This is a post-peer-review, pre-copyedit version of an article published in *Netnomics*. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11066-018-9125-2.

DataGorri: A Tool for Automated Data Collection of Tabular Web Content*

Julian Hackinger†

29th October 2018

Abstract

The era of the internet has been a boon for empirical and evidence-based research. By providing ever increasing amounts of data, the internet offers numerous opportunities for new empirical studies. While some research questions require data that was previously more time-consuming to collect, other data was simply not available before the creation of the internet. However, publicly available information is still often unstructured and its collection can be highly resource-intensive. In this paper we present DataGorri, a software enabling the user-friendly and automated collection of repetitive and non-repetitive tabular data that is freely available on websites. This paper lays out the motivation underlying the software's creation, describes its usage, and discusses its advantages and limitations.

Keywords: Software; DataGorri; Web Scraper; Data Scraper; Crawler; Data collection.

JEL Classification: C81; C82; C89.

*We would like to thank everyone who has contributed to current or previous versions of DataGorri: Ivaylo Dimitrov, Matthias Franze, Stefan Hentschel, Lukas Holzner, Florian Kreitmair, Daniel Krieger, Michael Legenc, and Marc Müller. A list of DataGorri's developers and contributors can also be found at `https://www.julianhackinger.com/software/datagorri/`. Furthermore, we thank Christian Feilcke, Miriam Leidinger, and two anonymous reviewers for comments, and Alexander Schlimm for research assistance.

†Email: julian.hackinger@tum.de. Postal address: Technical University of Munich, Chair of Economics, Arcisstr. 21, 80333 Munich, Germany. Phone: +49 89 289 25707. ORCID: 0000-0002-6011-892X.

1 Introduction

"The ultimate goal of economic science is to improve the living conditions of people in their everyday lives" (Samuelson and Nordhaus 1998, p. 7). To realise this goal, it uses theoretical models to derive predictions and analyses data in order to test hypotheses.

In recent decades, the number of empirical studies has grown tremendously. Between 1963 and 2011, the share of empirical articles published in the American Economic Review (AER), Journal of Political Economy (JPE), and Quarterly Journal of Economics (QJE), all of which are top journals, has increased by 34 percent from 47.8 to 63.9 percent. This rise is mainly attributable to the expanding feasibility and popularity of using individually assembled data. Since 1993 in particular, the share of studies using own data instead of publicly provided data has quadrupled (Hamermesh 2013).

This development coincides with the introduction of the World Wide Web in 1991. The internet has proven to be a primary resource for empirical research, augmenting the amount of available data and lowering the costs of access to it (Edelman 2012). As of August 2018, Netcraft (2018) counted over 184 million active websites. Many of these provide information that is valuable to economic scholars and can provide further insights to open questions. Some of the available data even is entirely novel and allows new research projects.

However, usually, information on websites is not gathered and presented for scientific use. Also, required data is often provided by many different sources. As a result, a critical lack of structure seems to be the norm (Einav and Levin 2014a,b). This makes the task of manually compiling online data very time-consuming. Unfortunately, the bulk of software that automates such processes is often too expensive for academic use. Moreover, software must be tailored to specific projects, which further increases costs and decreases scope. Thus, many hours of scholarly work have been used to copy and paste numbers, tables, and texts. Those researchers that are gifted with coding skills may have spent hours creating lines to simplify this job. Yet, in comparison to research data, which is increasingly made public by authors, such code snippets or entire software packages often seem to be kept private and are thus rather difficult to find.

In order to facilitate further research with internet data, we decided to develop and share a software package that might benefit others in their data collection. In this paper we introduce DataGorri[1], a free-to-use software that is generically applicable and can collect data from almost all standardised tables on the web.[2]

The software can be used free of charge. However, by accepting the license agreement when downloading DataGorri, the user agrees to cite this technical paper whenever DataGorri has been used for research purposes (cite ware). The package and documentation can be downloaded from `www.julianhackinger.com/software/datagorri/` and `https://github.com/julhac/datagorri`.

DataGorri is by no means a final product. As only its application can uncover bugs or further potential, anyone is kindly invited to contribute and to send in suggestions for improvement. For this purpose and for problems or questions, please consult the FAQ on the website (`www.julianhackinger.com/software/datagorri/faq/`) or contact the author.

