1．1　背景

「交通の要衝」という言葉がある．例えば「この地は古来より交通の要衝とされ，地域の中心として発達してきた」といった使われ方がされるが，学術用語というわけではなく，賑やかそうな印象を表現しただけに見える．だが「交通の要衝」とされる地域は長年にわたって多少なりとも周囲よりも繁栄の程度が高く，その地位を長期維持していることが多いと思われる．つまり，「交通の要衝」には何らかの特徴があり，その特徴を捉えて交通整備や地域整備をすることは，地域の長期的発展に寄与する可能性が高く，今後のインフラ整備の基本方針に大きな影響を与える可能性がある．

では，「交通の要衝」の条件とは何だろうか．「交通の要衝」が街道の分岐点に発達する例は多く，通過交通量（交流量）が関係することは容易に想像できる．だが，それだけならば大都市近郊などは全て交通の要衝である．一方，船着き場などが「交通の要衝」と呼ばれることもあり，待ち時間等の一時的な滞留も「交通の要衝」と認識することに関係している可能性は高い．これら交流量や滞留時間はいずれも「量」に関しており，すでに多数の地域分析に関する学術研究が存在している．

だが，他に見落としている観点はないだろうか．「交通の要衝」が街道の分かれ道附近に発達したケースが多いことを考えると，量の視点以外にも交流の多様性（他の地点に比べて行き先の数が多いこと）が関係している可能性はないだろうか．この多様性が長期的繁栄を維持してきた原因になってきた可能性はないだろうか．

### 1．2　本研究の目的および本稿の内容

都市・地域計画学分野でも「多様」という表現はしばしば見かけるが，「多様」そのものについての研究は極めて少ない．本来は「多様」とは種類が多いことなので情報量に関する概念であり，貨幣価値のような概念とは互いに独立であるはずだが，換算可能な概念だと誤解されている可能性も高い．

そこで本研究では，都市・地域計画学分野で「多様」がどのように捉えられ，地域に影響してきたかを初歩的段階から実証的に研究することとした．その端緒として「交通の要衝」に着目し，日常生活交通レベルから全国的交通レベルまで，具体的に何が「交通の要衝」の要件になっているのか（また，これに多様の概念が含まれているかどうか）を確認する．また，「交通の要衝」の要件の程度（多様度の水準）を具体的に計測し，地域の持続的発展（サステイナビリティ）に与える影響やその構造について分析する．

以上が本研究全体としての目指すところであるが，本稿はこのうち「何が交通の要衝の要件になっているのか」を調査する部分について，webからの情報収集作業をした部分についての

状況をまとめたものである.

## ２．地域情報収集の基本方針

### ２．１　地域情報の収集方法とweb空間に存在する情報について

地域情報は地図や航空写真，統計情報等が基礎資料になることが多い．だが，本研究のような地域に対する認識を調査対象とする場合は，これら資料からはわからないため，現状調査ならば住民等にアンケート調査を実施し，過去の状況に対する調査ならば文献調査が標準的な情報収集方法になる．交通の要衝は歴史的に形成されてきたことを考慮すると，通常の方法ならば歴史的な地域情報に関する文献, すなわち地誌や郷土誌が調査対象になると思われる.（**表１**）

ところが，何が交通の要衝の要件なのかを調査するには，かなり広域的に多数の事例を収集する必要があり，全国津々浦々の地誌や郷土誌を多数収集して記載事項を確認して整理する方法は膨大な時間と費用と労力を要し，研究方法として現実的ではない．特に，地誌や郷土誌は出版部数が少なく，これらを全国的網羅的に収集することは極めて困難と思われる.

そこで，本研究では全国的な情報を容易に収集可能なweb上に存在する情報に注目し，これを収集･分析することとした．ただし，web上に存在する情報は基本的には現在の情報であり，歴史的な過去の蓄積はあまり期待できない．また，地誌や郷土誌が電子情報化されて掲載されている可能性も小さく，サイトの内容についても正確性が保証されていないことも多い.

