Developments in the Field

Big Data on BHR: Innovative Approaches to Analysing the Business & Human Rights Resource Centre Database

Nicole JANZ*¶§, James ALLEN-ROBERTSON**‡¶§, Rajeshwari MAJUMDAR***¶§ and Shareen HERTEL****¶§

Keywords: Big data, Business impact, Human rights, Industries, Machine learning

I. Overview

Data limitations in the field of business and human rights (BHR) are a significant challenge to developing transparent and replicable scholarship. Difficulties stem from the proprietary nature of much BHR data; incomparability of information across firms, sectors or regions; and coding challenges of existing information. Together, issues like

¶ Conflicts of interest: The authors declare none.

‡ James Allen-Robertson acknowledges support from 'The Human Rights and Big Data Technology Project' Economic and Social Research Council grant number ES/M010236/1.

§ We thank colleagues at the Business and Human Rights Resource Centre for their expert advice on the use of their database and for insightful conversations about scholarship and activism in the BHR field. We thank an anonymous reviewer for valuable feedback.

* Nicole Janz is an Assistant Professor in International Relations (with Quantitative Methods) at the University of Nottingham, UK, where she is affiliated with the Rights Lab's Business and Economies Programme. She researches and publishes on globalisation, human rights and corruption, where she focuses on improving measurement and research transparency.

** James Allen-Robertson is a Computational Sociologist and Senior Lecturer in the Department of Sociology at the University of Essex, UK. His early book *Digital Culture Industry* (2013, Palgrave Macmillan) charted the role of piracy in the development of digital media distribution; more recent work focuses on the development of new methods in computational social science, digital cultures, technology and power.

*** Rajeshwari Majumdar‡‡ is a doctoral student in the Department of Politics at New York University, USA. She studies political behaviour in developing countries, focusing on group identity and misperceptions about politics. She graduated from the University of Connecticut in 2018 with Honours in Political Science, Economics and Mathematics-Statistics.

**** Shareen Hertel is a Professor of Political Science and Human Rights at the University of Connecticut, USA. She is editor of *The Journal of Human Rights*, co-editor of the *International Studies Intensives* book series (Routledge) and author, most recently, of *Tethered Fates: Companies, Communities and Rights at Stake* (Oxford, 2019).

‡‡ This article was originally published with incorrect author information. A notice detailing this has been published and the error rectified in the online PDF and HTML copies.

Business and Human Rights Journal, 6 (2021), pp. 120–126
doi:10.1017/bhj.2020.18

these constrain the nature of cumulative research. Absent standard protocols for sharing and validating large-scale data, for example, make it difficult for scholars to progress in generating new conceptual or policy breakthroughs in the field of BHR.

We identify a promising development: namely, the emergence of large-scale automated coding of company-related 'Stories' from the award-wining and publicly available Business and Human Rights Resource Centre (BHRRC) database. The BHRRC tracks the performance of over 9,000 companies worldwide. Its open-access website features a rich archive of media articles and non-governmental organization (NGO) reports concerning allegations against companies, along with company responses to such allegations. Data on the site range across individual countries, regions and industries globally.

However, the 'Stories' on the BHRRC site are difficult to analyse in a comparative and systematic manner, in part because hand-coding related content is time consuming and complex. Research teams at the University of Connecticut, the University of Nottingham and the University of Essex have now developed and used application programming interfaces (APIs) in order to code units of data from the BHRRC website. Their resulting analyses of trends in stakeholder consultation (Hertel, 2019)[1] and trends in industry patterns of rights issues demonstrate the potential the BHRRC database holds for transforming quantitative and mixed-methods scholarship in BHR and its broader policy implications.

This piece introduces new approaches to using big data in the BHR field, using the BHRRC repository as an example. We analyse the relative usefulness and richness of the information gathered. We then point toward ways of reshaping and analysing these large-scale data via custom-built analytic and data visualization tools. We illustrate our advances with concrete examples and illustrations. Finally, we discuss some of the challenges involved in working with this kind of data (such as under-reporting of certain issues) and we demonstrate approaches to mitigating them.

