«Comunicazioni sociali», 2020, n. 3, 346-358

MATTEO TARANTINO*

BRIDGING THE GREEN DATASCAPE: DATA SCRAPING FOR SUSTAINABILITY PURPOSES

Abstract
The abundance of provision of environmental data and their diffusion on the Internet through idiosyncratic methods without unified standards for disclosure, has brought about a situation in which data is available but difficult to aggregate, synthesize and interpret. This article explores the roots and implications of practices of "scraping", i.e. automatic unauthorized collection of data published on the web, enacted by public and private subjects for the purposes of sustainability. Drawing from the concept of 'datascape' to describe the overall socio-technical environment this data circulates, the paper explores two case studies. The first, EDGI/DataRefuge, deals with a systematic attempt to collect and preserve environmental data and documents published by environmental management agencies, which is subject of cancellation by US Government policies. The second case, WorldAQI, examines a platform collecting, refining and publishing on maps the air quality indexes of hundreds of countries in the world. The first case allows us to see the human component of large-scale web scraping efforts of highly heterogeneous data, which highlights the need for resources. The second case highlights how the processes of collation, formatting and normalization of heterogeneous data to maximize readability have implications for data quality and representativeness. In conclusion, we observe how through data scraping stakeholders can enhance the spatial and temporal comparability of data and provide new avenues for public participation into complex decision-making processes.

Keywords
Information retrieval; scraping; harvesting; environmental data; trust.
ISSN: 03928667 (print) 18277969 (digital)
DOI: 10.26350/001200_000100

1. INTRODUCTION

This article will explore some of the implications of practices of unauthorized data extraction and aggregation from online sources ("scraping"[1]), performed by public and private actors for the purposes of sustainability. "Scraping" refers to the practice through which data not originally intended for offline use is downloaded – from websites or mobile applications – and aggregated. "Data" is a rather blurry concept in itself; for the

* Università Cattolica del Sacro Cuore, Milan – matteo.tarantino@unicatt.it.

[1] In the field, the terminology is still a bit imprecise. The concept of "scraping" is sometime referred to also as "harvesting", and in some contexts is included into "data mining". "Scraping" in turn tends to include "Crawling" or "Spidering", which are generally used to describe software designed to transverse HTML links (but not necessarily capture of any data).

purposes of our work, we'll define here "data" following Jennex in his revision of the literature on Ackoff's Data-Information-Wisdom hierarchy, as "basic, discrete, objective facts about something such as who, what, when, where" (2017, p. 70). Specifically, in the context of scraping, data refers to any information that is collected and stored from the websites.

Scraping is performed either manually, by copy/pasting data from websites, or, more frequently, by deploying automatic software which load up the required data source (typically a web page) and collect the necessary data. "Scrapers" have been defined by Marres and Weltwerde as "bits of software code that makes it possible to automatically download data from the Web, and to capture some of the large quantities of data […] that are available on online platforms"[2]. In doing so, they potentially ease the systematic and continuous collection of data. In a previous article focusing on scraping environmental data in China, we stressed that even after code is deployed, scraping can entail significant human labor, regarding not only the need for redesigns (in response to scraping-blocking measures from websites), but also the location of data and data validation after it is collected. Overall, scraping is a part of the family of 'data ingestion' methods, that is practices aimed to the acquisition and preliminary aggregation of data for further use[3]. Data ingestion represents a preliminary step for any data-related operation, and as such a significant actor in its success or failure. It is only after data is located, collected and ingested that it can be processed. However, as Leonelli's concept of "data journey"[4] and Edwards' notion of "data friction"[5] both indicate, data is often transformed as it moves across surfaces. As the growing literature on the sociology of data underlines[6], focusing on the mundane side of data operations, such as data ingestion, enables a deconstruction of the taken-for-granted nature of data revealing what Creel referred to as a 'deeply social nature'[7]. As 'social nature' we hereby mean the power, economies and social dynamics that materially shape the data upon which contemporary societies have come to rely more and more to describe their realities[8]. As Ruppert *et al.* observe, within this frame data is approached not as a representation of reality but as a social and political product, whose production is caught into a web of power relations which need to be examined to understand their role in broader social and economic processes[9]. Yet, within the field, scraping represents a still understudied practice.

As Marres and Weltwerde observe, scrapers are instruments to re-order dispersed

[2] N. Marres - E. Weltevrede, "Scraping the Social?", *Journal of Cultural Economy*, 6, 3 (2013): 313-335. See also R.N. Landers, R.C. Brusso, K.J. Cavanaugh - A.B. Collmus, "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research", *Psychological Methods*, 21, 4 (2016): 475.