**表１　地域情報の資料とその特徴**

| 種類 | 利点 | 欠点 |
|---|---|---|
| 地図・航空写真・統計等 | ・数値や施設等の配置情報が得られる | ・認識に関する情報は得られない |
| アンケート調査 | ・現状の認識が得られる | ・過去の認識は得にくい |
| 地誌・郷土誌等 | ・過去の状況が得られる | ・収集が困難なことがある |
| webサイト | ・広範な情報が得られる | ・基本的には現在の情報のみ<br>・正確性の保証がないことがある |

### ２．２　地域情報の収集方針と収集対象

本研究では，ある程度内容についての正確性が保証されているサイトの情報を利用する方針とし，基本的には各地方自治体の公式webサイトを利用することとした．自治体のサイトには自地域を紹介する内容が含まれていることが多く，その沿革を紹介する中で「交通の要衝」について言及されている可能性がある．本研究では市区町村（区は東京都特別区のみ）のwebサイトを丸ごとダウンロードしていったん保存し，そのデータを用いて分析することとした.

市区町村の中には昭和から平成にかけて合併により消滅している自治体もある．これらについても合併後の自治体が現市町村の一部として紹介している可能性があるので現行市区町村と

同様に調査対象とする．

一方，市区町村の公式webサイトだけでは「交通の要衝」についての情報が不足する場合があり得る．そこで，情報の正確性は低下する可能性はあるが，多数の情報を集める方法として，各市区町村名と「交通の要衝」の２つのキーワードを使ってweb検索サイトからURLを取得し，これを補助的な情報源としてダウンロードして保存し，分析することとした．

上記の２種の収集方法ともに，保存および分析対象はテキストによる記述を中心とし，自動的な分析が困難である画像や動画等は除いた．

### ２．３　地域情報収集の技術的課題と解決方針

webからの情報収集時の課題としては，まず，調査対象サイト数が非常に多くなることである．現在の市区町村数約1,700および全都道府県数47に加え，昭和後期および平成の大合併以前に存在した市町村数が約1,800あり，これらすべてについて手作業で隅々までサイトを調べて情報収集することは困難である．そこで，収集作業は自動化することにする．全体の流れのイメージを**図１**に示す．

市区町村等の調査対象サイトそのものについても，URL一覧がどこかに準備されているわけではなく，検索サイト等を利用した情報収集が必要である．上述のように多数あるので手作業は困難であり，一覧の作成を自動化することにする．**図１**の左上の点線の枠内部分がこれに該当する．作成したURL一覧について，データをダウンロードするが，これも自動化する（**図１**の左下の点線の枠内部分）．

さらに各市区町村名（もしくは廃止された市町村名）と「交通の要衝」のキーワードを使って検索サイトからURLを取得する作業についても手作業は困難であるため，自動化することにする．**図１**の左下の点線の枠内部分がこれに該当する．

なお，検索作業等を自動化すると，調査対象サイトや検索サイトに多大な負荷をかけることになって好ましくないため，自動化はするものの手作業と同程度になるように情報収集速度を調整することにした．詳しくは次章以降で説明する．