In this piece, we combine findings and insights from our partnerships with stakeholder groups and highlight the importance of new ways of thinking, and sound empirical methods, for scholarship in BHR as well as policy work. As scholars, we bring a range of skill sets to our research. We work across the fields of political science, sociology, economics and mathematics. We research corporate activity carried out in the Global South, along with headquarters-based dynamics in the Global North. One of us edits a major journal in the human rights field that has developed standardized quantitative data-sharing and replication protocols (including development of a dedicated segment of the Harvard Dataverse site).[2] All of us are committed to working on research that has both scholarly and policy value.

This piece offers a glimpse at the benefits to be gained (and challenges along the way) of pioneering new research using the award-winning resources of a leading international non-governmental research organization, the BHRRC.[3] We offer a glimpse 'under the hood' of work in progress and invite the scholarly community to join us on the journey.

---

[1] Shareen Hertel, *Tethered Fates: Companies, Communities, and Rights at Stake* (Oxford: Oxford University Press, 2019).

[2] To access all datasets associated with articles published in *The Journal of Human Rights*, which include statistical modelling, see: https://dataverse.harvard.edu/dataverse/jhr (accessed 25 May 2020).

[3] The BHRRC was awarded the Thomas J Dodd Prize in International Justice and Human Rights in 2013, which highlighted the website's central role in the organization's mission. See Kenneth Best, 'Dodd Prize Highlights Business

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press

II. Early Phase: From Hand-Coding to API

The BHRRC began developing an API to efficiently access information on their website in early 2018. Web APIs can be an extremely useful tool for social science research: they provide users with a convenient, standardized method for handling large amounts of information that have been produced online or collected offline and then digitized. While social scientists and journalists have extensively used APIs in the past to interact with and collect data from international organizations (including United Nations agencies and the World Bank), government offices, newspapers and social media websites (such as Twitter and Facebook), the BHRRC's API was one of the first initiatives in the realm of human rights[4] and the first that can be used in the study of business and human rights.

The earliest version of the BHRRC's API, available as a Private Beta, was used to analyse trends in stakeholder consultation in Hertel (2019). Hertel and Majumdar were able to systematically travel through and analyse thousands of articles, reports, web references, company responses and other documents collected by the BHRRC since its founding in 2002 in a matter of seconds. Although this version of the API was not fully developed when we began using it in early 2018 (and we collaborated directly with BHRRC staff to refine the tool), it nevertheless constituted a remarkable improvement over our 'pre-API' strategy to hand-code the same data, as other members of our research team had done for upwards of two years (from 2016 to 2018). To put the efficiency of the API in context, we were able to create a complete and error-free dataset with over 11,000 observations that were necessary for our statistical analysis in just one weekend; our manual coding efforts yielded only a small fraction of that data after weeks of scrolling through the BHRRC website.

The data obtained through the API[5] is not difficult to navigate; in what follows, we provide a brief description of the data structure. In its most basic and accessible form, the API enables users to query the entire database and receive information in standard JavaScript Object Notation (JSON) format, which can then be read by any standard programming language or software used by social scientists (such as R, Python or STATA). The API is built around four types of objects that are central to how the BHRRC stores and presents its data: Categories, Companies, Components and Stories. As explained more fully in Hertel (2019, chapter 3), a 'Story' can be a stand-alone news article or report, or a collection of related articles, reports, statements and/or web references about a particular event involving a company's conduct towards human rights. Every Story has at least one 'Component' such as an NGO report, a media article, a lawsuit summary, or a company response, so that allegations are described from multiple sources and viewpoints.[6] Both Components and

Link to Global Human Rights', *UConn Today* (15 November 2013), https://today.uconn.edu/2013/11/dodd-prize-highlights-business-link-to-global-human-rights/ (accessed 14 May 2020).