In the following section, we describe how DataGorri works in theory and how it can assist in the data collection process. Subsequently, in Section 3, we put the theory into practice and use DataGorri to download data on institutions in the RePEc archive. Section 4 points out advantages and limitations and discusses further possible improvements. Section 5 concludes.

2 DataGorri

DataGorri is an application used to extract data from tables found on websites. It has the ability to run through a list of predefined links and save specified information from tables. Importantly, the respective tables must always be located in the same place of each link and of the same format (the same number of columns; the number of rows is irrelevant). This applies for instance to academic rankings by region (e.g. `https://ideas.repec.org/top/top.usa-ma.html`), sporting squads and statistics by team,

[1]Katagorri = Basque name for squirrel; DataGorri collects data like a squirrel gathers nuts.

[2]Before scraping websites, please ensure that you have the permission to do so.

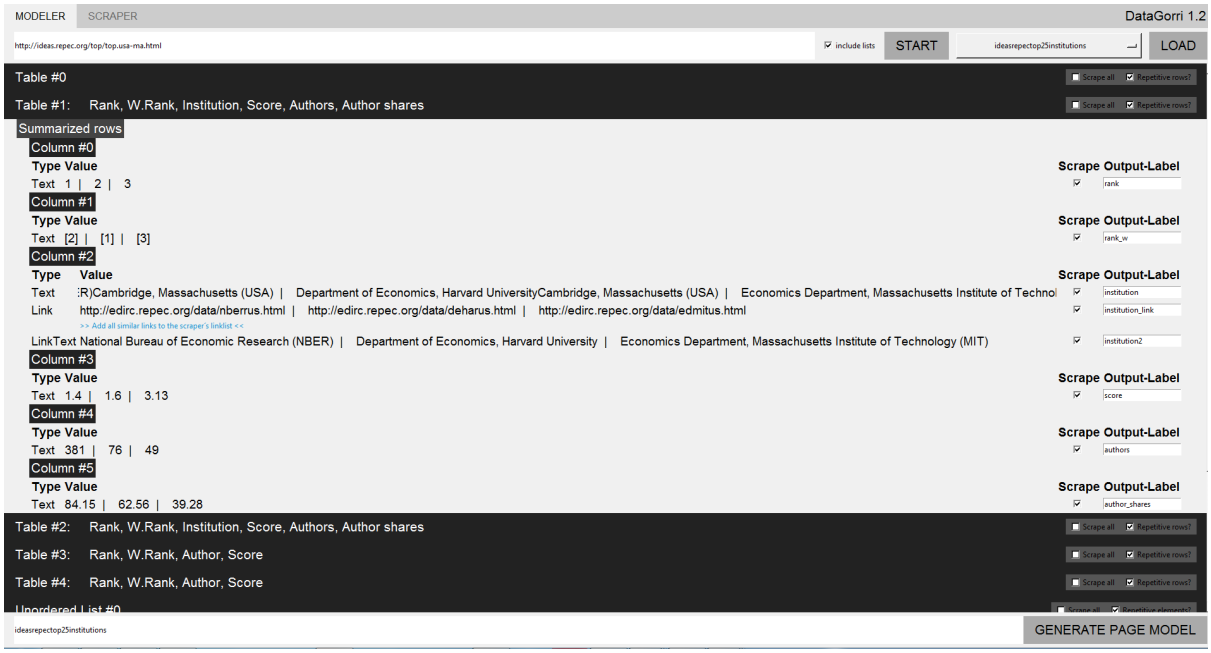

Figure 1: The modeler is used to create a page model that is applied to a list of websites.

or year (e.g. `https://www.transfermarkt.com/1-bundesliga/tabelle/wettbewerb/L1?saison_id=2016`), and monthly weather tables (e.g. `https://en.tutiempo.net/climate/01-2017/ws-108660.html`). At the end of a scraping task, the data is saved to a .csv file format that can be read by common statistics packages.

In order to set up a scraping task, two steps are necessary:

1. Create a page model to define the content of interest.

2. Input links of websites that should be scraped.

2.1 The Page Model

The first step is handled by DataGorri's modeler (Figure 1) which can be found under the tab "MODELER". Here, the user has to input an URL and inspect the website's structure to define the contents which she is interested in. To this end, the modeler displays all tables contained on the respective page. On the first level, it lists mother tables including consecutive number, their headers if available, and a tick box to define whether the table is repetitive or non-repetitive.

