A SAGE
White Paper

The Ecosystem of Technologies for Social Science Research

Daniela Duca, PhD
Product Manager, SAGE Publishing

Katie Metzler
Associate Vice President Product Innovation, SAGE Publishing

www.sagepublishing.com

SAGE
Publishing

Contents

Suggested citation: Duca, D., & Metzler, K. (2019). *The ecosystem of technologies for social science research* (White paper). London, UK: Sage. doi: 10.4135/wp191101

Acknowledgments

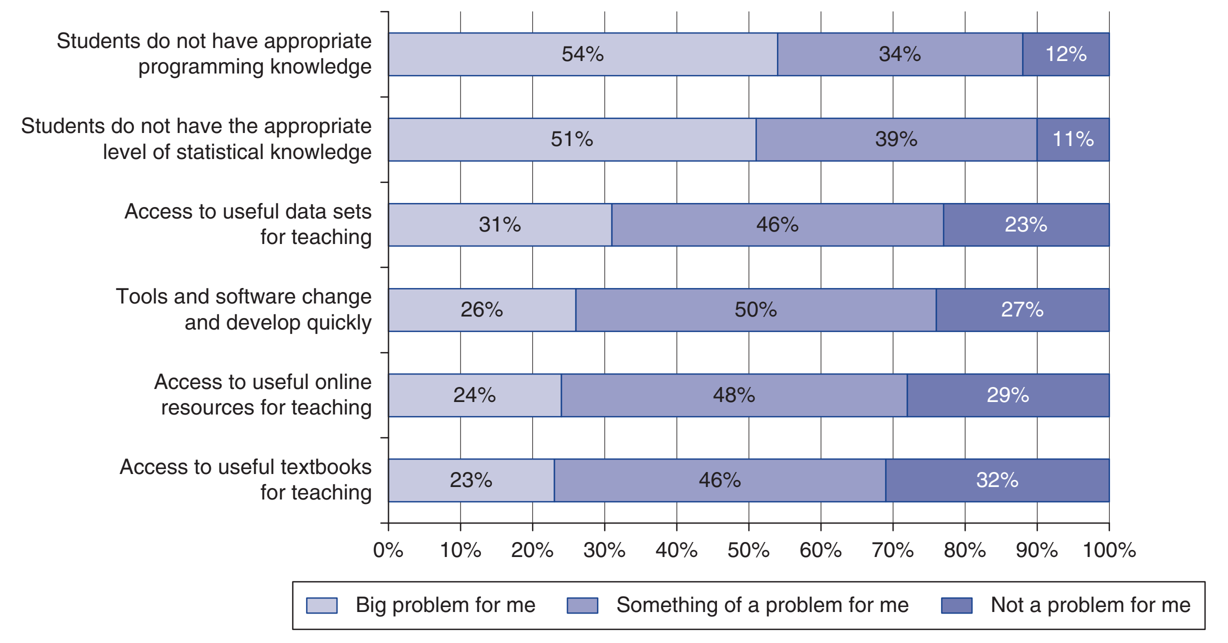
This paper would not have been possible without the work of Shulin Hu and Lily Davies, both master's students within the Digital Humanities programme at University College London, and Eve Kraicer, master's student in Gender Studies at the London School of Economics, whose work on this project involved data cleaning and looking at gender distribution. We also want to thank all the researchers, developers, and startup founders who endured and answered with patience and insight our endless questions.

Overview

The growth in digitally borne data, combined with increasingly accessible means of developing software, has resulted in a proliferation of software to support the research lifecycle. There is now a range of software and tools custom-built for very specific tasks, and the tools supporting common research methods have improved and expanded. Moreover, progress in machine learning models—especially around natural language processing, speech recognition, and the application of graph and network theory—has led to an explosion in new tools and has enabled social science researchers to borrow tools and technologies from other disciplines.

The availability and accessibility of new technologies for research is promising. But how can researchers and educators keep up with the changing landscape of tools and software? This challenge became apparent in a survey we conducted in 2016 with close to ten thousand researchers in the social sciences, who told us that the pace of change was an obstacle to teaching new methods to students (Figure 1; Metzler, Kim, Allum, & Denman, 2016). Moreover, the rapid evolution of tools for big data research in particular was seen as a barrier to researchers looking to move into the new and growing field of computational social science (Figure 2).

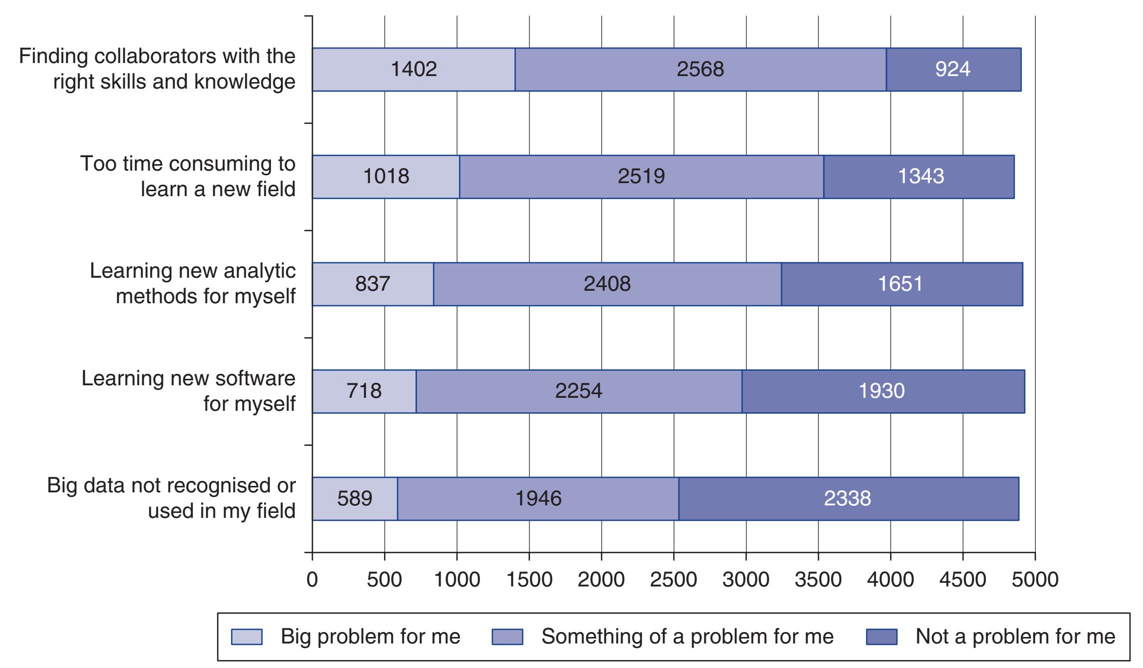
Figure 1 Challenges facing educators teaching big data (*n* ≅ 1212)



Through subsequent interviews with researchers and students, we gained an understanding of the challenges facing social scientists who want to prepare themselves for a more data-intensive future in research. In response to this and the 2016 survey results, SAGE Publishing launched the SAGE Ocean initiative,[1] with the mission to support social science by equipping social scientists with the skills, tools, and resources they need to work with big data and new technology.

Over a period of 10 months, SAGE Ocean reviewed 418 tools and software packages used by social science researchers, which we sourced from research papers, tools directories,

Figure 2 Challenges facing those wishing to enter into big data research (*n* ≅ 4894)



company databases like Crunchbase, Wikipedia, researcher and lab blogs, and other websites. We were interested to find out more about:

- How researchers discover tools
- How researchers decide which tools to adopt for their research
- How tool developers fund and maintain their tools
- How developers are recognised for their efforts
- What role software development plays within the academic ecosystem

We explored the various features of these tools and technologies, as well as the key people and organisations that supported their development. We conducted detailed analyses of tools for text annotation, recruiting and surveying research participants, and collecting and analysing social media data.

From this work, SAGE Ocean has built a Research Tools Directory[2] to help researchers navigate the landscape of tools and software, and launched a Concept Grant scheme[3] to support the builders of tools and software for social science research. We will continue this research and share our findings as we expand our list

Clarification: Throughout this paper, we refer to a tool, technology, software service, or package interchangeably. We acknowledge these terms mean different things, and our list of 418 includes all of them: cloud and locally installed apps, software applications, packages and code, licensed and open source—essentially any piece of code, whether neatly packaged within an interface or delivered raw on GitHub and used by academic researchers.

of tools for research. We believe this insight and knowledge is vital for a future in which more research is carried out with the help of technology, and in which researchers may increasingly become tool builders themselves.

What We Learned About Technologies for Social Science Research

Our research identified a wide variety of tools and software packages used by social science researchers. Software is important for researchers in this field; many develop—or hire someone to develop—tools for their projects, but they are equally likely to use commercial tools or other open source tools. Among the 418 tools we reviewed, close to 50% are United States based, just over 50% are free to use for researchers and can be developed by private (50%), big tech (5%), public sector, or individual projects (45%). Whilst many commercial tools are available, we note that the more innovative ones are coming out of academia.

Our sample is skewed towards tools for social media analysis, text labelling and annotation, and surveying and recruiting participants, with a growing list in the text mining cluster. Overall, we note a diversity in the types of features the tools offer, growing concomitantly with support from various stakeholders.