[4] Some major national and international institutions such as the United States Department of Justice and United Nations agencies that promote human rights issues among other activities have APIs to access their collections. However, there do not appear to be any institutional APIs developed and released for the sole purpose of accessing as rich a set of human rights data as the records maintained by the BHRRC.

[5] The BHRRC has a website with instructions on how to use the API: https://www.business-humanrights.org/en/using-our-business-and-human-rights-data-to-bring-about-change (accessed 7 July 2020).

[6] For stand-alone articles and reports, the Story and Component are one and the same, whereas for collections, every item within the collection is a single Component and all of the Components fall under the umbrella of the Story.

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press

Stories can have 'Categories' and specific 'Companies' linked to them. Categories include countries, regions, sectors, principles, organizations and company policy steps. A researcher perusing the BHRRC database typically uses these 'Categories' (of which there are over 700 in total) to seek information on the issues they are interested in exploring.

Figure 1 shows a flowchart to illustrate the four types of objects one can retrieve from the API. The leftmost box contains the Story about worker injuries in India. This Story is linked to two companies that were involved (centre, top). The Story is also associated with a range of Categories such as the region, industry and human rights issue areas (centre, bottom). Finally, the box on the right illustrates the Components. Components include items such as such as NGO reports, media articles, lawsuit summaries, or if the BHRRC received a company response, which form the evidence base for the overall Story. The API returns identifiers for all these objects as well as metadata such as publication date and language. This structure allows users to match Components to Stories, Categories to Stories, Components to Companies, etc. This provides a great deal of flexibility in the granularity of analysis, as API users can retrieve lists of all Stories and Components linked to different combinations of Companies and Categories.

Drawing on the API to build a flexible dataset for analysis in itself offers a vast potential of options given the diversity of data types available such as categorical data, text extracts, entity associations and dates. However, there is also potential to integrate the BHRRC API with other databases and other APIs in order to both cross-validate data and expand the range of research options. Whilst there is not a specific key against which different datasets can be matched, there is potential for creative intersection of different sources. One example might be to draw on a source of news article text such as LexisNexis, and match stories to articles using company names, date ranges and keywords extracted using named-entity-recognition.

Equipped with this knowledge about the scope and structure of the data, one can easily register with BHRRC and start using the API for any number of purposes with minimal investment in terms of time and effort. In the next section, we will introduce ongoing efforts to interact with the latest version of the API to collect and analyse significantly larger amounts of data in a more sophisticated manner.