Repetitive tables (Table 1) contain information for multiple observations with observations being below one another. In repetitive tables, each column contains information

Table 1: Repetitive table.

Object	**Cost**	**Availability**
table	250	yes
chair	50	yes
bed	150	no

Table 2: Non-repetitive table.

Object	table
Cost	250
Availability	yes

Table 3: Mother table with child tables.

Object	**Cost**	**Information (child tables)**	
		Colour	**Availability**
table	250	black	100
		white	0
		red	70
chair	50	black	80
		white	10
		red	70
bed	150	black	0
		white	0
		red	50

belonging to one variable. Hence, variable names are ordered horizontally on the top. In contrast, non-repetitive tables (Table 2) usually contain only information on one object or observation. Here, the variable names are ordered vertically on the left. The user must specify whether the table is repetitive or not for the data to be displayed and downloaded correctly.

By clicking the header of each table, the modeler provides more detail on the information contained in the table. Some tables contain so-called child tables (tables within tables, see for example Table 3 in which the column "Information" contains child tables) which can equally be expanded to show the contained data. The user then simply selects the desired contents of one or more of these tables and, by saving it, creates a page model for this specific page structure. The model can then be used for all similarly structured pages.

2.2 Link List

The second step is collecting one or more URLs that should be scraped with a certain page model. These links should be entered below one another in the "SCRAPER" tab

(Figure 2). To be able to run the same request at a later time, we recommend saving the list of links under a meaningful name.