What We Learned About the Individuals and Organisations Developing and Supporting Technologies for Social Science Research

Only 10% of the key people involved in designing and developing software tools used by social science researchers are women. It is critical that there be diversity among the teams developing tools and the organisations supporting them, and we believe this should be a priority in the continued sharing and development of software tools.

Other challenges face the creators and developers of tools: There is a lack of peer review, and only a few academics follow the citation principles available for software. More work needs to be done to facilitate code sharing and the maintenance and financial sustainability of these tools. All of these challenges determine the use of tools and their future in general, especially for those coming out of academia.

When it comes to supporting the development of research tools, a growing number of communities, organisations, and consortia offer support, guidance, and training. Among these are the Software Sustainability Institute, the discipline-specific Digital Methods Initiative, NUMFocus and Pelagios Commons, and the regional NeCTAR and CESSDA. Where tools have applicability or a primary focus in the business intelligence world, we find top venture capitalists involved, such as Sequoia and Index Ventures. With the exception of Prolific, only a few startups coming out of university incubators target social science researchers.

Methodology

We reviewed 418 software tools we know are used by social science researchers. This is by no means a comprehensive list, and we intend to keep adding to it. We acknowledge that our sample is biased towards English-speaking countries, and more work is needed to identify the tools that exist outside of that space. Our main selection criteria for identifying software packages, tools, and apps related to their purpose and the user; the tool had to be useful in the research process—for example, by enabling data collection, analysis, visualisation, or running experiments on or offline. We also required evidence that the tool was used by social science researchers, demonstrated through references in research papers or recommendations from researchers or researchers' blogs.

While expanding the list to 418 tools, we looked in particular at tools for surveying, social media research, and text annotation and analysis. This paper describes in more detail the trends we identified in these clusters of tools.

We researched the following information where it was easily available:

- The team or key people involved in the tools' development
- Founding date
- Headquarters
- Description and history
- Features and integrations
- Pricing
- Funders and partner organisations or supporters
- Available papers

We used these key sources:

- Project/company websites
- Conference websites and archives
- Prominent labs and initiatives, such as the Social Media Lab at Ryerson University,[4] the Digital Methods Initiative,[5] Lazer Lab,[6] and the Public Data Lab[7]
- Other tool directories and lists, including Dirt Directory,[8] SourceForge,[9] GitHub,[10] LabWorm,[11] and data from the 2016 survey on Innovation in Scholarly Communication (Bosman & Kramer, 2016)
- Academic papers referencing or describing the tools
- Wikipedia[12]
- Crunchbase[13]

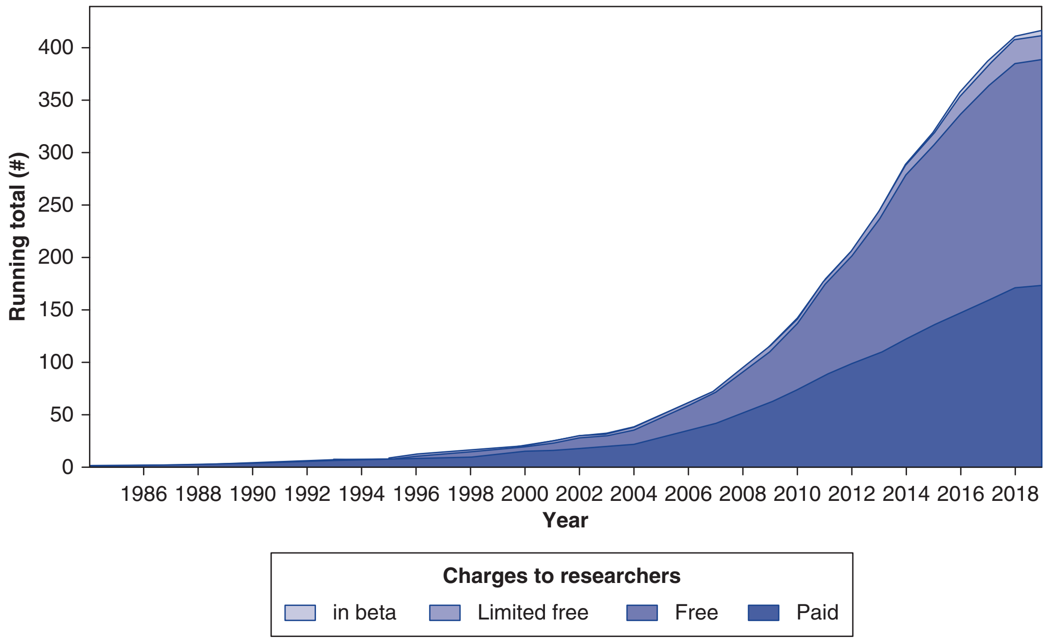
Analysis

Tools and Technologies: Trends Over Time

Software and computational tools are critical for social science research. In the most recent Software Sustainability Institute Survey (Hettrick, 2019) conducted at the University of Southampton in the UK, 83% of respondents working in the social sciences said they make use of research software. Just over 60% claim this software is vital to their research, and at least a fifth develop their own software with or without grant funding.[14] SAGE Ocean recently ran a version of this survey and found that just under 90% of the 148 respondents (most of whom are based out of social science and humanities faculties) use research software, and more than half believe it is important or very important to their work.[15] Significantly, over a fifth have hired someone specifically to develop software.

Our data suggests the number of research tools available has grown rapidly (see Figure 3), particularly in the past 10 years. This could be a result of researchers expanding their range of skills and DIY software development becoming more accessible thanks to advances in computer science. There is a similar number of paid and free tools available (197 vs. 215) in our dataset, but the number of free tools has seen a marginally steeper increase in recent years. This may be a combined effect of the push for open source and a greater number of individuals developing their own tools. Similarly, it is becoming increasingly common to see software available for free on a limited basis—for example, for a limited time, with limited features, or for certain customer groups (most frequently academics).

Figure 3 Volume of tools over time by charge, 1984–2019



In terms of global distribution (see Figure 4), almost half the tools are based in the United States (209), followed by the UK (69), Germany (17), and Canada (13).

The bubble graph (Figure 5) illustrates the distribution of tools in our sample dataset by competitive cluster. To create the clusters, we used keywords to classify the tools that are most closely related or that would be used as alternatives to each other by researchers. Whilst the sample covers a variety of tools for social science research, it is biased towards social media analysis and annotation, as these are the areas we investigated most comprehensively.

Characteristics of Tools by Cluster

1. Tools for Surveying and Sourcing Participants

Surveys are probably the most common method for data collection in the social sciences. Researchers used to carry clipboards, but now there are more than 50 different software tools for surveying. The timeline in Figure 6 shows a select number of survey tools and illustrates how varied this space is: There are free and open source tools, tools developed specifically for