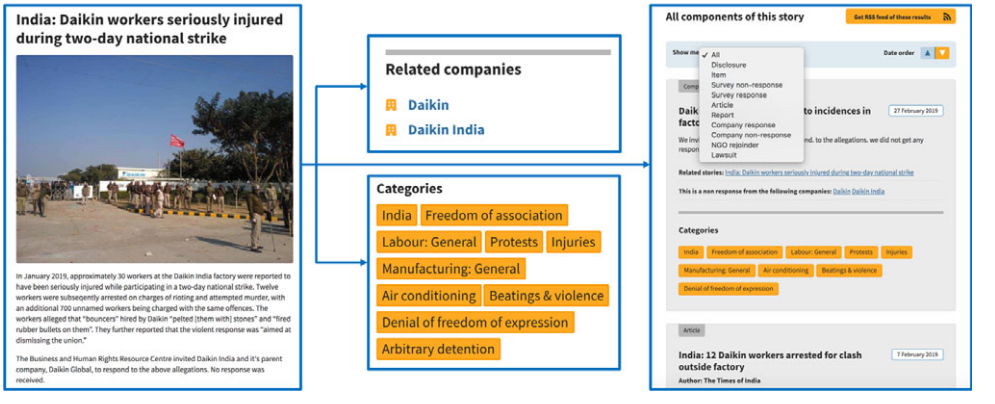

Figure 1. Flowchart of information that can be retrieved via the API

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press

III. Second Phase: Analysing and Visualizing Stories on the Big Data Scale

Janz and Allen-Robertson have begun utilizing the API infrastructure to push data collection to the 'big data' scale.[7] This allows us to perform systematic analysis of corporate human rights issues on an unprecedented scale. We have downloaded over 50,000 individual stories with roughly 70,000 components such as NGO reports or media articles that are linked to each story. We can now use the 'Categories' to map how rights allegations differ within and across industries, show central and peripheral issues, uncover clusters of rights that tend to be violated together, display the strength of relationship between industry activities and rights issues, and examine particular companies' allegations. We can also learn about NGO and media strategies in reports concerning firms' human rights abuse, and track where the public focus shifts over time as well as how global regions compare with respect to reporting trends.

For example, following the industry-sector framework presented in Janz (2018)[8], we are interested in the differences between human rights allegations against companies within and across industries such as natural resources, finance or textiles. The data contain stories from over 15 industries and include more than 20 different human rights issues tracked by the BHRRC over time (see Table 1). We filtered the original database's stories into these industry and human rights categories and counted frequencies of stories that fell into each group. In the next step, we examined how often a story is related to a particular industry (e.g., natural resources) and at the same time, a particular right (e.g., child labour). This has allowed us to visualize these rights/sector relationships as networks, indicating which human rights are commonly affected by firms in particular industries. We were also able to see which rights seemed to be commonly violated together (e.g., child labour and modern slavery) via these networks and additional clustered topic graphs. Finally, we can create timelines to map which types of human rights (in which industries) receive the most attention by NGOs and the media over time, and in which regions of the world (see Fig. 2). The opportunities to explore these data are endless, and the data can be used to confirm whether issues from anecdotal evidence and single case studies are representative of global patterns. We can also detect previously unknown trends and clusters of allegations that may not have received scholarly attention because they were previously 'hidden' by the sheer volume of information about companies' wrongdoings worldwide.

The main challenge involved in analysing this kind of data is under-reporting. The BHRRC only captures stories that NGOs and media publish, so that certain issues may remain undetected. Moreover, the stories themselves may miss critical nuance either in terms of the nature of corporate action or in the scope of stakeholders affected. Scholarship in the business and human rights literature can be affected by multiple

[7] Replication materials can be found at: https://dataverse.harvard.edu/dataverse/NicoleJanz

[8] Nicole Janz, 'Foreign Direct Investment and Repression: An Analysis Across Industry Sectors' (2018) 17:2 *The Journal of Human Rights* 163–183. See also Krishna C. Vadlamannati, Nicole Janz, and Indra de Soysa, 'U.S. Multinationals and Human Rights: A Theoretical and Empirical Assessment of Extractive Versus Nonextractive Sectors' (2020) Business & Society. First published online July 3, 2020.

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press

Table 1. Industries and types of human rights issues (2000–2019) based on BHRRC stories

Industry sectors:	
Apparel and textile	Media and publishing
Chemical	Military and security equipment
Construction and building materials	Natural resources
Consumer products and retail	Services
Finance	Shipping and handling
Health sector	Technology
Leisure	Transport
Manufacturing	Travel
	Utilities
Human rights-related issues:	
Abduction and disappearances	Group rights
Access to water	Health
Arbitrary detention	Injuries
Beatings, violence, torture, ill-treatment	Intellectual property
Cultural issues	Intimidation and threats
Denial of freedom of expression	Killings, death
Discrimination	Labour (forced labour, modern slavery, child labour, living wages, human trafficking)
Displacement of communities, land rights, denial of freedom of movement	Privacy
Education	Rape and sexual abuse
Environment	Right to food
Freedom of association	Sexual harassment
Genocide	Unfair trial