[3] A good discussion on data ingestion can be found in J. Meehan, C. Aslantas, S. Zdonik, N. Tatbul, J. Du, *Data Ingestion for the Connected World*, 2017.

[4] S. Leonelli, "Learning from Data Journeys", in *Data Journeys in the Sciences*, edited by S. Leonelli and N. Tempini, Cham: Springer, 2020.

[5] J. Bates, "The Politics of Data Friction", *Journal of Documentation*, 74, 2 (2018): 412-429; P.N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, The MIT Press, 2010.

[6] See K.A. Creel, "Transparency in Complex Computational Systems", *Philosophy of Science*, 0, ja (2020); A.J. Lee, P.S. Cook, "The Myth of the "Data-Driven" Society: Exploring the Interactions of Data Interfaces, Circulations, and Abstractions", *Sociology Compass*, 14, 1 (2019); S. Leonelli, B. Rappert, G. Davies, "Data Shadows", *Science, Technology, & Human Values*, 42, 2 (2016): 191-202.

[7] Creel, "Transparency in Complex Computational Systems".

[8] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, Sage, 2014.

[9] E. Ruppert, E. Isin, D. Bigo, "Data Politics", *Big Data & Society*, 4, 2 (2017).

data into ordered datasets[10]. This proposition appears enticing with respect to environmental information, as wide public availability of digitalized environmental information since the 1990s has come with more than significative levels of disorder. Two lines of discussion in literature appear pertinent to describe this situation: the first regards availability of information; the second its degree of fragmentation. Concerning the former, as Gupta and Mason underscore, we are living in an age of unprecedented transparency[11], especially in the environmental domain. The steady reduction in the price/performance ratio of environmental data capture, processing and storage has rendered IT-based management of many environmental processes accessible to a wide range of public and private institutions. This high availability is also related to the fact that data generation in the environmental domain is being progressively more automated. In particular, sensors monitoring water, air and noise pollution as well as satellite imagery have been declining in price, allowing for wide accessibility. This accessibility has been consolidating what Arthur Mol, building on Manuel Castells, calls "informational environmental governance"[12]. In this mode, traditional rule-centered environmental governance is being increasingly replaced by multi-layered governance structures pivoting on access and exchange of information. In this scenario, as Kitchin as observed, the monopoly of central governments in the management of environmental information has been challenged[13] and multiple actors (including civil society, such environmental NGOs or even private citizens, and private sector players) have become important brokers of environmental data – acquiring, processing and disclosing information to pursue their own agendas[14]. Within global governance, this transparency turn[15] is closely related to the broader trend towards decentralization. Growing recourse to devolved governance structures (from financial markets to countries themselves) has been often accompanied by elicitation towards public and private entities to disclose information about their proceedings. The implications of this intertwining and manifold, but can be summarized as follows: while accommodating public demands for participation, transparency also improves the governability of decentralized environments by increasing the accountability of local actors *vis-à-vis* central ones, and thus reduces the necessity for strong centralized regulatory efforts[16]. The resulting 'panopticon effect' is alleged by some authors

[10] Marres and Weltevrede, *ibid*. The authors' reflection deals with social media data, but it can be applied to most domains where scrapers are used.

[11] A. Gupta, M. Mason, *Transparency in Global Environmental Governance: Critical Perspectives*, The MIT Press, 2014, and Id., "A Transparency Turn in Global Environmental Governance", in *Transparency in Global Environmental Governance: Critical Perspectives*, edited by A. Gupta and M. Mason, The MIT Press, 2014; M. Mason, "Transparency for Whom? Information Disclosure and Power in Global Environmental Governance", *Global Environmental Politics*, 8, 2 (2008): 8-13.

[12] A.P. Mol, "Environmental Governance in the Information Age: The Emergence of Informational Governance", *Environment and Planning C: Government and Policy*, 24, 4 (2006): 497-514; A.P. Mol, *Environmental Reform in the Information Age. The Contours of Informational Governance*, Cambridge (UK): Cambridge University Press, 2008 and "Environmental Governance through Information: China and Vietnam", *Singapore Journal of Tropical Geography*, 30, 1 (2009): 114-129; A.P. Mol, G. He, L. Zhang, "Information Disclosure in Environmental Risk Management: Developments in China", *Journal of Current Chinese Affairs*, 40, 3 (2011): 163-192. For the concept of "Informational Governance" see M. Castells, *The Rise of the Network Society*, John Wiley & Sons, 2011.

[13] E. Ruppert, E. Isin, D. Bigo, "Data Politics", *Big Data & Society*, 4, 2 (2017).