Figure 4 Number of tools by country

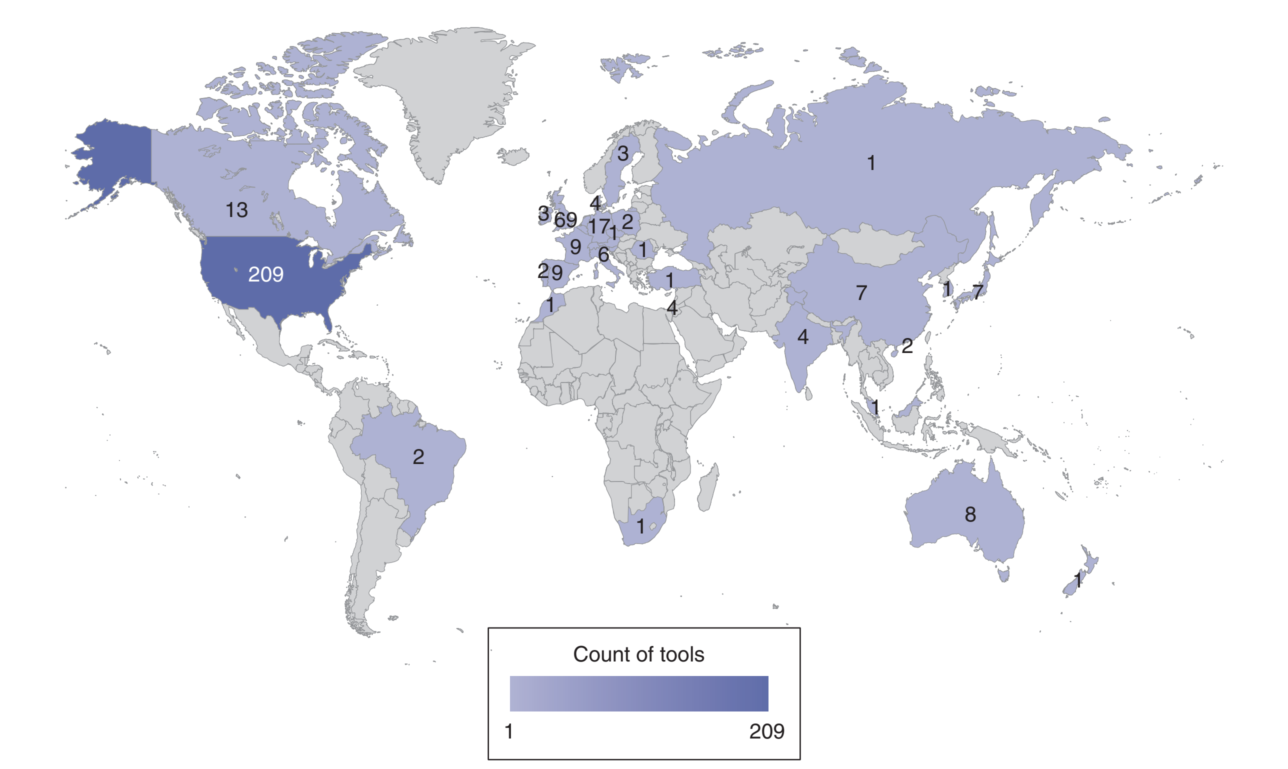
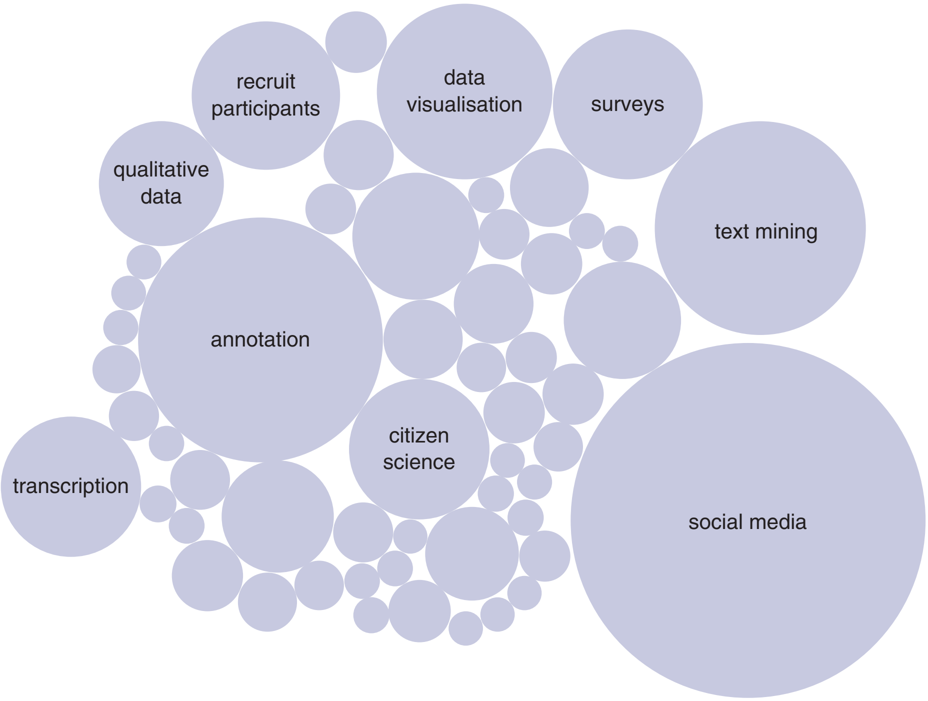
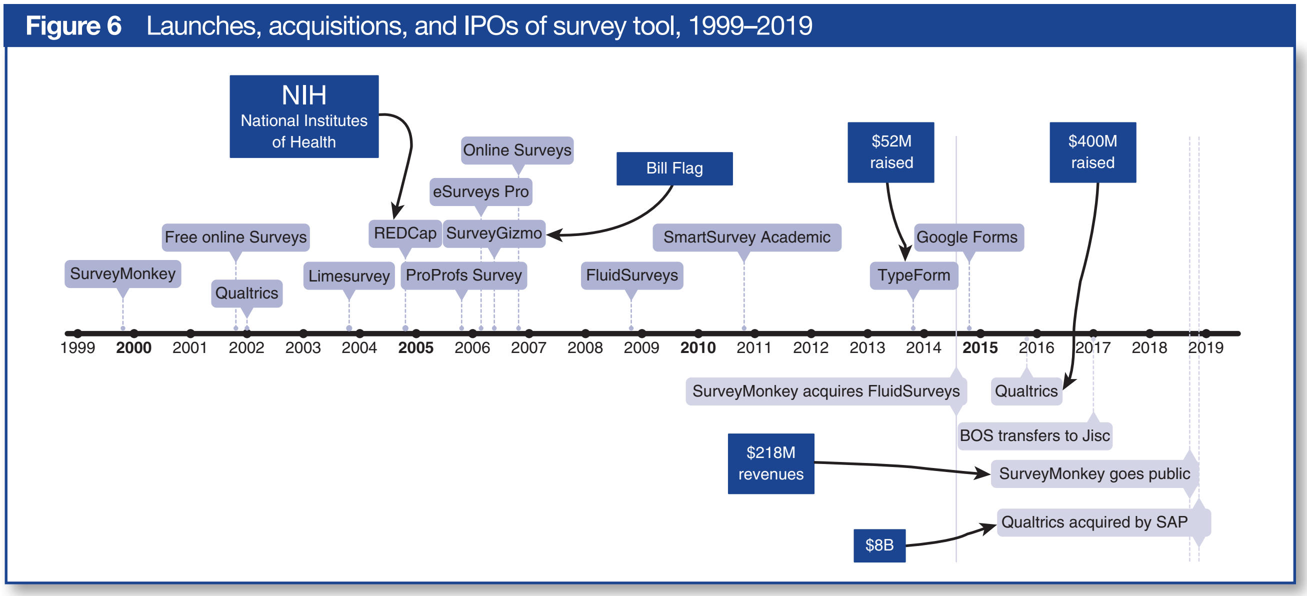


Figure 5 Distribution by cluster



academics (REDCap, Bristol Online Surveys, SmartSurvey Academic), some that developed quickly and raised a lot of money (TypeForm), and others that have been less successful. Some acquired competitors and went public (SurveyMonkey), and some had backing from angel investors (SurveyGizmo). This part of the market is still shifting, with old and new entrants, acquisitions, and fundraising, and yet more will come, with machine learning having the potential to transform the way we conduct surveys.

Figure 6 Launches, acquisitions, and IPOs of survey tool, 1999–2019

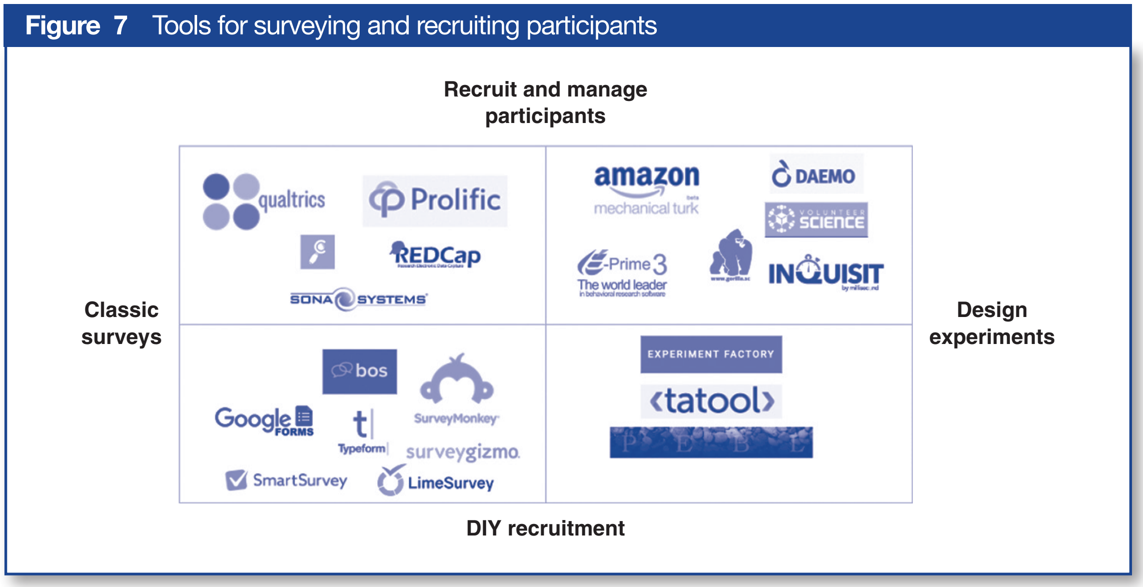


Of the surveying tools developed specifically for academics, the most popular one is REDCap. REDCap is an open source survey and client relationship management tool (CRM) developed initially for clinical research and funded by the National Institutes of Health (NIH). The tool grew a large supporting community and is now used across disciplines by researchers who are comfortable with coding and concerned with privacy, i.e., where person-identifiable data is stored.

Qualtrics may appear to be a recent entrant (their first press release was in 2012), but they have been earning revenue since 2002, and by late 2016 they had raised almost $400M to develop an advanced surveying platform for research and business. Qualtrics is popular among social science researchers and is considered an innovative company (Qualtrics, 2019); it can source respondents and interface with almost anything, including games and custom-built tools. At the end of 2018 it was acquired by a business solutions company for $8 billion (SAP, 2019), after having turned down another offer for $500k a few years earlier, according to a TechCrunch story (Andersen, 2013).