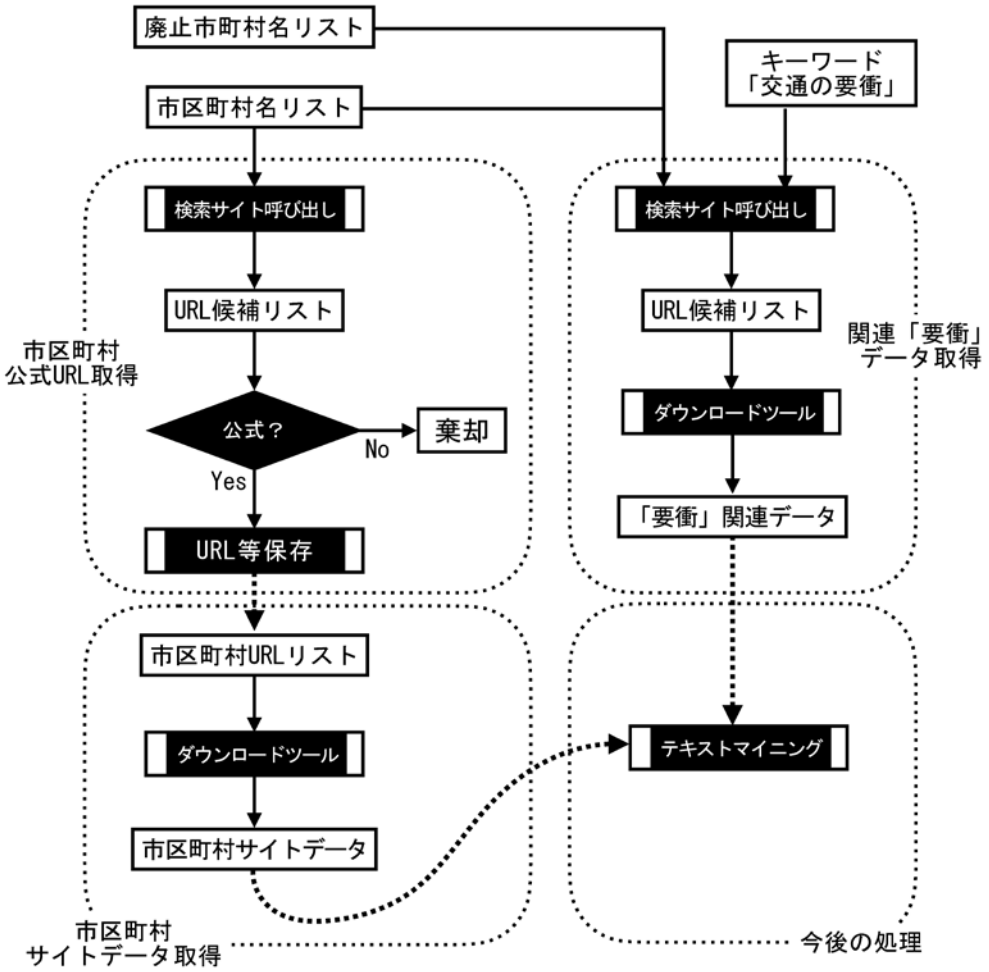



**図１　情報収集作業全体の流れのイメージ**

## ３．市町村の公式webサイトURLの取得

以下は主として**図１**の左上の点線枠内およびその上部に書かれているデータ作成についての説明である．

### ３．１　市区町村名リストの作成

調査対象とする現在の市区町村一覧については，平成27年（2015年）の国勢調査における都道府県・市区町村別主要統計表をもとに「市区町村名リスト」のデータを作成した．データ数は1,741件である．また，平成24年（2012年）版の地方財政白書をもとに，昭和60年（1985年）度以降に合併を実施したこと等によって名称が消滅した市町村一覧を作成し，これを「廃止市

町村名リスト」のデータとした．データ数は1,822件である．

両者とも，データの内容としては，市区町村コード，含まれる都道府県，市区町村名であり，「廃止市町村名リスト」における市区町村コードは仮の番号を新たに割り振った．データに都道府県を含めているのは，同じ都道府県下に同一名称の町村は存在しないが，全国的には存在しているためである．データの形式はMS-EXCELのワークシート形式（.xlsx）である．