[14] A. Gupta, "Transparency under Scrutiny: Information Disclosure in Global Environmental Governance", *Global Environmental Politics*, 8, 2 (2008): 1-7.

[15] A. Gupta, M. Mason, "A Transparency Turn in Global Environmental Governance", in *Transparency in Global Environmental Governance: Critical Perspectives*, edited by A. Gupta and M. Mason, The MIT Press, 2014.

[16] Several commentators have stressed the need to take seriously the link between neoliberal modes of

to keep local actors (public and private) in check due to the increased visibility of their actions[17]; moreover, the complex nature of environmental issues has shown the limits of market-based and command-and-control governance methods, fostering what Florini has labeled "regulation by revelation"[18]. This decentralization has profound impacts on the way the data is disclosed.

The environmental domain has been particularly invested by the described drive towards transparency[19]. Public and private actors in the field have been widely recurring to the Internet as a platform for disclosure. However, disclosure efforts by public institutions (as well as those by private companies) are seldom coordinated and harmonized[20]. This phenomenon tends to happen at all scales from global (e.g. among countries) to local (e.g. among cities within the same region or province). This lack of harmonization can be due to differences in such factors as disclosure policies, technical means, knowledge capital, measurement techniques, standards, reporting strategies, political needs etc. The result is a highly inhomogeneous ensemble of data flows, released in different formats, through different outlets and at different times, with different levels of detail. To describe this scenario, we proposed a reframing of the concept of "datascape", which he used to designate the "space" traced by thematically-bound data traversing informational infrastructures; he also stressed the fragmentation of the "environmental datascape" and explored some of the costs that institutions run into when trying to aggregate data in a fragmented datascape[21].

One of the key variances determining this fragmentation format. As 'format' of data we hereby mean, according to the EU definition, "the way in which the data is structured and made available for humans and machines"[22]. Formats can possess varying degrees of machine-readability. More machine-readable formats tend to be more granular, reusable and re-processable. They are usually delivered either through downloads (for instance, of tabular files such as CSV or Excel files) or via software interfaces which can be interrogated directly by the user (through APIs or database connections). As such, this data is harder to understand for humans but can be fed easily into other software to be

production and the "transparency wave", wherein transparency supports 'light government' approaches to governance, which in turn tend to favor free-market agents. See for example Haufler's discussion on transparency in the extractive industries: V. Haufler, "Disclosure as Governance: The Extractive Industries Transparency Initiative and Resource Management in the Developing World", *Global Environmental Politics*, 10, 3 (2010): 53-73.To that, I would add that autocratic states appear to have also seized the opportunities offered by transparency to increase local resistances to central directives in check. This, of course, can also have positive implications, as is the case of China when re-energizing the traditionally dismal environmental performance of some local administrators through transparency. See L. Zhang, A.P. Mol, G. He, "Transparency and Information Disclosure in China's Environmental Governance", *Current Opinion in Environmental Sustainability*, 18 (2016): 17-24.

[17] For a critical approach to this phenomenon and the underlying ideology, see U. Golob, W.J. Elving, A.E. Nielsen, C. Thomsen, F. Schultz, K. Podnar, W.T. Coombs, S.J. Holladay, "The Pseudo-Panopticon: The Illusion Created by CSR-Related Transparency and the Internet", *Corporate Communications: An International Journal*, 18, 2 (2013):

[18] A. Florini, "The End of Secrecy", *Foreign Policy*, 111 (1998): 50-63.

[19] This drive has been ratified also at the supra-national level: the Aarhus convention, proposed by the UN Economic Commission for Europe, adopted in 1998 and enforced since 2001, binds 47 countries (to varying degrees) to release to the public environmental information. See M. Mason, "Information Disclosure and Environmental Rights: The Aarhus Convention", *Global Environmental Politics*, 10, 3 (2010): 10-31.

[20] J.B. Braden, M.C. Jolejole-Foreman, D.W. Schneider, "Humans and the Water Environment: The Need for Coordinated Data Collection", *Water*, 6, 1 (2014): 1-16.

[21] M. Tarantino, "Navigating a Datascape: Challenges in Automating Environmental Data Disclosure in China", *Journal of Environmental Planning and Management*, 63, 1 (2020): 67-86.