Whilst with most existing surveying tools researchers are responsible for sourcing their own participants, there are now a number of tools that can help with recruitment (Figure 7). The most commonly used is Amazon's Mechanical Turk, though arguably the best one for academic research is Prolific. Prolific (also known as Prolific Academic) was set up in 2014 by then-PhD student Ekaterina Damer, who was herself struggling with participant recruitment. Through the Oxford Innovation Incubator, Damer and her team developed a minimum viable product in four weeks, and in less than a year they grew both the participant pool and the researcher (user) community rapidly. Prolific integrates with a number of surveying and online experiment platforms (Lumsden, 2019) and, at the time of writing, has 45,000 participants from across the globe.

Figure 7 Tools for surveying and recruiting participants



2. Annotation, Labelling, and Coding Text

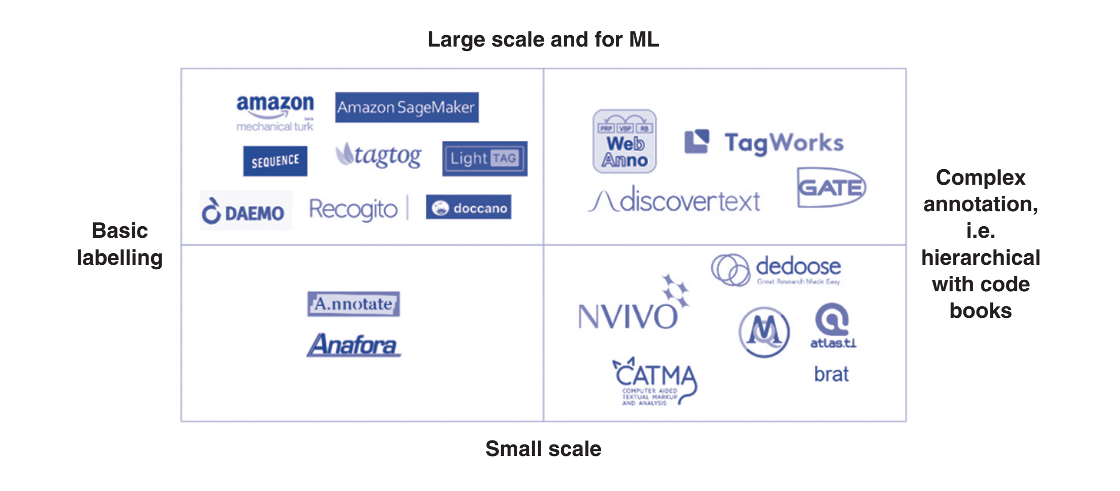
We found 48 tools (60 if including CAQDAS[16]) that offer labelling or annotation services for text. Ten (20%) of these have not been updated for at least a year. Thirty-three (69%) are free to use, and pricing for the remaining 15 (31%) ranges from approximately $60 to $10,000 for a 1-year project or to tag 2,500 documents with three annotators.[17]

The oldest annotation tools (besides the CAQDAS) are GATE and Open Calais. GATE is open source and very popular with researchers using complex coding schemas and larger text corpora, while Open Calais is for active work—labelling as you write. Given the proliferation of machine learning, especially around text mining, the need to annotate larger corpora to serve as training datasets has grown. This has driven a boom in new text annotation tools, each with unique selling points and improved user experience.

Broadly speaking, there are two types of tools in this space: the recent tools that support semiautomated tagging (i.e., if you label a word or a phrase, it will label it in the remainder of your documents, automating a part of the task) and the classic annotation tools that support custom coding schemas. Tagtog, LighTag, Doccano, and Dataturks are relatively new semiautomated tagging tools. Integrated with Mechanical Turk, the most recent entrant to this space is Amazon's SageMaker, which promises to beat most other tools in rapidly crowdsourcing labels to train your algorithm. For researchers looking for a more complex set of labels and who are building a hierarchical coding schema, TagWorks[18] (beta) was developed to support these functions. Figure 8 maps a select number of the tools for annotating text based on complexity of labelling and how well they scale.

A range of open source, free to use, publicly funded tools provide similar manual annotation services, with a greater focus on inter-annotator agreements and developing robust code books. They rely on community input, and although they have been developed within old infrastructure, they are actively used. The oldest, GATE, can score up to thirty thousand downloads a year. Brat has 218 forks on GitHub and more than 200 members within their

Figure 8 Tools for annotating text



contributing community, and along with TAMS Analyzer, we estimate they are cited more often than the other annotation tools (between 130 and 630, vs. less than 100 for the rest[19]). GATE, Brat, and TAMS Analyzer appear in the first two pages in a Google search for “text annotation,” and they are often recommended on Quora and ResearchGate as the best options for a social science labelling task.

We are now interested in exploring text mining and large-scale text-analysis tools, including how they compare with qualitative analysis and where there might be gaps within this space. We used an adaptation of Aaron Tay’s (2019) performance vs. skill level quadrant to map these tools in Figure 9. Most of the tools we identified for serious large-scale text analysis require a high skill level; only three require a low level of coding and are intuitive enough for a non-programmer; however, performance and outputs are still basic. For example, the following tools are capable of basic automated text analysis, such as bag-of-words, basic sentiment analysis, or the pre-cleaning of the data (for example removing the stop words):

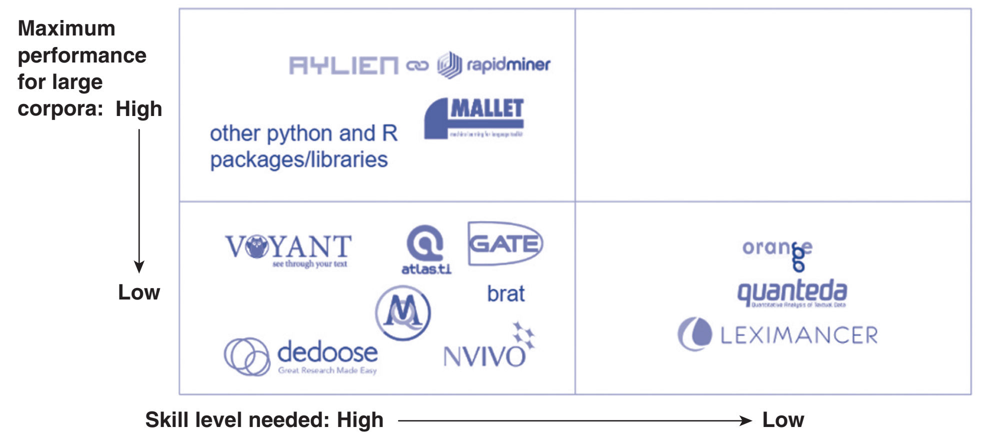
- Orange data miner (developed at Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia, together with the open source community)
- Quanteda (currently in beta, under development by Ken Benoit with support from SAGE)
- LEXIMANCER (one of the older commercial text-mining tools, developed in Australia)

We are only at the beginning of this project and will share our results in our SAGE Ocean blog.[20]

3. Tools for Social Media Research

Over 30% of the world’s population uses social media. We spend a vast number of hours attached to our devices; every minute in the United States, 2.1 million snaps are created on Snapchat and one million people log in to Facebook (Clement, 2019). This use generates an enormous amount of data about our behavior, opinions, and preferences. Some of this data can be collected and explored, not only for research purposes but also for insights that can help companies sell to or influence people. For example, even fast food chains are no longer

Figure 9 Tools for text mining



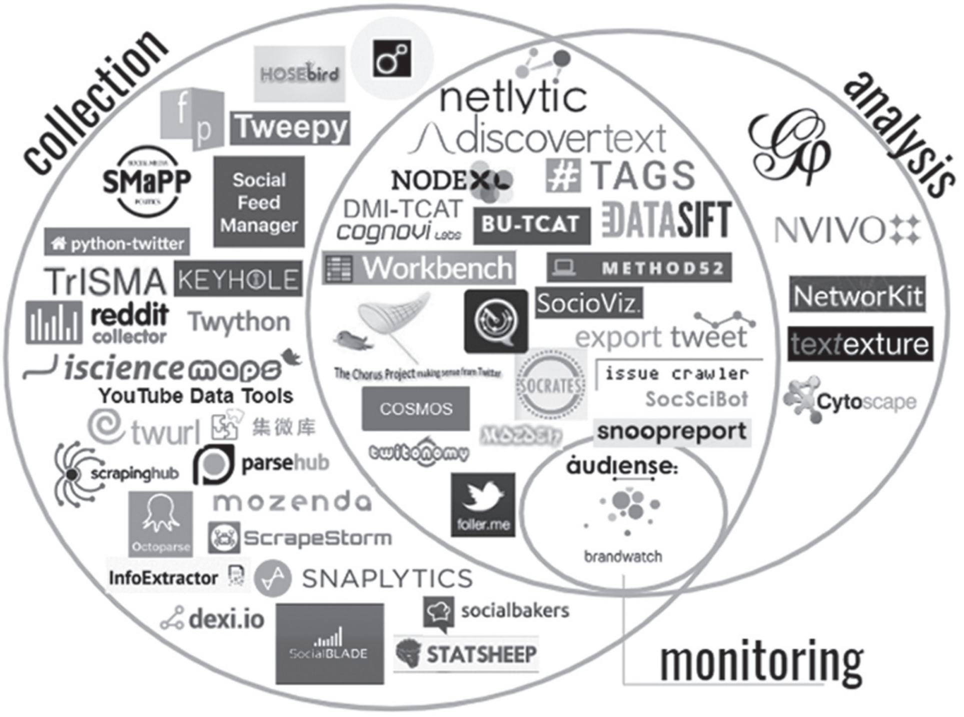
waiting for their customers to complain but are parsing social media posts to pinpoint the location of restaurants that are likely to have caused food-borne disease (Wiggers, 2019).