本章で説明する作業に使用するのは前者の「市区町村名リスト」であり，「廃止市町村名リスト」については第５章で説明する作業で使用する．

### ３．２　検索サイトの利用と自動化

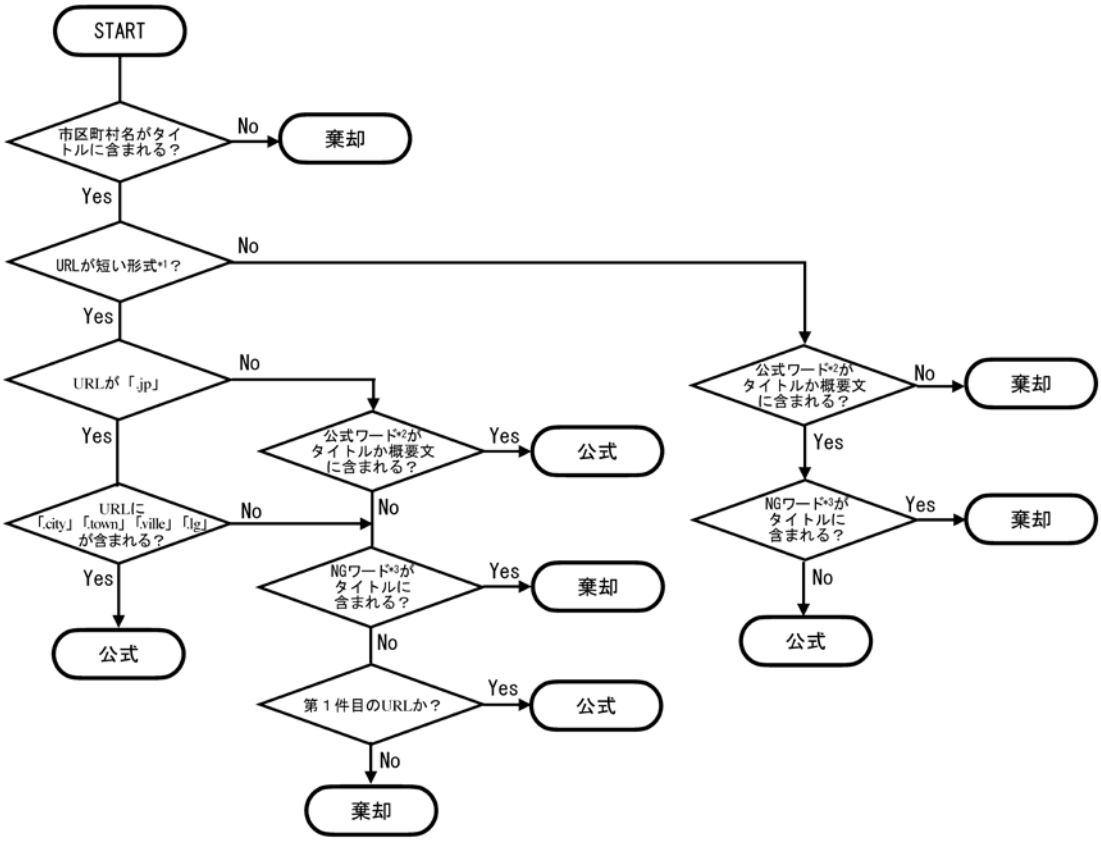
以下は前述のように，調査対象とする市区町村の公式webサイトのURL一覧が信頼できる形でどこかに準備されているわけではないため，検索サイト等を利用した情報収集を行った．

自動化するためにPythonでプログラムを作成し，このプログラム内からwebブラウザのGoogle Chromeを自動操作（Seleniumを使用）した．検索条件として都道府県名と市区町村名とを指定し，webブラウザに画面として表示される情報を取得する．10件程度のURLが示されるため，プログラム内に取り込んだ上で次節において説明する方法で公式サイトの判別を行い，市区町村コード，都道府県，市区町村名，URL，web画面に表示される概要文をファイル（MS-EXCELのワークシート形式）に保存した．実行環境はMac OS（Mojave）のターミナル上である．

そのままプログラムを走らせた場合，短い時間内に多数回の検索を繰り返すことになり，検索サイトや学内ネットに大きな負荷をかけることになる上，検索サイトから接続を拒否される可能性もある．そこで，ほぼ手作業での検索と同程度の操作速度となるように，新規ウィンドウを開く操作，検索キーワードを入力する操作，検索開始のボタン操作，内容の取得操作などの主要操作ごとに数秒の待ち時間を設けるようにプログラムを記述した．これにより，連続して無人の自動検索を実施した．自動作業は概ね１日あれば完了した．

### ３．３　URL候補の選別方法

前節の説明の方法で取得した検索結果画面には10件程度のサイトが示され，かなりの率で第１件目が市区町村の公式URLではあるが，必ずしも第１件目が適切とは限らない場合もある．そこで，webのタイトルや概要文，URLの形式などからプログラムによって以下の方法（**図２**）でサイトを判別した．なお，同図のフローは，試行錯誤の結果である．




*1 「http://hoge.hoge/」または「http://hoge.hoge/index.???」の形式
*2 「公式」「オフィシャル」「役場」「役所」
*3 「観光」「振興局」「商工会」「交通局」「開発局」「図書館」「秒後に」「水道」

**図２　市区町村公式URL判定フロー**

## ４　市町村の公式webサイト内容の取得・保存

以下は主として**図１**の左下の点線枠内についての説明である.