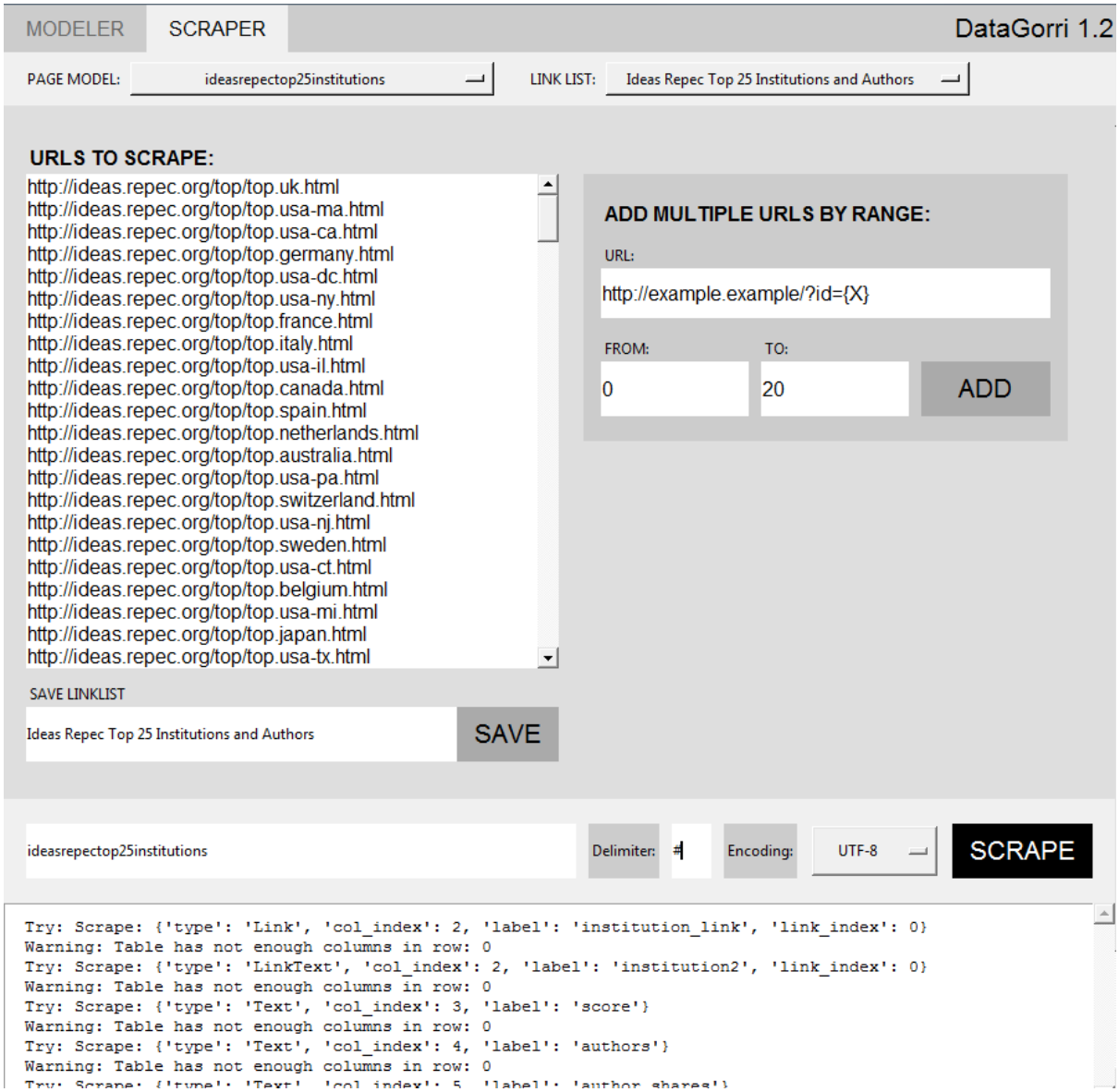

Figure 2: In the subsection of the scraper, the user can enter a list of websites that contain formally identical tables which will be scraped.

In order to facilitate the collection of links, DataGorri includes two methods to quickly collect multiple similar URLs. First, this is the link generator at the right hand side of the tab "SCRAPER". Many websites are structured in a way such that a main URL is followed by a count variable that incrementally increases (e.g. page number or year: `www.example.com/data/2016`). By replacing the counter with "{X}" and defining the corresponding range for "X", one can add multiple URLs at once.

The second method can be found in DataGorri's modeler tab. After having loaded a page structure, the modeler displays the option "Add all similar links to the scraper's linklist" whenever it encounters a hyperlink. Once one has found an overview page that includes links to websites that should all be scraped, the second option can be used to easily add several links to the link list.

Ultimately, upon clicking "SCRAPE", DataGorri will go through the selected list of URLs and extract all data located in the predefined cells or positions. After completion, the scraped result is automatically saved to a .csv file format. Please bear in mind that existing files with an identical name are overwritten.

All files (page models, link lists, and result files) are saved in folders on the desktop that are automatically created upon the first execution of DataGorri. The files can also be shared, which facilitates the replication of studies that used DataGorri.

With these basic functions, DataGorri is highly flexible but remains user-friendly and comprehensible even for non-experienced users. For more detailed information on its functionality, please consider the user manual and documentation of DataGorri.

3 Application

3.1 Research question

So far, we provided the motivation to develop and use DataGorri and described its usage in theory. However, it certainly helps interested readers to see how DataGorri can be applied in practice. For demonstration purposes, we picked an exemplary research question.

We are interested in whether and how the number of authors or academics per institution is correlated with the institution's average research performance. In a broader sense, efficiency and economies of scale in research are frequently discussed topics (Abramo et al. 2012). Wuchty et al. (2007) show that research output benefits from collaborations. A larger institution implies a larger pool of scholars that might have matching research interests. Therefore, it seems reasonable that larger institutions facilitate conducting projects in teams and finding co-authors. Moreover, a better match in research interests

or a higher number of matches should also increase the likelihood to receive valuable input from colleagues. Hence, research output and quality should increase with the size of institutions.

While Jordan et al. (1988, 1989) and Meador et al. (1992) conclude that "publishing productivity rises with faculty size at a diminishing rate" (p.347), Golden and Carstensen (1992a,b) dispute the impact of department size on per capita publications (see Abramo et al. 2012, for a summary on that topic and a recent study). Studying a Stackelberg differential game between journal editors and authors, Faria and Goel (2010) propose that a larger network (e.g. being at a larger department) has a positive effect on an author's number of citations but not on her number of publications or research quality.

3.2 Data

In order to investigate this question, we resort to IDEAS (`ideas.repec.org`). Based on the RePEc archive, IDEAS is the largest bibliographic database dedicated to economics as it indexes over 2,600,000 items of research and more than 50,000 authors (as of 03 September 2018). Among other things, IDEAS also ranks institutions and authors by a performance score. We will use that score for our analysis.