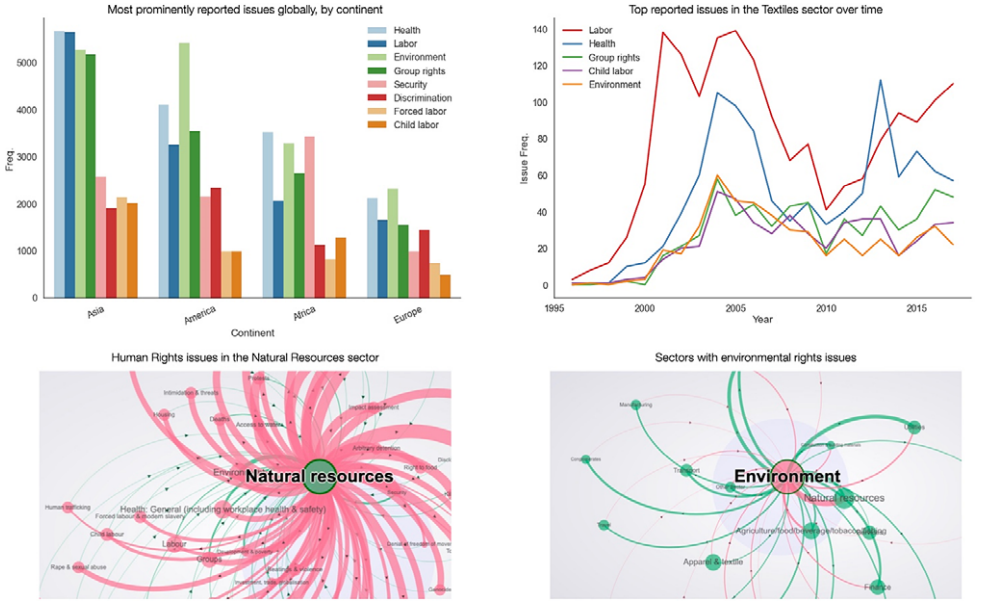

Figure 2. Global patterns in human rights issues based on BHRRC stories

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press

forms of under-reporting bias: companies have reason to avoid publication of their difficulties; hand-coders of the conventional human rights indices such as the Political Terror Scale always rely on 'reported' violations; and self-reporting tools for companies' corporate social responsibility activities are likely to be biased. However, the BHRRC mitigates this problem by placing researchers across the globe, who search for information in different languages from international and local news outlets as well as NGOs. The number of potentially omitted cases decreases with the sheer volume of information captured in the BHRRC data archive. Now that we can download all of these big data in bulk for analysis, the likelihood of missing particular stories still exists, but is considerably reduced compared with conventional data collection methods in the field. In addition, scholars whose research is informed by the BHRRC data can integrate data available through APIs provided by other research outlets, thus expanding the scope of coverage further. The volume of relationships that this approach enables us to explore and visualize – between companies, industries, geographies and patterns of violations – has the potential to enrich the field of business and human rights research dramatically.

IV. Implications for Scholarship and Practice

Our experience working with 'big data' illustrates both the academic value of an API (for example, in addressing omitted variable bias) and its practical value. Identifying problems with greater precision along with potential avenues for change on the ground can happen more quickly if we use the types of tools discussed in this article. The data visualization presented in Fig. 1, for example, conveys compellingly and quickly for the specialist or average reader alike the types of trends that the BHRRC data captures.

As an international non-profit, the BHRRC itself will no doubt continue to refine its data gathering, reporting, analysis and advocacy efforts over time. New technologies will no doubt emerge that enable further precision and enlarged scope in data collection, processing and analysis along with wider information-sharing with the public. Indeed, the richness of the BHRRC database is not only the millions of bits of information searchable quickly, but also the free and open nature of that information for public use.

When we began collaborating on this article, the COVID-19 pandemic was not even on the horizon. We do not yet have a full sense of how this or other global threats will shape decision-making across societies and in private sector entities over the long term. However, we can assume that researchers could use the API discussed in this article or similar tools to map corporate responses to the virus, or to track shifts in public concern about specific companies in relation to emerging technologies, supply chain disruption, or other issues relevant in this crisis. Our work here is an invitation to others to explore big data over the course of what will likely be a much longer arc of scholarship on business and human rights.

https://doi.org/10.1017/bhj.2020.18 Published online by Cambridge University Press