How should web scraping be organised for official statistics?

Nigel Swier
Office for National Statistics, Newport, United Kingdom – nigel.swier@ons.gov.uk

1. Introduction

Many National Statistics Institutes (NSIs) are in the process of establishing what big data means for official statistics. Gaining access to big data is a major challenge since private companies hold much of it. However, Internet web pages are in the public domain and are relatively accessible. This makes web scraping a logical departure point for NSIs starting out in the world of big data.

There are many examples of web scraping taking place within NSIs. There are numerous applications in the domain of price statistics covering items such as clothing, house prices, airline fares, cinema tickets (Ten Bosch, 2016), as well as groceries (Breton et al, 2016). Another application of interest is web scraping on-line job advertisements for labour market statistics (UNECE, 2015), (Körner et al., 2016). Bacaroli et al, (2014) propose that some enterprise statistics (e.g. e-commerce) could be produced using data scraped from company websites rather than through traditional survey instruments.

Not all NSI web scraping use cases involve direct use for official statistics. The UK Office for National Statistics (ONS) is experimenting with web scraped property sales data to enhance its Census address register. ONS have also undertaken small projects that have scraped data of consumer electronics for making hedonic adjustments for price statistics and council tax bands[1] for enriching other data sets.

Despite its potential, most NSI web scraping projects are small-scale and experimental in nature. Larger scale organisational approaches for web scraping are not well established and it is not clear how NSIs should scale up for statistical production. Within ONS there is a debate as to whether web scraping should be done in-house or whether it should be outsourced to a third party.

The aim of this paper to is describe some potential approaches for organising web scraping for official statistics and to identify the main factors that could help shape organisational decisions. Web scraping can take many different forms and so appropriate arrangements may involve a mix of approaches. However, strategies for moving web scraping into production should generally be aiming to rationalise activity to achieve economies of scale.

2. What is web scraping?

Web scraping can be defined as the process of automatically collecting information from the Internet, using tools (called scrapers, internet robots, crawlers, spiders etc.) that navigate, extract the content of websites and store scraped data in local databases for subsequent elaboration purposes (Bacaroli et al, 2014).

[1] Council tax bands are part of the system of local property taxation in England, Wales and Scotland.

There are several broad technical approaches to web scraping:

2.1 Programmatic web scraping:

Programmatic web scraping approaches involve the design and development of computer programmes that navigate the structure of a website and extract information from a web page. There are different approaches varying in complexity and a range of tools for performing all, or part of the process. Some commonly used tools include Python Scrapy, Selenium and Apache Nutch. Programmatic web scraping solutions are suitable for developing resilient applications for statistical production.

A further distinction can be drawn between specific programmatic approaches that target parts of a website and more generic approaches that first crawl and download web content, and then separately parse and extract the target information. For example, a robot developed to collect specific product information from a retail website is specific web scraping. In contrast, an application designed to identify websites engaged in e-commerce by analysing website content requires a more generic approach.

2.2 “Point and click” web scraping:

Point and click web scraping tools enable the extraction of data from web pages using an intuitive user interface. The ‘point and click’ actions of the user generate scraping code that can then be executed either for a one-off collection or to be reused subsequently. Some examples include Import.io, Dexi.io, and Kimono Labs. These tools are designed for simple applications although they may also offer more complex functionality (e.g. crawling).

These tools are best suited for one-off or smaller scale web scraping projects. The main advantage of these tools over programmatic approaches is that clerical staff can learn to use them very quickly. For NSIs this presents an opportunity to increase the efficiency of existing clerical processes involving the manual extraction of data from websites.

2.3 Application Programming Interfaces (APIs):

Some websites allow access to an underlying database via an API. Although this is not strictly web scraping, it is a means of acquiring similar type data. These APIs exist to support the business model of the website. For example, a large job or property search engine may offer API functionality to enable other portals to publish links to advertisements elsewhere and then receive a fee on a click-per-view basis.

This is generally an easier and more resilient method of extracting web content than programmatic approaches. However, APIs often have limitations; for example, some limit the amount of data returned from a single query. In addition, APIs should be used in accordance with their terms and conditions, since they are not provided for the purpose of supporting statistical production.

3. Legal Issues

Although the official statistics community has identified many legal and ethical issues for web scraping, there is not yet an agreed position. On one hand, there is a concern that web scraping could potentially breach database rights of website owners (European Commission, 2014). On the other, national statistical laws and codes of practice empower NSIs to collect data for statistical purposes, and web scraping may be consistent with these frameworks (Stateva et al, 2016). In the face of this uncertainty, most NSI web scraping projects follow the advice from their own legal departments.

