10. Social Data APIs

Origin, Types, Issues

Cornelius Puschmann & Julian Ausserhofer

Introduction

Application programming interfaces (APIs) represent an increasingly relevant form of data access in both academic research and beyond. Many popular social media services provide APIs to developers, which can also be used to collect information relevant to social scientists, industry researchers and journalists, as do some more traditional data providers, such as archives and databases. In science and scholarship, fields that study digital media, and some which are concerned with other areas of inquiry (economics, climatology, medicine), have taken up the use of APIs. As more and more information becomes available online, providing access in a standardized way is a convenient and efficient way for turning data that is generated by users (Facebook, Twitter), routinely collected by organizations (public research institutes, national statistics offices), and generated and enriched by cultural institutions (GLAM: galleries, libraries, archives, museums) and news organizations into information that can be used in a number of ways both for academic and applied research, as well as for tackling real-world problems.

In this chapter, we discuss different aspects of APIs from the perspectives of social scientists who use APIs for data collection. We describe (1) the origin of APIs in software development, (2) conduct a survey of popular Web APIs by type, and (3) discuss issues with regard to the reliability, validity and representativeness of data retrieved from APIs. We close by pointing to future developments in this area.

API Origins

The Oxford English Dictionary defines an API as 'the interface between the operating system and an application program; the protocol to be observed by the writer of an application program designed to run under a particular operating system'. Beal (2016) similarly speaks of 'a set of routines, protocols, and tools for building software applications' and goes on to say that it 'specifies how software components should interact'. APIs have a long

history in computer programming. They stand in contrast to application binary interfaces (ABIs) which are based on binary, rather than interpreted, code. An API can make it easier to extend existing applications with new features by providing a framework that relies on pre-established functions. APIs also facilitate access to data, usually in order to provide new functions, which translates into greater utility of the service or application. Data access through Web APIs, in contrast to the broader understanding of an API as a programming framework that implements a set of standard behaviours (for languages such as C or Java), will be our focus in this chapter. APIs are often provided in the form of a package or library that includes specifications for data structures, functions and object classes. In the case of Web-based SOAP and REST services which underpin many popular social media services, a Web API is simply a specification of remote calls available to users of the API. SOAP and REST are the two most widely used exchange standards for doing this on the Web, with REST much more popular. REST is able to return results in different formats, most notably JSON, while SOAP supports only XML. JSON has proven particularly popular, being both able to represent complex data structures and being (relatively) readable to humans, while XML is comparatively less easy to learn. Web APIs based on SOAP or REST afford essential CRUD operations (create, read, update and delete) that underpin most data interactions. In other words, they are much narrower in their ability but also in their complexity, than are general purpose programming APIs.

Popular Web APIs

Web APIs are a fairly recent innovation in comparison to programming APIs in a broader sense. Still more recent is their adoption in social science research. The number of REST-based Web APIs listed by ProgrammableWeb, a global directory for APIs, surpassed 14,000 in 2015. The list of APIs tracked by ProgrammableWeb includes areas such as finance, science, education, mapping, games and messaging.

In addition to private companies, public institutions such as cultural heritage organizations and statistics offices increasingly offer APIs. In these organizations, APIs are usually part of larger open data strategies. Important providers include the UN[1], the WHO and the World Bank. Also, many federal

1 http://data.un.org/Host.aspx?Content=API. The other APIs mentioned below can be found through ProgrammableWeb's directory or the search engine of your choice, using 'API' and the organization's name as search terms.

and regional governments have implemented Web APIs. Many of these can be accessed through open data portals that have been implemented in the past years (e.g. data.gov, data.gov.uk, open-data.europa.eu, etc.). CKAN, the open-source system behind many open data portals, provides the framework for most of these APIs.

A handful of influential global news organizations such as the BBC, The New York Times, The Guardian, NPR, USA Today and ZEIT Online have also started to offer parts of their content through APIs. The biggest beneficiaries of this step seem to be the organizations themselves though, since the offerings of APIs make internal R&D efforts more efficient, help to further commercialize the news content, and facilitate external networks of open innovation (Aitamurto & Lewis 2013). Nevertheless, some content and (meta)data can also be used fruitfully for research.

Areas with high volumes of data creation, such as the (life) sciences, are particularly open to APIs to facilitate data exchange and enable new forms of information reuse. The rOpenSci collection offers a number of API libraries for the R programming language. Platforms for publishing and storing research data such as Dryad or figshare can be queried, as can be archives of scientific articles such as arXiv and PLoS. Countless sources of scientific data, but also cultural heritage material from Europeana, countless museums and the Internet Archive are available.

In addition to the above-mentioned APIs provided in different sectors, the APIs of large social media companies have been of great importance for the social sciences and humanities in the past years. Some of the most popular services, not only for research, include the APIs offered by Facebook, Twitter, Reddit and Instagram in the social network category, Google Maps and Yelp in the geolocation category, and Spotify and Soundcloud in the music category (Brennan 2015).