Given that there is interest in this data from those conducting both academic and consumer research, a great quantity of tools (full apps, wrappers, or pieces of code) is available to explore social media data. At the time of writing, we have found and reviewed 104 tools that we know researchers have used (see Figure 10). Close to 70% of the tools (70 out of 104) interface with Twitter, even though this is not considered the most popular social media platform, with far fewer monthly users than Facebook or Instagram.[21] We believe this is because it is much easier to access data from Twitter through their application programme interface (API) than from other platforms.

Generally, social media platforms have a policy against using their API for research purposes. This is true for LinkedIn and Facebook. However, the spread of misinformation has driven Facebook to enable access to their data for research purposes. In 2018, together with a number of partners, Facebook set up Social Science One,[22] a nonprofit organisation led by a group of senior researchers from around the world who selected and gave grants to 12 teams of researchers to work with Facebook data. A couple of years earlier, and now on their second round of research programmes, LinkedIn set up the Economic Graph Research Program[23] to enable researchers to access and work with their data.

More than half (56) of the social media tools we identified are either free apps or open source packages on GitHub, and 10 have limited free functionality. One tool that particularly caught our attention is DocNow, developed by Shift Design and a group of researchers at the University of Maryland and the University of Virginia, and funded by the Andrew W. Mellon Foundation. DocNow[24] provides a collection of tools that respond to the role of social media in reporting historically significant events, or “documenting the now.” Prioritising ethical collection, use, and preservation of social media data in academic research, their crowd-sourced catalog of tweet ID datasets and “rehydration” tool[25] allow access to historical tweets, whilst ensuring they are used in line with the creator’s decisions.

Figure 10 Tools for working with social media data



We published a series of blog posts with more insights from our research:

- A review of the current landscape of social media tools and the labs involved in developing them (Davies, 2019)
- How researchers are using LinkedIn data (Duca, 2019)
- How researchers are using Weibo data (Hu, 2019)
- Tips on collecting social media data for research (Radford, 2019)
- Information about the teams given access to Facebook data through Social Science One (SAGE Ocean, 2019)

Challenges Facing Tool Creators and Users

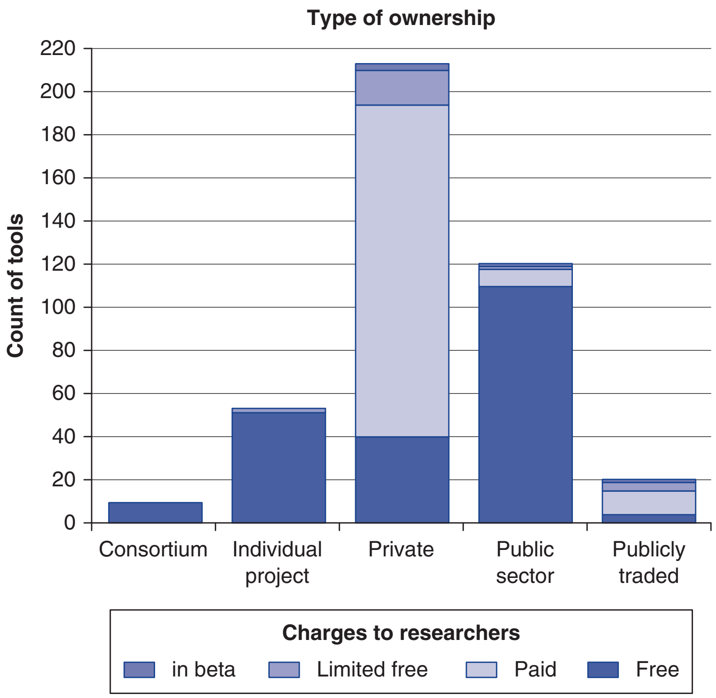
We were interested in learning about the people creating and developing these tools, in order to determine patterns that could drive the tools' success, as well as to identify any emerging trends or challenges facing tool creators. We also wanted to understand how SAGE Ocean could offer support in these areas and promote best practices. We were curious to know whether there is more appetite for tools developed by researchers and/or within universities, as well as how these tools compare to commercial ones. We collected data on affiliation (private or public sector) and diversity of the leadership team.

Ownership Type

Out of 417 tools for which we could confidently attribute an ownership type, 214 (51%) are developed by private companies, startups, or other medium enterprises (see Figure 11). Anything developed or acquired by Google, Microsoft, Amazon, or other publicly traded companies falls into the “publicly traded” category, which amounts to 20 tools (5%). Some tools coming out of publicly traded or private companies charge a fee, but 44 (11% of total, or close to 20% of privately owned and publicly traded) are free to use for researchers.

Interestingly, 9 tools (2%)—a small but growing number—were created or are maintained by consortia and are free to use for researchers. Under consortia, we included groups of organisations that decide to work together to actively maintain the tools.

Figure 11 Distribution of tools by ownership and charge



Given breakthroughs in hardware, the accessibility of software development, and the changing needs of researchers to work with large and more complex data, the number of tools being developed by individuals or as side projects has grown rapidly in the last 10 years, now making up 13% of the tools we surveyed (see Figure 12). This figure will no doubt rise sharply when it becomes commonplace to share software and code. We observed that, most often, the tools that are developed as a hobby or individual side-project only survive when the owner takes full responsibility for the tool and continues to use and maintain it.

Just 120 tools (30%) were developed by teams or labs within universities or other public-sector organisations. These are usually funded by grants for research and sometimes are a result of cross-disciplinary collaborations.

Diversity of Leadership Teams

We were curious about the diversity of the teams and individuals behind the tools, so we collected data on the tools’ founders or founding teams. Based on 271 tools (or just over 65% of the entire list) that had information available, only 17 (6%) were led by women, and 19 (7%) had mixed-gender teams (Figure 13). Of the 410 people involved in creating these 271

Figure 12 Tools totals by ownership type, 1984–2019

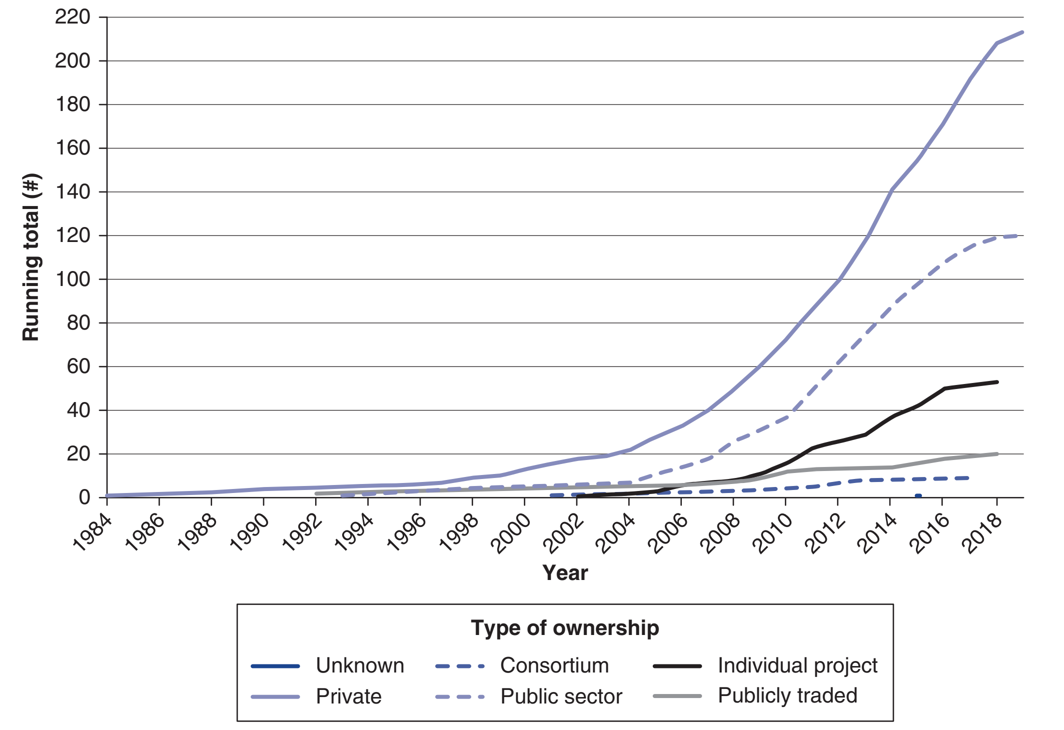
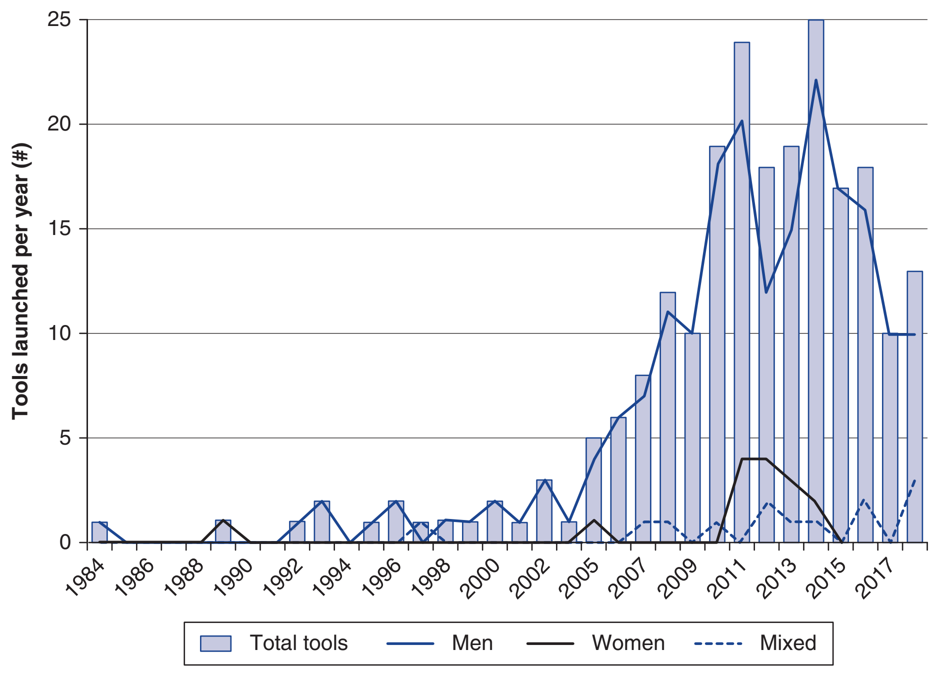