However, good practice for NSI web scraping projects may include the following (ibid):

- Respect the robots.txt exclusion protocol: this enables website owners to declare specific parts of their website they do not want scraped.
- Create user-agent strings to identify your organisation to the website owner and a means of contact.
- Be transparent about web scraping activities, for example, by declaring such activities on a website.
- Notify owners if scraping large volumes of data on a regular basis.
- Take steps to minimise the impact of web scraping activities on website owners. For example, add idle time between requests and scheduling activity when the website load is low (e.g. in the early morning).

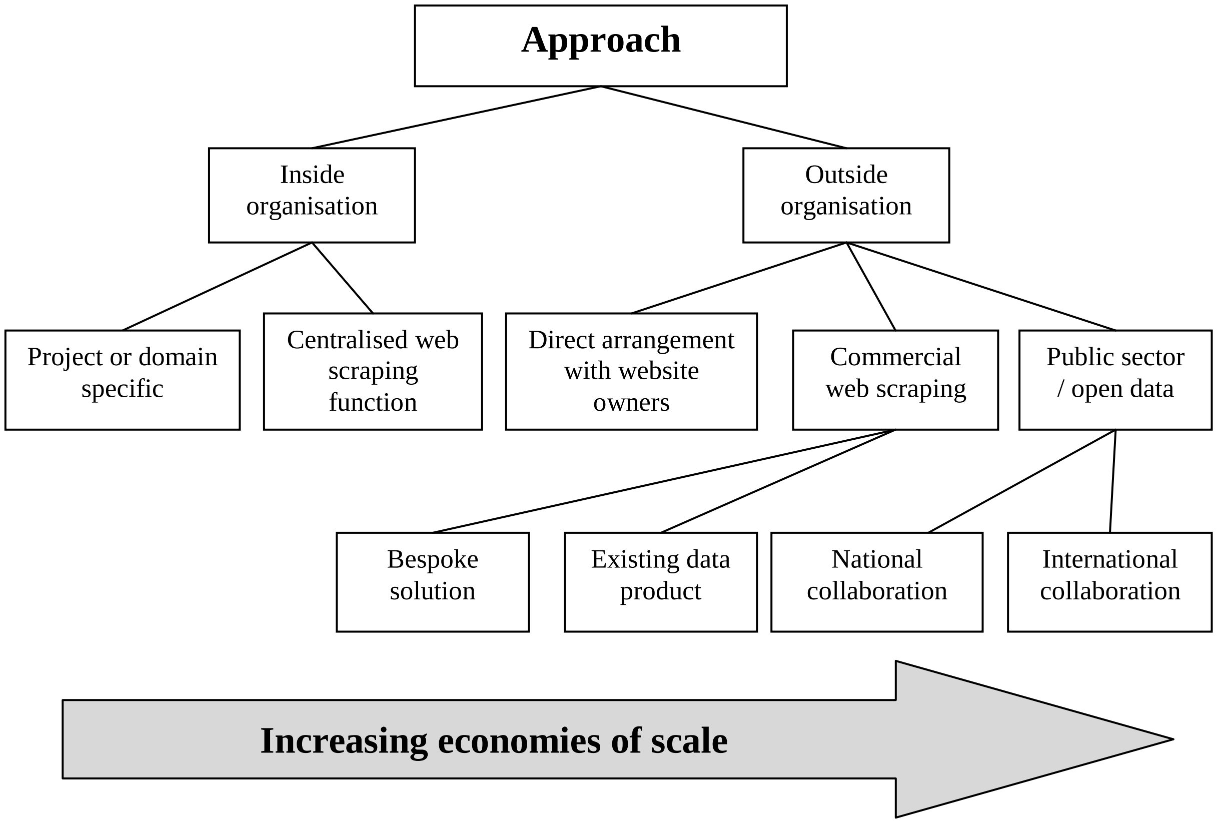
4. Organisational approaches to web scraping

There is a distinction between web scraping as an activity and how it is done in an organisational context. The different approaches are outlined in Figure 1. First, web scraping can be done from either inside or outside the organisation. If done from inside it could be carried out within a domain area, or as part of a centralised function. If the web scraping is done externally this could be through a commercial arrangement with a third party or part of a wider public sector (or even an open data) initiative.

Commercial web scraping arrangements can be broken down into either bespoke solutions or procurement of an existing data product. A bespoke solution would involve contracting out specific web scraping activities to a third party. The procurement of existing data products may be problematic since NSIs are increasingly taking a position of not paying for data. In addition, commercial data products often use proprietary black-box methods, which are incompatible with official statistics principles.