While these APIs 'provide new ways of sharing and participating, they also provide a means […] to achieve market dominance, as well as undermine privacy, data security, contextual integrity, user autonomy and freedom' (Bodle 2011: 320). Therefore, Web APIs cannot be seen solely as support software systems. Because they shape the organizations that provide them and format the rules under which external software developers can make use of them, APIs can be seen as powerful mediators in a datafied society (Ausserhofer [forthcoming]; Bucher 2013).

Through the establishment of social data APIs, social media companies seek to set up an open innovation ecosystem that draws application developers to the platform. The companies invest considerable resources to keep external programmers engaged with the API because they believe that this improves

their internal innovation capacity. Some companies even argue that by giving third parties such as researchers access to social trace data, they contribute to the public good. While this may be true for some cases, often this can be seen as measures for PR purposes, a form of 'open washing' (Villum 2014).

Social media platforms process billions of API requests annually. The central function of such requests is to provide derivative services or functionalities that increase the usefulness of the social media platform. However, as company policies change, a company's data management regime may become stricter. Twitter is an example of this approach. Initially offering broad access to data in the first years of its operation in order to encourage development of derivate services, such as software clients for unsupported platforms, the company reasserted its control by making access to data more restrictive in several successive steps over recent years (Puschmann & Burgess 2014). This shift took place alongside acquisitions (Tweetdeck, Gnip) and a number of derivate service providers going out of business, merging or changing their business model.

Strategic reasons are not the only motivators behind such changes. Facebook has greatly restricted access to user data through the API out of privacy concerns, as have other platforms. When dubious actors acquire large amounts of data that are clearly not used for the API's intended purpose, this often leads to a tightening of policies by the API's operators, if only because providing and sustaining the performance of an API is not trivial computationally. When Twitter greatly enhanced the ability of its search API, it was largely because the engineering feat of making historical Twitter data indexable was very difficult to resolve (Zhuang 2014). APIs, in other words, incur significant costs to businesses which may be invisible to users, who may be under the impression that data sits in the company archive like books on a shelf, ready to be picked up. Social media data, in addition to being available directly from platforms such as Facebook, Twitter and Instagram, are also stored, indexed and repackaged by dedicated social analytics providers such as Gnip (owned by Twitter) or Datasift (partnered with Facebook).

Reliability, Validity and Representativeness of API Data

We have so far argued that APIs are a useful data source for scientific research. There is, however, also reason for scepticism. Commercial platforms such as Facebook and Twitter do not provide their APIs as a service to researchers, but have other uses which inhibit reproducible sampling and frequently render data sets incomplete (González-Bailón et al. 2014;

Gerlitz & Rieder 2013). Capturing data from such platforms is also computationally resource-intensive, imposing limitations on research. Below, we pose a list of questions for scholars who engage in research that is based on data from an API. These questions are intended to highlight issues that typically arise in research designs that draw upon digital data sources. While these are very similar to standard social science tenants of research design, it is worthwhile to reiterate some of these issues in the context of data API.

'How purposeful is the sampling strategy?'

By purposeful we mean, 'What is the impact of technical constraints on sampling?' How do language implementations of the API, content fields for data and metadata, rate limitations and the availability of APIs for certain types of content all shape the sampling strategy? Consider this in the context of Twitter. The streaming and search APIs are well-supported in different languages. Content is provided in the form of tweets which are the preferred unit of analysis over discussion turns, topical frames or other, more conceptually-grounded units of analysis. Rate limitations for Twitter have become stricter over time, but are still quite lenient. Extracting and analysing Twitter data is easier and more popular than Facebook data, even though Facebook is far more popular than Twitter (Tufekci 2014).

'How clear is the sampling procedure?'

By clarity we mean, 'How clear is it what steps were undertaken to arrive at the sample?' Random stratified sampling is traditionally a pillar of empirical analysis, but this fails in many instances when sampling from social media sources. As Ruths and Pfeffer (2014) have pointed out, random Twitter samples are non-random in the sense that the server collecting the data and fluctuations in message volume both have an impact on the randomness of a sample. Since randomness is difficult to achieve for Twitter researchers who do not have access to a large volume (or ideally the entirety) of tweets, much sampling relies on snowball sampling or other convenience strategies. This is both bad for the reliability of results and raises complexity issues.

'How reliable is the sampling?'

By this we mean, 'Would the same query to the API at different times or from different people return similar results?' This is much more straightforward

in some APIs than in others, depending on the overall volume of content. APIs for archives, online news or public records will be much more reliable than commercial APIs for social media content.

'How valid is the operationalization undertaken in the research?'