The IDEAS average rank score is determined by taking a harmonic mean of the institution's rank relative to a corresponding sample (e.g. within a region) in each RePEc criterion. On IDEAS, authors, institutions, journals, and countries are ranked according to (variations of) the number of works registered with RePEc, citation counts, journal page counts, abstract views and downloads, and the author's network (see Zimmermann 2013, for a description of all criteria). Across criteria, the rank of a specific author or institution might vary. According to Zimmermann (2013) this entails the risk of cherry picking by authors and institutions themselves, editors, and publishers. Therefore, IDEAS uses the harmonic mean of the ranks of all criteria to calculate a score. Aggregating ranks, a lower score is better than a higher one.

3.3 Data collection

For the analysis we first use DataGorri's feature to add similar links to the scraper's link list. The website `https://ideas.repec.org/top/top.country.all.html` lists all countries that have research output catalogued in the RePEc archive. Entering this link into the modeler returns the corresponding table including the links to all countries with research institutions or authors in RePEc. Clicking on "Add all similar links to the scraper's linklist" copies all links to the link list. At this step an exceptional issue arises. The links provided in the table lack "/top/" to form complete links like `https://ideas.repec.org/top/top.usa-ma.html`. Instead links without target like `https://ideas.repec.orgtop.usa-ma.html` are provided. The missing part can be inserted between "org" and "top" manually using any word processor. This demonstrates that, despite its convenience, DataGorri still provides enough flexibility to its users.[3] Afterwards, we copy the list of links to DataGorri's link list and save it (we will refer to this list as country link list).

Second, we take the first link to the top 25% institutions and authors in Massachusetts, USA (`http://ideas.repec.org/top/top.usa-ma.html`, or any other link from the country link list we want to scrape) and enter it in the modeler. DataGorri returns the two tables on the page containing the top 25% institutions (Figure 3) and the top 25% authors (which we will not use) in Massachusetts. We will scrape the corresponding tables listing each country's top 25% institutions, country by country. Clicking on the first header opens the content of the respective table: rank in the corresponding country, worldwide ranking, institution, score, number of authors, and author shares. We select to scrape all variables by ticking the corresponding boxes and assign meaningful output labels. Finally, we save the generated page model for this table and define it as the institutions page model. This model can now be applied to all items in the country link list.

Now, with the institutions page model and the country link list at hand we return to the "SCRAPER" tab. First, we select the institutions page model in the drop-down menu

[3]The option to select a different delimiter than the default (;), and to choose between UTF-8 and Latin-1 character encoding are further features that increase DataGorri's flexibility.

Top 25% institutions in Massachusetts (United States), all authors, all publication years

Rank	W.Rank	Institution	Score	Authors	Author shares
1	[2]	National Bureau of Economic Research (NBER) Cambridge, Massachusetts (USA)	1.4	381	84.15
2	[1]	Department of Economics, Harvard University Cambridge, Massachusetts (USA)	1.6	76	62.56
3	[3]	Economics Department, Massachusetts Institute of Technology (MIT) Cambridge, Massachusetts (USA)	3.13	49	39.28
4	[4]	Kennedy School of Government, Harvard University Cambridge, Massachusetts (USA)	4.45	74	40.19
5	[5]	Department of Economics, Boston University Boston, Massachusetts (USA)	5.33	58	52.74

Figure 3: Excerpt of a screenshot of the top 25% institutions in Massachusetts (United States) on `https://ideas.repec.org/top/top.usa-ma.html` (Accessed 03 September 2018).

for page models. Next, from the drop-down menu for link lists, we select the country link list containing the links to all countries in RePEc. Finally, we choose a meaningful name for the result file and click on scrape. On our machine[4], the download took three minutes. The request results in a .csv file containing observations on 2,633 institutions representing the top 25% in their respective country (as of 03 September 2018). The file can now be imported to any common statistics software and analysed.

3.4 Results

To examine the correlation between an institution's performance and its size, we consider the institutions' IDEAS scores and their number of authors on RePEc.

As Figure 4 shows, the IDEAS score improves with the logarithmic number of authors.[5] This relationship is highly significant (Pearson's Correlation coefficient $= -0.2106$, $p < 0.0000$) and is further substantiated in regressions that control for country effects (Table 4 and Figure 4).[6] Since the number of authors is in log scale, an increase

[4]Windows 7, 64 Bit, 3.60 GHz, 32 GB Ram, 100 Mbit/s.