Figure 13 Tool leadership by gender, 1984–2019



tools, 38 are women (10%) and 372 men (90%). This data was collected manually from the tools' websites, Crunchbase, GitHub, and any papers describing the tools. The gender was verified manually based on individual profiles.

Challenges

The variety of tools, their diverse sources and developers, and their unclear status in the academic ecosystem means multiple challenges face the people building these tools, as well as the researchers trying to use them. Even when a tool exists that is to some extent suitable for a certain aspect of a social science research project, it can be difficult and time-consuming to find it; many researchers don't know where to look, and some research papers don't provide details of the technical packages used. Many academics decide it's quicker to develop their own tool for a research project, particularly when they have the required skills, a grant to cover the costs and time, and the flexibility to hire a graduate student to support. Without any widely accepted and widely implemented process to credit software reuse, many of these new tools will not be documented or made discoverable, with other research commitments taking priority. Thus any researchers looking for a similar tool in the future will not be able to find those that have already been created, and the cycle continues—another tool is developed, money, time, and effort will be invested again, but few others will ever find it or use it. As such, the development of these tools comes with challenges—both for the researchers using the tools, as well as those developing them. We discuss here four of the most immediate challenges and their impact on the ecosystem.

1. Poor Citation Practices

The software tools and packages we reviewed are still relatively hard to search for in academic papers, even when their names are sufficiently unique. Some of the tools ask for a specific paper to be cited when a package is used, and this makes it easier for the researcher using the tool as well as for the original developers, who would get credit as a result. GATE, for example, is very clear about which paper to cite for which package.[26]

In other instances, researchers reference the software tools and packages they use either in the methodology section or in footnotes. This is not ideal, as it makes it hard for others to find research based on specific tools, and for the developers to get credit, but it's still better than no reference. Even if they do mention the tools, researchers often find themselves in a situation where they have to remove the mention, because "it is too technical."

A group of researchers from across the world has worked, as part of the FORCE11 Software Citation Working Group,[27] to define software citation principles, which we would encourage you to read (Smith, 2016). The Software Sustainability Institute also provides easy-to-use guidance on how to cite software (Jackson, 2019).

2. Difficulty Navigating the Ecosystem

The sample we reviewed is not comprehensive, but it is already too big and too difficult to navigate. Researchers often rely on their peers, other academic papers, or their institution or department's immediate license to find and use a tool for their research. Alternatively, and if their skills permit, they will search for and use a package someone else developed for the programming language they are working in.

There are a number of directories out there that ease parts of this search process. For example, we created one[28] with a select number of tools that you can browse by theme or topic. The Dirt Directory[29] is a database of tools supported by the Andrew W. Mellon Foundation. It was specifically developed for digital humanities based on the TaDiRaH taxonomy,[30] but it is well known and useful for the social sciences as well. The Digital Methods Initiative also maintains a list of tools,[31] mostly developed by their collaborators. The tools list from the Digital Humanities Lab[32] combines licensed and open source packages. Other useful databases are SourceForge[33] for open source tools and LabWorm,[34] a community-curated list of tools that started in bioinformatics and is now expanding beyond that discipline.

3. Sustainability and Open Source

Most often, overwhelmed by the range of software tools available, unsure of their quality and often of what is behind the computation, many researchers resort to developing their own tools, tailored for very specific use cases. Thus the number of tools and packages continues to grow, but their lifetimes are shortening. In rare cases, researchers or developers manage to maintain their tools as a side project (see Laurence Anthony's tools[35]) and sometimes build an entire community while keeping the tool free and open source (see Gephi[36]).

It is becoming clear that open source and financially sustainable are not necessarily exclusive. Increasingly, researchers are finding ways of developing tools that are both open source and capable of making revenue. The most successful in this space is RapidMiner (2015), which follows an open core model, licensing parts of the code to enable scaling to enterprise levels. There are pros and cons to all the existing open source operational models, and there is no golden straitjacket that will solve the sustainability challenge. However, there are a few options that have already been successfully implemented. Joseph Jacks has a great overview in his blog on open source business models (2018).

Two key organisations are supporting the teams that are either thinking about or already working on different models for their open-source tools. The Software Sustainability Institute provides guidance and support around open source code for research—anything from tips on building a community to continuous integration.[37] The Apereo Foundation[38] is a membership organisation that offers guidance and incubation opportunities for teams working on open source technologies for learning and research, to be used within higher-education institutions.

4. Lack of Peer-Review for Tools

The proliferation of tools and software packages, combined with the difficulty of navigating this ecosystem and poor citation practices, lends itself to a need for peer-review of these tools. We believe there should be a standard model with a set of minimum requirements (in addition to citations in academic papers) that could be tested or evaluated. This would reduce the time some researchers already spend on comparing different tools and packages before deciding which one to use for their research.

For example, any new and existing tools should be tested on some standard datasets. We have already seen this happening with summarisation tools; the Connected Experience Lab[39] based at Cornell University, with support from Oath and Google, developed a dataset of 1.3 million news articles to be used for testing, training, and evaluating summarisation algorithms (Grusky, Naaman, & Artzi, 2018).

5. The Challenges of Big Technology

Big technology companies are posing a number of challenges for researchers in the social sciences and the tools they use. These companies control the data: how it is released and who can use it. This means that any tools that rely on Facebook, Twitter, LinkedIn, and other APIs have to keep up to date with the legalese and make sure they update their own features as new API versions are released. This puts the onus on open-source alternatives and precludes researchers from accessing the data. Similarly, big tech companies are increasingly taking control of the newer tools appearing on the market, via acquisitions (GNIP by Twitter, bloomsbury.ai by Facebook) which limits further access for academic research. Finally, these companies attract high-caliber social science researchers to work for them. This is where the research purpose and the incentives suddenly shift towards fast experimentation with immediate impact.

There is a silver lining. We've seen initiatives that look to archive individuals' digital footprints with proper informed consent for research. An increasing number of startups consider the value of the data they collect for research and actively collaborate with academics. And regulations around the ownership and use of personal data are being strengthened, with more consideration for the current state that big tech companies have brought about.

Who Supports Tools and Technologies for Social Science Research?

Over one hundred of the tools in our list were developed within university research projects, by dedicated communities, consortia, or other nonprofit organisations. We looked further into these organisations and communities to understand what they do and how they support these technologies:

- NumFOCUS[40] is a nonprofit organisation that helps teams develop open-source software by providing advice around tax relief and sustainability programmes. It has supported tools like rOpenSci.[41]
- The Software Sustainability Institute[42] is a non-profit organisation based in the UK that provides guidance and support to developing open source software.
- Pelagios Commons[43] is a discipline-specific consortium (humanities) that develops and supports infrastructure for Linked Open Geodata. One tool in their portfolio is Recogito,[44] for collaborative annotation.
- The Foundation for Open Access Statistics[45] is a discipline-specific nonprofit that offers nonmonetary support for projects that align with its mission: free open-source tools and reproducible research in statistics, like the Shiny package for R.[46]
- The Open Knowledge Foundation[47] is building a community around open knowledge and advocating the release of data and information. It supports projects like ckan.[48]
- The Digital Methods Initiative,[49] directed by Richard Rogers, professor of New Media and Digital Culture at the University of Amsterdam, collaborates with research groups across the world and other nongovernmental organisations to develop tools and carry out studies on digitally borne data. The initiative has also collaborated with the Amsterdam-based govcom.org, a foundation dedicated to creating and hosting political tools on the Web. DMI-TCAT[50] is among their portfolio.
- The Corporation for Digital Scholarship[51] is a charitable organisation that supports the open-source tools Omeka and Zotero.
- The Social Media Research Foundation[52] is a charitable organisation, or "a group of researchers who create tools, generate and host data, and support open scholarship related to social media," rhR developed NodeXL.