### ４.１　ダウンロードツールの利用

前章で取得した各市区町村のサイトURLにおいて,「交通の要衝」に関する記述の記載位置が不明なため,サイト内のデータをそのままダウンロードして保存した.ダウンロード後のファイルを使って「交通の要衝」の使われ方を分析することになるが，様々な条件を変えて複数回サイトにアクセスする可能性があるため，分析条件を変更する都度アクセスするよりはサイトデータを一括してダウンロードした後に分析を行った方が利便性が高く，ネット上のデータ転送量も少なくなると考えた．基本的にはサイトをそのまま保存するだけであるので（Python

でもBeautiful Soupというライブラリを組み込んでコードを作成すれば階層的な構造でもダウンロード可能であるが），独自のプログラム作成はせずに既存のダウンロードツール（wgetコマンド）を利用した.

ダウンロード対象は前章で取得した各市区町村の公式URLのサイトであり，トップページおよび５階層下までを対象とし，階層構造を維持したまま保存した．そのままツールを実行すると，アクセス先および学内ネットに大幅な負荷をかけることになり，特にアクセス先の市区町村の業務に支障を来す恐れがある．そこで，２MB/secの帯域制限を課すとともに取得間隔を１～２秒おきとした（後に２MB/secの帯域制限はあまり意味がないことがわかったので制限解除）．また，サイト内でのリンクが同じファイルを指し示していて既に取得済みの場合はダウンロードを省略した.

### ４．２　バッチ処理と並行処理

前章で取得した各市区町村のサイトURLをもとにバッチファイルを作成し，実際のダウンロード作業は，Windows10のコマンドプロンプト上でバッチ処理により実施した.

多数のサイトを下位階層までダウンロード対象としているためデータ量が多く，また帯域制限を課しているため，処理の完了までにかなりの時間を要することが予想された．そこで，ダウンロード作業に用いるPCを20台程度準備し（元々は遺伝的アルゴリズムの計算処理用），同時に並行して作業を行った．また，作業期間を授業期間外に設定した．これにより，アクセス先および学内ネットにおける負荷の問題を軽減した. データのダウンロード開始から完了まで，概ね３ヶ月を要した.

### ４．３　不要データの削除と保存

本研究ではテキストマイニングにより分析を行おうとしているため，基本的にはテキストデータ（HTMLファイルやPDFファイルに記載されている内容等も含む）を必要としているものの，それ以外の画像ファイルや動画等は処理できないことが予想されるため，不要である．サイトを深い階層までダウンロードした場合，処理対象外のデータも同時にダウンロードされるため，分析作業に先立つ処理として以下の種類のファイルを保存対象から除外して削除した．不要データ削除後のデータ総容量は約7.3TB（市区町村のweb）となった.

①画像ファイル（拡張子がjpeg，jpg，gif，bmp，png，svg，tif，tiffのもの）
②音声ファイル（拡張子がwav，mp3のもの）
③動画ファイル（拡張子がmp4，mov，wmvのもの）
④書式情報やスクリプト等（拡張子がcss，xsl，jsのもの）
⑤フォントファイル（拡張子がeot?，ttf，woff，woff?のもの）

## 5　キーワードを含む様々なサイトのURLの取得と内容の取得・保存

以下は主として**図1**の右上の点線枠内についての説明である.

### 5.1　キーワードの選定とダウンロード対象

市区町村の公式サイトの情報以外に，各市区町村名（もしくは廃止された市町村名）と「交通の要衝」のキーワードを使って検索サイトからURLを取得し,そのサイトの内容を取得した.「交通の要衝」については，しばしば同じ意味で「交通の要所」と表記されているケースが散見されることを考慮し，以下の2つを検索のキーワードとした.