[5]Note that the IDEAS rank per criterion and, thus, also the IDEAS score is calculated for each country in our country link list separately. Hence, each country has distinct rankings for all criteria, which are also aggregated on country level only.

[6]As the variable IDEAS Score exhibits overdispersion (its variance is greater than its mean), a negative binomial regression is more appropriate than a poisson regression.

in the number of authors per institution of equal size is associated with a larger improvement of the IDEAS score the fewer authors an institution comprises. One can consider this as decreasing returns to scale.

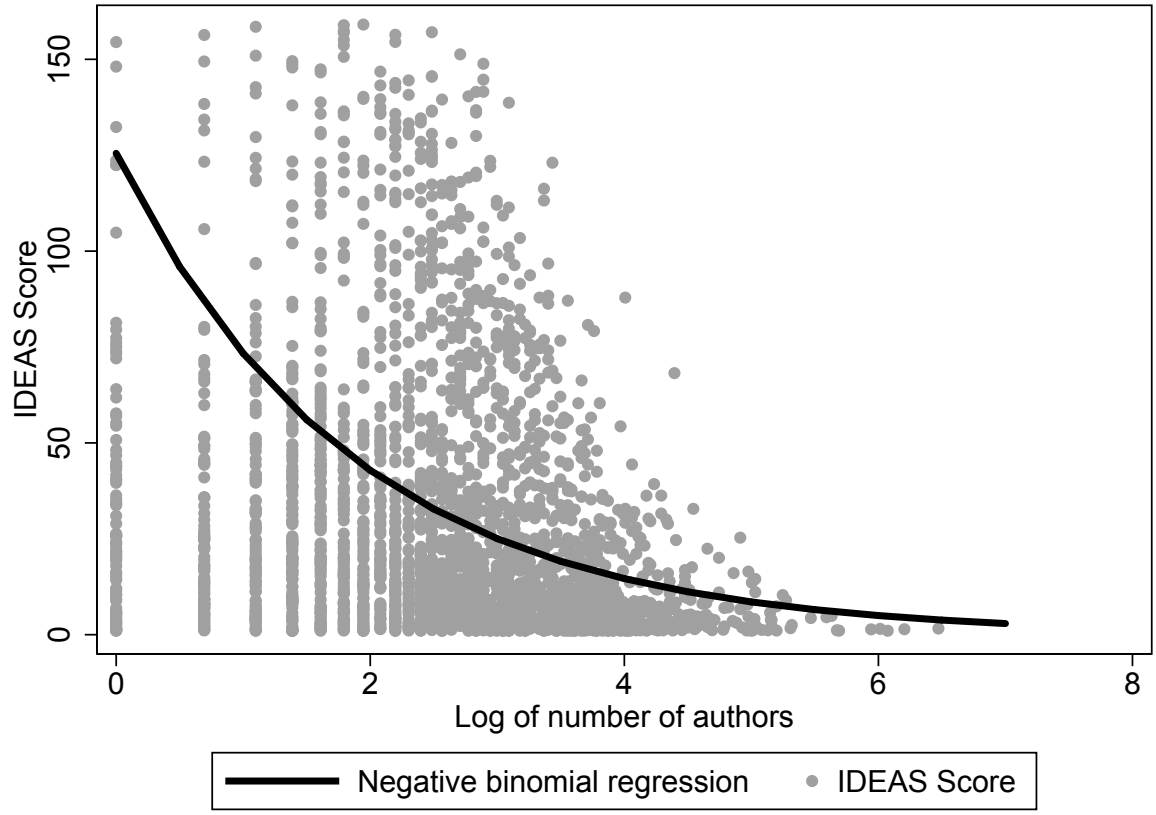

Figure 4: Log of number of authors per institution and institution's IDEAS score on `ideas.repec.org`.