Another possibility is direct arrangements with the website owners to get access to the data underpinning the website. There are various possibilities for delivering data including bespoke file extracts or access via an API or secure web server. Direct arrangements will usually be necessary for applications requiring historical data. A further possibility is the deployment of widgets in commercial web sites to capture, not just web content, but also system interactions such as page views or even sales data. Although these approaches go well beyond the normal definition of web scraping, it is useful to consider these possibilities as part of a spectrum of options from web scraping sites without formal agreement through to deeply integrated arrangements with website owners.

Figure 1: Approaches for acquiring web scraped data


Approach
Inside organisation
Outside organisation
Project or domain specific
Centralised web scraping function
Direct arrangement with website owners
Commercial web scraping
Public sector / open data
Bespoke solution
Existing data product
National collaboration
International collaboration
Increasing economies of scale

Coordinated public sector initiatives for using web scraped data are not well developed at a national or international scale. However, the European Centre for Vocational Training (CEDEFOP) is leading one such project. This involves web scraping job portals in EU member states to inform the design of training programs using the skills information published in on-line job advertisements (Kvetan, 2016). A 2015 pilot involving five countries will now be extended to cover all EU member states. There are also some initial discussions about the possibility of wider cooperation between CEDEFOP and others with an interest in this kind of data, including elements of the European Statistical System (ESS). This cooperation could conceivably lead to job vacancy data being available to service a wide range of public sector organisations across Europe.

While many web scraping projects start at a small scale, production level systems require more investment. Production level web scraping applications are complex systems that require a robust IT infrastructure and ongoing maintenance by specialized technicians. It may not be financially viable to do this for a single domain area within an NSI. However, it may become viable as the overall scale of web scraping activity expands and economies of scale increase. This expansion could involve either the rationalization of activity across domains, across organisations, or both. NSIs should develop strategies that consider the overall strategic scope and configuration of current and future web scraping activities.

An interesting feature of web scraping is that, unlike traditional data collection methods, it is not constrained by national boundaries. Official surveys are subject to national statistical laws while administrative data derive from systems designed to support the operation of national governments. In contrast, web data does not face any geographical constraints. The Billion Prices Project[2] is a clear example of how one organisation can collect web data across national boundaries. Therefore, there is an opportunity for international bodies to lead collaborative efforts to collect and process web data for the benefit of the international user community.

It is clear that there are a number of approaches to acquiring web scraped data and there are a range of factors to consider when determining which approaches are suitable. The main factors include:

- What is the purpose? Is the data needed research or production purposes?
- What is the scale and complexity of the activity?
- How many target websites are involved?
- Will production be infrequent or an ongoing activity?
- Is there a requirement for historical data?
- Who are all the potential users of the data?
- What is the level of integration with existing statistical systems?

An example of where integration issues are important would be the example for web scraping for enterprise statistics referred to the introduction. A feature of this approach is the use of enterprise website URLs linked to a statistical business register. Given the level of integration with existing statistical infrastructure, outsourcing this web scraping activity is not really an option.

5. Conclusion

Given the various situations for using web scraped data for statistical purposes, it is not possible to recommend a single organisational approach. Point and click applications require minimal investment and can easily adopted in-house. In contrast, web scraping requirements for more complex applications could be met in different ways, including through a centralised web scraping function, outsourcing, or multi-organisational arrangements. However, certain highly integrated approaches may need an in-house solution.

Long-term strategies for web scraping should therefore consider the overall scope and configuration of current and future activities both within and outside the organisation. One critical question for NSIs is whether they see themselves as simply a customer for web scraped data or whether they could have a additional role as a supplier, particularly to other parts of government. Given the Internet's lack of boundaries, international bodies could play a key role in leading and coordinating the collection of web data for the benefit of the international data community.

[2] http://bpp.mit.edu/

References:

Barcaroli G., Scannapieco M., Summa D., Scarnò M. (2014). Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies, Eurostat

Breton R., Flower T., Mayhew M., Metcalfe E., Milliken N., Payne C., Smith T., Winton J. Woods A. (2016). Research indices using web scraped data: May 2016 update, Office for National Statistics

European Commission (2014). Analysis of methodologies for using the Internet for the collection of information society and other statistics

Körner T., Rengers M., Swier N., Metcalfe L., Jansson I., Wu D., Nikic N., Pierrakou C. (2016). Inventory and Qualitative Assessment of Job Portals, Eurostat

Kvetan V. (2016). Exploring on-line vacancies: What have we learnt so far?, Eurostat Conference for Social Statistics

Ten Bosch O. (2016). Webscraping at Statistics Netherlands, Eurostat

Stateva G. Ten Bosch, O., Maślankowski J., Righi A., Scannapieco M., Greenaway M., Swier N., Jansson I., Wu D. (2016). Legal aspects related to web scraping of enterprise websites, Eurostat

UNECE (2015). Experimental Report: Job Vacancies