By this we mean, 'Is it analytically sound to operationalize a data variable in a particular way?' Examples would be to characterize the number of followers a Twitter user has as a measure of her influence or the number of reads that an article receives as a measure of its popularity. The issue hardly ends there, but research that is based on digital data faces particularly complex questions of operationalization because in contrast to data sources such as surveys or interviews, the data comes in a highly suggestive pre-packaged form. Many social media metrics lead a dual life of meaning for users and platform providers with both parties influencing them deliberately or unintentionally.

'How representative is the sample of the population?'

By this we mean both, 'How well does the sample represent that platform from which it was drawn?' and, 'How well does the platform represent other platforms, users or sources of information?' In the case of Twitter and Facebook, it is a nontrivial problem to draw samples that are representative of either platform. Secondly, it is equally challenging to formulate valid assumptions about how well these samples represent groups of people more broadly.

'How reproducible is the research in total?'

By this we mean, 'How hard would it be to conduct similar research that tests the findings of the study?' In the case of exclusively big data samples, it is quite hard, just as it would be with smaller but historical samples of social media data. APIs as such do much to greatly improve reproducibility, by providing a common source of access to researchers. Proprietary data sets on CD-ROM or with strict access protection do much to effectively limit access, even when there is a general agreement that those who want to can gain access. On the other hand, hurdles exist both in relation to the computational feasibility of such research and to the technical skills required to make use of APIs.

APIs in the Future

We have sought to show that APIs are an increasingly relevant form of data access, both in academic and applied research, and for civil society more broadly. In addition to the Web APIs provided by large internet companies such as Facebook, Google and Twitter, APIs also proliferate among governments, scientific organizations and NGOs. They are likely to become a more widely used channel, assuming that more people are able to access them. Competency is the key issue here: Web APIs require basic programming knowledge to enable access. This requirement represents a significant hurdle, and one that cannot be overcome easily. The use of a programming language is what makes data access through an API efficient, and alternative, more intuitive forms of access incur costs to the data providers and are unlikely to scale efficiently. A second hurdle is the ability of data suppliers to control access and the ability to distinguish and, if needed, discriminate between users. As society becomes increasingly 'datafied', the relatively informal relationship between API providers and API users will need to be codified in a way that resembles the relationship between providers and users of other (public) services. Public services, such as libraries, and private utilities such as the telephone network point into the direction that this codified relationship may take. As APIs become more and more mundane outside of software development, and our reliance on them increases, the issue of their reliability too will become ever more important.

Acknowledgements

Both authors gratefully acknowledge the support of Volkswagen Foundation.

References

Aitamurto, Tanja & Seth C. Lewis. 2013. "Open Innovation in Digital Journalism: Examining the Impact of Open APIs at Four News Organizations." *New Media & Society* 15 (2): 314–31.

Ausserhofer, Julian. forthcoming. "Die Datenbank verdient die Hauptrolle: Bausteine einer Methodologie für Open Digital Humanities." In *Aufgehoben? Speicherorte, -diskurse und -medien von Literatur*, ed. Susanne Eichhorn, Bernhard Oberreither, Marina Rauchenbacher, Isabella Schwentner & Katharina Serles. Würzburg: Königshausen & Neumann.

Beal, Vangie. 2016. "API – Application Program Interface." Webopedia. Accessed 30 March 2016. www.webopedia.com/TERM/A/API.html.

Bodle, Robert. 2011. "Regimes of Sharing: Open APIs, Interoperability, and Facebook." *Information, Communication & Society* 14 (3): 320–37.

Brennan, Martin W. 2015. "Most Popular APIs Used at Hackathons." *ProgrammableWeb*. Accessed 10 April 2016. www.programmableweb.com/news/most-popular-apis-used-hackathons/elsewhere-web/2015/10/04.

Bucher, Taina. 2013. "Objects of Intense Feeling: The Case of the Twitter API." *Computational Culture* 3 (November). http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api.

Gerlitz, Carolin & Bernhard Rieder. 2013. "Mining One Percent of Twitter: Collections, Baselines, Sampling." *M/C Journal* 16 (2). www.journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620.

González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer & Yamir Moreno. 2014. "Assessing the Bias in Samples of Large Online Networks." *Social Networks* 38 (July): 16–27.

Puschmann, Cornelius & Jean Burgess. 2014. "The Politics of Twitter Data." In *Twitter and Society*, ed. Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt & Cornelius Puschmann, 43–54. Digital Formations 89. New York: Peter Lang.

Ruths, Derek & Jürgen Pfeffer. 2014. "Social Media for Large Studies of behaviour." *Science* 346 (6213): 1063–64.

Tufekci, Zeynep. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.* www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Villum, Christian. 2014. "'Open-Washing' – The Difference between Opening Your Data and Simply Making Them Available." Open Knowledge Blog. October 3. http://blog.okfn.org/2014/03/10/open-washing-the-difference-between-opening-your-data-and-simply-making-them-available/.

Zhuang, Yi. 2014. "Building a Complete Tweet Index." Twitter Blogs. November 18. https://blog.twitter.com/2014/building-a-complete-tweet-index.