A recent article published by The Institute of Electrical and Electronics Engineers (IEEE) describes a few of these communities in more detail, also proposing a conceptualisation of the United States equivalent of the Software Sustainability Institute (Katz et al., 2019).

We found a number of regional programmes and consortia that support infrastructure for the social sciences or broader academic use:

- Nectar[53] in Australia (National eResearch Collaboration Tools and Resources project) was established in 2011. The project "provides an online infrastructure that supports researchers to connect with colleagues in Australia and around the world, allowing them to collaborate and share ideas and research outcomes, which will ultimately contribute to our collective knowledge and make a significant impact on our society."
- The Center for Open Science (COS)[54] is a United States nonprofit organisation established in 2013 to promote open science, globally. Brian Nosek, one of the cofounders, initiated the Psychology Reproducibility Project, coordinating more than 200 researchers to reproduce psychology papers. COS are also behind Open Science Framework, the open source academic collaboration platform.
- In Europe there are a number of infrastructure-driven collaborations. CLARIN[55] supports language-related infrastructure (services, tools, and data) for humanities and social sciences. DARIAH[56] is another research infrastructure consortium in Europe that supports services, tools, data, and training materials across the arts and humanities. CESSDA[57] is a research infrastructure consortium for tools, services, and training in the social sciences, with a strong focus on research data.

When looking into financial support, we found no standard or most popular way that the tools in this space are funded. As Yvonne Campfens (2019) also notes in her analysis of innovation in scholarly communication, support can come from outside of the publishing space and academia. Given that in many cases it is businesses, not academics, who are the target customers for tools for surveying, annotation, text mining, and social media data collection and analysis, it is not unusual to see the involvement of top venture capitalists.

Generally, we found that the tools follow one or a combination of the following strategies, unless of course, they invest their own time and money:

- Grants
- Crowdfunding (only a few)
- Venture capital (VC) and private investors

Grants are offered as a one-off or over multiple years and, based on this sample, an individual tool, company, or team can get anything from $10,000 to $6 million. The Institute of Museum and Library Services and the National Endowment for the Humanities are popular supporters. Some funders are establishing programmes directed at startups and promoting the development of sustainable and commercial digital services, such as the Knight Prototype Grant and the NEH Digital Humanities Startup Grant. In the United States, the Knight Foundation, Sloan, Omidyar Network, Mellon, and Mozilla foundations are very active. In Europe, the H2020 (and previously FP7) infrastructure funding have supported the development of services for research. The South African Shuttleworth Foundation has also been involved in supporting some of the tools.

With the SAGE Concept Grants,[58] which we launched in 2018, we have now supported four different tools for social science researchers, including Quanteda for text mining and Text Wash for redacting and anonymising textual data. Focusing on the entire scholarly communications cycle, the Digital Science Catalyst Grants[59] have funded more than a dozen startups in the academic space.

Crowdfunding, although more popular for science projects, has been used for a few tools in the social sciences. Two of these are Open Collective[60] and Superior Ideas.[61]

The Knight Enterprise Fund is a VC-type programme established by the Knight Foundation to support early stage startups in news media and publishing. They have been active for about 5 years (estimate at least $20M size); portfolio companies include Authorea and had 3 exits.

VC funding is popular among teams that intend to commercialise or make a profit from tools that cover a wider market than the academic sector. A startup can get anything from $250,000 to $400 million over one, two, or three series of investment. Some companies have been able to attract big-name investors like Sequoia, Index Ventures, the Northwest Fund, and OpenOcean. There are a number of investors that focus on EdTech, such as Edvinca and Emerge Education. However, most VCs are not so specific about the sector they invest in and prefer to focus on "digital" as an umbrella term for all software services regardless of industry.

Examples of VC-funded tools

Riffyn: $10M VC

Import.io: $22M VC

paralleldots: $2M VC

RapidMiner: $38M VC

SciStarter: $1M mix

qualtrics: $400M VC acquired

Tableau: $45M VC acquired

While a number of universities mostly in the United States and the UK, and a few around the world, have incubators and some investment capacity, they don't normally see applications from many software for social science research. Driven by market demand and budget size, large university incubators such as Imperial Innovations (£300M, with at least 91 investments), SETSquared (£1B+), Velocity (157 investments), and Creative Destruction Lab (194 investments)[62] fund and accelerate startups with underlying patents pending, or those focusing on biotech and industrial technologies.

Conclusion

In a previous paper we identified that a majority of social science researchers find it challenging to learn to use new software, especially because these technologies are changing rapidly. Many researchers are developing new tools to serve specific project requirements. We wanted to understand this ecosystem and share the big picture with the research community, identify best practices and successful tools, find out where more support is needed, and think about how SAGE Ocean can help to promote tools for the social sciences. In this paper we reviewed 418 software tools and technologies used by social science researchers. We explored who leads the development of these tools, where the supporting communities and investors are, and what challenges people face. This is a first step towards making software tools and technologies more accessible to social science researchers.

The real innovation in the tools space is, in our opinion, still a result of academic practice. Surveys are a case in point. Successful survey tools like Qualtrics, SurveyMonkey, and TypeForm do the basic job of survey management, enable an easier-to-use interface in designing complex questionnaires, interface with other games and experimental sites, and help recruit participants. The next most fascinating development, however, will address the effectiveness and efficiency of these surveys. Matt Salganik and his team developed allourideas.org, a free surveying tool that enables researchers to engage their respondents to contribute to the survey development while also collecting answers. A group of computer scientists and social scientists from the University of Washington Madison developed NEXT,[63] a surveying tool powered by an algorithm that adapts the survey sample and questions as more people answer them to get better results faster and without having to rerun the survey. Academics are at the forefront of these projects, developing new methodologies and tools that will eventually be taken up by the private sector. To do that more effectively, they need a community of users, financial support, consortia and other organisations that are able to host and scale up their tools. This will ensure that more researchers can use and build on existing tools, as well as enable the development of sustainable models and the growth of the community of users.

We found that, as with startups in other sectors, the research teams developing successful tools and technologies used by social science researchers work fast. The UK-based Prolific team, for example, came up with its minimum viable product in just four weeks. Additionally, RapidMiner is a successful tool that researchers developed in relatively rapid iterations, raising more than $30M in funding[64] while also open sourcing the code. Where tools have gone through significant exits (IPO for SurveyMonkey[65] in 2018, acquisition for Qualtrics[66] in 2018, Tableau[67] and FigureEight[68] in 2019) an executive with a background in business and/or consultancy was always involved. Whilst many of these tools are used by social science researchers, their success is ultimately driven by a growing need to explore and exploit the market for business intelligence.

The Future of Tools and Social Science

What will the future of the tech ecosystem for the social sciences look like? Will social scientists use algorithms to research humans and society? Will they simply deploy machines to monitor people's behavior online or extract information from surveillance videos in driverless cars or fully connected houses? These are all questions that we are exploring with SAGE Ocean as we envision ways to help social scientists work with big data and new technologies. Any guesses as to the future of technologies in this field will almost certainly be wrong, but if we were to speculate about the next 20 years we could consider: What will society be like or be obsessed with? Where are the data about society going to be, or what types of institutions will hold it? Where will the most promising social scientists be working? Who will be developing the tools, and who will be funding them?

According to the Imperial College Tech Foresight (2019), in 20 years' time, our new best friends will be avatar companions that will know everything about us, we will be uploading new data to our brains, and recording our dreams. Self-writing software will be easily available. Someone or something will be monitoring the public mood and releasing forecasts almost like weather forecasts. We may even have an interface to control thoughts, along the lines of CTRL-Labs[69] and Emotiv.[70] A lot more data will be available, and possibly accessible, for research. But who will own the data and where will the talent and academic experts who can explore it be?

In a potential worst case scenario, we will see the most promising social researchers working for big tech, performing ongoing experimentation, advancing insights into consumer behavior, and analysing data from implantable trackers on employees—all to drive shareholder profits at the expense of society. The tools ecosystem will be driven by commercial priorities, immediate impact, and the need for robust results, combined with abundant budgets and skills, with no limits on automation.

In a best-case scenario, robust research will still happen within independent centres of excellence, like universities, that are not affiliated with or influenced by state or private interests. Data, collected by a variety of private and public organisations, will be brokered by regulated intermediaries (on some type of blockchain or another distributed technology) that prioritise inclusive research. In 20+ years, we will go through a convergence of tooling and skills followed by a further specialisation in the development of new tools. In other words, we will first see more social science researchers acquire computational skills (a continuation of the current trend), and then a standardisation of computable models as integral methodologies to social science research. As research cycles go, this will reach a plateau, and the really interesting outputs will start appearing at the fringes again—where computational boundaries and social theory are pushed. The predominant tools in 20 years' time will have an interface that is intelligible and usable by social science researchers, as well as a backend that can be exploited further due to advances in computing.