①3.1で作成方法を説明した「市区町村名リスト」もしくは「廃止市町村名リスト」に記載の市区町村名.

②「交通の要衝」もしくは「交通の要所」.

検索結果として取得できたURLについては，そこに「交通の要衝」に関する記述が存在することがわかっているため，さらに下層のデータまで取得する必要がない．そこで，この章でのダウンロード対象は（リンク先に補足説明が存在する可能性などを考慮して）1階層下までとした.

### 5.2　検索サイトの利用と自動化

web空間上のどこに「交通の要衝」という記載があるかどうか不明なため，ここでも検索サイトを利用した情報収集を行った．Pythonで作成されたプログラム内からChromeを自動操作し，前節①②の検索キーワードを与えてgoogleで検索し，ブラウザに画面として表示される情報を取得する．検索結果は1面あたり10件程度表示されるため，タイトル，URL，概要文を取得し，市区町村コード，都道府県名，市区町村名，表示順位などの情報とともにファイル（MS-EXCELのワークシート形式）に保存する.

1面分の処理のたびに取得したURLにアクセスし，プログラム内からダウンロードツールを呼び出し，次節で説明する方法でサイトのデータを取得して保存した．1組の検索キーワードに対して検索結果は次面を次々と表示させることで10〜13面分程度表示させることが可能なため,これらを可能な限り取得する（計120サイト程度）．実行環境はMac OS（Mojave）のターミナル上である.

この処理でも，短い時間内に多数回の検索を繰り返せば検索サイトに負荷をかけることになるため，Chromeの主要操作ごとに数秒の待ち時間を設けるようにプログラムを記述した.Chromeの画面表示1面分ごとに示されたURLからのダウンロード処理（10件程度）を挟んだため，検索サイトへのアクセス頻度はさらに下がっており，手動操作なみのアクセス頻度以下になった．これにより，連続して無人の自動検索を実施できた.

現行の市区町村と廃止市町村合計3,563それぞれにつき100～120サイトの情報があるため，４台のPCで並行処理を行った．この作業についてもデータのダウンロード開始から完了まで，概ね３ヶ月を要した．総容量は約1.6TB（キーワードに基づく検索結果）となった．

### ５．３　ダウンロードツールの利用

この章で説明しているデータ取得についても，既存のツール（wgetコマンド）を使用する．理由は４．１での説明と同様で，条件を変えて複数回分析する可能性があるため，都度アクセスするより保存したデータを使用した方が適切と考えたからである．また，データ取得段階ではツールで準備されている機能以上の作業は要求されないためでもある．

ダウンロード対象は５．２で取得したURLのサイトであり，トップページおよび１階層下までを対象とし，階層構造を維持したまま保存した．アクセス先および学内ネットの負荷軽減のため，データ取得間隔を１秒おきとし，作業期間を本学の授業期間外に設定した．

## ６　まとめと今後の分析について

本稿では，地域の多様に関する概念に関する研究としての「交通の要衝」についての調査のうち，web上に存在する情報を多数収集する方法についてまとめた．このような作業については，手間がかかる割には主題となっているテーマに関する研究論文をまとめる際には端折って説明されるだけであるため，この機会にまとめておくこととした．他の研究で同様の作業をする際に参考となれば幸いである．

データを取得した結果，データ量が当初予想以上に多くなってしまい，バックアップを含めて大容量のハードディスクを複数台用いる状況になっており，データのコピーや移動だけでそれぞれ数日を要する状況に陥っている．

現在，取得したデータに含まれているキーワードの使われ方について分析作業を行っているが，分析結果が出そろうまでにはまだ幾分の時間を要する状況にある．完了し次第，何らかの方法で公表の予定である．

**謝辞：**

本研究は令和元年度分野別研究組織「交通の要衝の概念に端を発する交流の多様性が地域構造に与えた影響に関する研究（代表：波床正敏）」のうち，web上に存在する情報を多数収集する方法についてまとめたものである．ここに記して感謝の意を示したい．