Hence, similar to Jordan et al. (1988, 1989) and Meador et al. (1992), we find a positive correlation between the number of authors per institution and the performance of institutions measured in IDEAS scores. However, this positive relationship is decreasing with the number of authors per institution. Obviously, the causality could go in both directions. Better institutions could attract more funding and could therefore also hire more scholars. Alternatively, a higher number of academics per institution could result in more interaction between them and could lead to better and more output. Identifying the causal direction thus requires further research. Also, the IDEAS score is an aggregate measure. Therefore, it can not be used to estimate how citations, quality, and number

Table 4: Poisson and negative binomial regression of the institutions' IDEAS score on their number of authors registered on RePEc.

	IDEAS Score	
	Poisson regression	Negative binomial regression
Log of authors	-0.430***	-0.537***
	(0.00353)	(0.0113)
Constant	0.958	1.055
	(0.545)	(0.620)
Country Effects	Yes	Yes
Log likelihood	-13,870.695	-9,328.452
Pseudo R^2	0.714	0.201
Observations	2,597	2,597

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

of publications are individually affected by institution size (c.f. Faria and Goel 2010). Further empirical work is necessary to uncover these relationships.

4 Advantages and Limitations

DataGorri's advantages are manifold. While it has already been tested and applied extensively, we are sure that further applications beyond our scope exist. In any case, DataGorri is able to save researchers a substantial amount of time. The more websites a scholar wants to compile data from, the more they benefit from DataGorri. Using the program merely requires a small amount of upfront effort in setting up the page model, but it can be scaled to an unlimited number of websites thereafter. The two options for gathering links described above help reduce the work necessary for the latter.

DataGorri is specialised to scrape tabular data. Hence, it cannot extract data contained in unstructured texts. Other tools exist for such purposes.

In order to provide some degree of automation, the tables in question need to look alike and be located at the same position within a website.

Furthermore, DataGorri does not recognise whether a table is repetitive or not. It therefore requires some user feedback. We aim to tackle some of these issues in future releases of the program.

5 Conclusion

In this paper, we introduced DataGorri, a software that enables researchers to collect repetitive and non-repetitive tabular data that is available on websites. For that purpose, DataGorri runs through a list of predefined links, which all contain the same type of table, and exports the tabular data to a .csv file format.

We are aware of the fact that compiling online data can be a cumbersome task and very time-consuming. We provide DataGorri free of charge. However, we require to be cited whenever DataGorri has been used for scientific research (cite ware). This ensures that more colleagues will learn about DataGorri and are able to benefit from using it. Sometimes scientific work is impeded by preparatory efforts. With DataGorri, we hope to lower this hurdle.

References

Abramo, G., Cicero, T., and D'Angelo, C. A. (2012). Revisiting size effects in higher education research productivity. *Higher Education*, 63(6):701–717.

Edelman, B. (2012). Using internet data for economic research. *Journal of Economic Perspectives*, 26(2):189–206.

Einav, L. and Levin, J. (2014a). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Einav, L. and Levin, J. (2014b). Economics in the age of big data. *Science*, 346(6210):1243089.

Faria, J. R. and Goel, R. K. (2010). Returns to networking in academia. *Netnomics*, 11(2):103–117.

Golden, J. and Carstensen, F. V. (1992a). Academic research productivity, department size and organization: Further results, comment. *Economics of Education Review*, 11(2):153–160.

Golden, J. and Carstensen, F. V. (1992b). Academic research productivity, department size and organization: Further results, rejoinder. *Economics of Education Review*, 11(2):169–171.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172.

Jordan, J. M., Meador, M., and Walters, S. J. (1988). Effects of department size and organization on the research productivity of academic economists. *Economics of Education Review*, 7(2):251–255.

Jordan, J. M., Meador, M., and Walters, S. J. (1989). Academic research productivity, department size and organization: Further results. *Economics of Education Review*, 8(4):345–352.

Meador, M., Walters, S. J., and Jordan, J. M. (1992). Academic research productivity: Reply, still further results. *Economics of Education Review*, 11(2):161–167.

Netcraft (2018). August 2018 web server survey. `https://news.netcraft.com/archives/2018/08/24/august-2018-web-server-survey.html`. Accessed: 03 September 2018.

Samuelson, P. A. and Nordhaus, W. D. (1998). *Economics*. Irwin/McGraw-Hill, Boston, Massachusetts, 16th edition.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science (New York, N.Y.)*, 316(5827):1036–1039.

Zimmermann, C. (2013). Academic rankings with RePEc. *Econometrics*, 1(3):249–280.