Want to talk about the future of tools, data, and methods in the social sciences? SAGE Ocean would love to hear from you. Contact us at ocean@sagepub.com or visit ocean.sagepub.com.

Notes

1. https://ocean.sagepub.com
2. https://ocean.sagepub.com/research-tools
3. https://ocean.sagepub.com/concept-grants
4. https://socialmedialab.ca/
5. https://wiki.digitalmethods.net/Dmi/DmiAbout
6. https://lazerlab.net/
7. https://publicdatalab.org/
8. https://dirtdirectory.org/
9. https://sourceforge.net/
10. https://github.com/
11. https://labworm.com/
12. https://www.wikipedia.org/
13. https://www.crunchbase.com/
14. For latest raw dataset and analysis, see this GitHub repository: https://github.com/Southampton-RSG/soton_software_survey_analysis_2019
15. For the latest raw and cleaned dataset, see this GitHub repository: https://github.com/softwaresaved/local-and-regional-software-surveys/tree/master/data/sage-ocean-social-science
16. Computer Assisted Qualitative Data Analysis Software generally refers to packages like NVIVO, MaxQDA, Atlas.ti, etc.
17. These criteria were adjusted according to the features of each pricing model.
18. SAGE Publications invested in Thusly, the company that offers TagWorks, in 2018.
19. The inferences are based on most popular papers.
20. https://ocean.sagepub.com/

21. Just over 300 million monthly users in 2019, compared to two billion on Facebook and one billion on Instagram, for more details: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

22. https://socialscience.one/

23. https://engineering.linkedin.com/teams/data/projects/economic-graph-research

24. https://www.docnow.io/

25. On how ‘rehydration’ works, see https://medium.com/on-archivy/on-forgetting-e01a2b95272

26. https://gate.ac.uk/gate/doc/papers.html

27. https://www.force11.org/group/software-citation-working-group

28. https://ocean.sagepub.com/research-tools

29. https://dirtdirectory.org/

30. https://github.com/dhtaxonomy/TaDiRAH

31. https://wiki.digitalmethods.net/Dmi/ToolDatabase

32. https://dighumlab.org/tools/

33. https://sourceforge.net/

34. https://labworm.com/

35. http://www.laurenceanthony.net/software.html

36. https://gephi.org/

37. https://www.software.ac.uk/resources/guides/guides-developers

38. https://www.apereo.org/

39. http://cx.jacobs.cornell.edu/

40. https://numfocus.org/

41. https://ropensci.org/

42. https://www.software.ac.uk/

43. http://commons.pelagios.org/

44. https://recogito.pelagios.org/

45. http://www.foastat.org/

46. https://shiny.rstudio.com/

47. https://okfn.org/

48. https://ckan.org/

49. https://wiki.digitalmethods.net/

50. https://github.com/digitalmethodsinitiative/dmi-tcat

51. http://digitalscholar.org/

52. https://www.smrfoundation.org/

53. https://nectar.org.au/

54. https://cos.io/

55. https://www.clarin.eu/

56. https://www.dariah.eu/

57. https://www.cessda.eu/About

58. https://ocean.sagepub.com/concept-grants

59. https://www.digital-science.com/investment/catalyst-grant/

60. https://opencollective.com/

61. https://www.superiorideas.org/

62. Data retrieved from https://www.crunchbase.com/ in July 2019

63. http://nextml.org/

64. Data retrieved from https://www.crunchbase.com/organization/rapidminer in July 2019

65. On SurveyMonkey's IPO, see: https://www.marketwatch.com/story/surveymonkey-ipo-5-things-to-know-about-the-survey-software-maker-2018-09-01

66. On SAP's acquisition of Qualtrics, see https://techcrunch.com/2018/11/11/sap-agrees-to-buy-qualtrics-for-8b-in-cash-just-before-the-survey-software-companys-ipo/

67. On Salesforce's acquisition of Tableau, see: https://seekingalpha.com/article/4271153-sales-force-com-tableau-acquisition-game-changer

68. On Appen's acquisition of FigureEight, see: https://techcrunch.com/2019/03/10/appen-acquires-figure-eight/

69. https://www.ctrl-labs.com/, acquired by Facebook in September 2019

70. https://www.emotiv.com/

References

Andersen, D. (2013, March 3). The story behind Qualtrics, the next great enterprise company [Online]. Retrieved from https://techcrunch.com/2013/03/02/the-story-behind-qualtrics-the-next-great-enterprise-company/

Bosman, J., & Kramer, B. (2016). Innovations in scholarly communication - Data of the global 2015–2016 survey [Dataset]. Zenodo. doi:10.5281/zenodo.49583

Campfens, Y. (2019). Market research report: What has become of new entrants in research workflows and scholarly communication? doi:10.31219/osf.io/a78zj

Clement, J. (2019). Social media – Statistics & facts. Retrieved from Statista: https://www.statista.com/topics/1164/social-networks/

Davies, L. (2019, June 4). Social media data in research: A review of the current landscape [Blog post]. Retrieved from https://ocean.sagepub.com/blog/social-media-data-in-research-a-review-of-the-current-landscape

Duca, D. (2019, June 27). Social scientists working with LinkedIn data [Blog post]. Retrieved from https://ocean.sagepub.com/blog/social-scientists-working-with-linkedin-data

Grusky, M., Naaman, M., & Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* doi:10.18653/v1/N18-1065

Hettrick, S. (2019). Software in Southampton: Presentation on the results of the University of Southampton software survey conducted in June 2019 [Slides]. Retrieved from https://slides.com/simonhettrick/software-in-southampton#/3/9

Hu, S. (2019, May 20). How researchers around the world are making use of Weibo data [Blog post]. Retrieved from https://ocean.sagepub.com/blog/how-researchers-around-the-world-are-making-use-of-weibo-data

Imperial Tech Foresight. (2019). Table of disruptive technologies [Infographic]. Retrieved from https://www.imperialtechforesight.com/visions/table-of-disruptive-technologies/

Jacks, J. (2018, July 9). #3: COSS business model progressions [Blog post]. Retrieved from https://medium.com/open-consensus/3-oss-business-model-progressions-dafd5837f2d

Jackson, M. (2019). How to cite and describe software [Online]. Retrieved from https://www.software.ac.uk/how-cite-software

Katz, D. S., McInnes, L. C., Bernholdt, D. E., Mayes, A. C., Hong, N. P., Duckles, J., Gesing, S., Heroux, M. A., Hettrick, S., Jimenez, R. C., Pierce, M. E., Weaver, B., & Wilkins-Diehr, N. (2019). Community organizations: Changing the culture in which research software is developed and sustained. *Computing in Science & Engineering*, *21*, 8–24. doi:10.1109/MCSE.2018.2883051

Lumsden, J. (2019, February 19). So, you want to run an online experiment? [Blog post]. Retrieved from https://ocean.sagepub.com/blog/how-to-run-an-online-experiment

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). Who is doing computational social science? Trends in big data research [White paper]. London, UK: SAGE Publishing. doi:10.4135/wp160926.Retrieved from https://us.sagepub.com/sites/default/files/CompSocSci.pdf

Qualtrics. (2019, July 11). Qualtrics ranked top 10 most innovative company in 2019 GRIT report [Press release]. Retrieved from https://www.qualtrics.com/news/qualtrics-ranked-top-10-most-innovative-company-in-2019-grit-report/

Radford, J. (2019, April 29). Collecting social media data for research [Blog post]. Retrieved from https://ocean.sagepub.com/blog/collecting-social-media-data-for-research

RapidMiner. (2015, September 1). RapidMiner embraces its community and open source culture delivering get-more-open-core predictive analytics [Press release]. Retrieved from https://rapidminer.com/news/rapidminerembracesitscommunity/

SAGE Ocean. (2019, May 2). Researchers are awarded grants to study Facebook data and its influence on elections [Blog post]. Retrieved from https://ocean.sagepub.com/blog/researchers-are-awarded-grants-to-study-facebook-data-and-its-influence-on-elections

SAP. (2018, November 11). SAP SE to acquire Qualtrics International Inc., sees experience management as the future of business [Press release]. Retrieved from https://news.sap.com/2018/11/sap-to-acquire-qualtrics-experience-management/

Smith, A. M., Katz, D. S., Niemeyer, K. E., & FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science* 2:e86. doi:10.7717/peerj-cs.86

Tay, A. (2019, July 4). Are Google & web scale discovery services — Low skill cap, low performance cap tools? [Blog post]. Retrieved from https://medium.com/@aarontay/are-google-web-scale-discovery-services-low-skill-cap-low-performance-cap-tools-afa232f0d196

Wiggers, K. (2019, May 23). Chick-fil-A's AI can spot signs of foodborne illness from social media posts with 78% accuracy. *VentureBeat.* Retrieved from https://venturebeat.com/2019/05/23/chick-fil-as-ai-can-spot-signs-of-foodborne-illness-from-social-media-posts-with-78-accuracy/