[22] https://www.europeandataportal.eu/elearning/en/module9/#/id/co-01. Accessed June 1, 2020.

aggregated, manipulated and transformed (for instance in graphs, maps etc.). Conversely, data disclosed in formats easier to read by humans (for example HTML web pages or PDF files) are more static and tend to have lower reusability. The underpinning data cannot be accessed directly; it must be extracted through dedicated processes. For instance, a user cannot easily take a table contained into a PDF document or in a web page and perform calculations on it; as a bare minimum, s/he will need to copy/paste the table into a spreadsheet (and many times this operation will fail or produce errors). This kind of format raises the costs for end-users wanting to perform any form of aggregation (for example for the purposes of comparison), transformation and processing.

In sum, while abundant in data, the environmental datascape at all scales tends to be characterized by a high degree of fragmentation and diversification; moreover, the cost of access is unevenly distributed, also depending on the choice in disclosure format. Given this scenario, using this data to inform scientific, economic or political action can be complicated, especially – but not only – for non-institutional actors. This can hinder meaningful action towards greater sustainability by both public and private entities. From this perspective, the fragmentation of environmental datascapes is also a matter of "data politics"[23], which see data as "generative of new forms of power relations and politics" and of "data justice"[24], that is the attempt to orient data production and consumption towards social justice. As Paul Edwards underscores:

> Data politics are not really about data per se. Instead, the ultimate stakes of how data are collected, stored, shared, altered, and destroyed lie in the quality of the knowledge they help to produce. The value of knowledge depends on trust in the infrastructures that create it, which themselves depend on trusted data[25].

In this scenario, scraping represents one of the practices through which public and private entities attempt to recollect and reorganize the data flows of this datascape for their own purposes. Sometimes scraping is the only way to obtain ordered, comparable data to inform action. Scraping represents therefore an important practice in environmental governance; yet, it often possesses very low visibility, both in academia and among practitioners. This is also because scraping is often prohibited by the websites' terms of use, which tend to frame them as unauthorized usages of contents. Breaching such terms can lead to legal action[26]. Therefore scrapers tend to lie in an ethical and legal gray area, and its usage is often not openly discussed, particularly by institutions[27].

[23] D. Beraldo, S. Milan, "From Data Politics to the Contentious Politics of Data", *Big Data & Society*, 6, 2 (2019): D. Bigo, E. Isin, E. Ruppert, *Data Politics: Worlds, Subjects, Rights*, Taylor & Francis, 2019; E. Ruppert, E. Isin, D. Bigo, "Data Politics", *Big Data & Society*, 4, 2 (2017).

[24] On the concept of 'data justice', which was originated by EDGI founder Michelle Murphy, see Taylor, "What Is Data Justice?".

[25] P.N. Edwards, "Knowledge Infrastructures under Siege", in *Data Politics*, edited by D. Bigo, E. Isin, E. Ruppert, London-New York: Routledge, 2019.

[26] See for example the case of Craiglist vs. PadMapper and Ryanair vs. PR Aviation. It must be said that in all these cases the plaintiffs were opposing the commercial reuse of data. See J.K. Hirschey, "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping", *Berkeley Tech. LJ*, 29 (2014): 897; G. Woodhead, "Could Ryanair Control Use of Its Flight Data by PR Aviation without Database Right?", *Journal of Direct, Data and Digital Marketing Practice*, 16, 3 (2015): 234-235.

[27] In practice, scraping is mostly blocked (or slowed down) via technical means. While formally scrapers should be coded to be perceived as such by servers (who can refuse them outright), most are coded to be perceived as human users. The relationship between scraper and scraped then becomes a game of mimicking: the scraper will try to imitate the behavior of a human being while trying to optimize the data harvesting; the server will try to detect the presence of a scraper from the traces it leaves in the server logs (such as: excessive

2. CASE STUDIES

Within the context outlined above, this article will attempt to explore some of the complexities entailed by the intertwining between web scraping and sustainability. This will be illustrated through two vignettes related to ongoing case studies, relating to scraping for preservation (par. 2.2.) and scraping for comparability (2.2).

For this exploratory analysis, we collected materials describing the activities of each involved entity: websites, social media accounts, data repositories, apps, existing interviews, news coverage have been collected and encoded using QSR's NVivo software. Categories employed in the analysis regarded the kind of data, purpose, timeframe of the scraping activity, destination of the scraped data, motivations, obstacles and connected activities.

2.1. *Scraping for preservation: EDGI and Data Refuge*

Scrapers can be designed to collect data before it disappears from the Internet, with the purpose of preserving its availability in the future. Content hosted on the World Wide Web, in particular, tends to be transient and impermanent: when web sites are updated, companies go out of business, entire web sites are erased and/or replaced. This raises important political and historiographical problems regarding the preservation of such materials, which some countries have been attempting to amend[28]. Outside of the institutional sphere, this question is addressed by projects such as the Californian NGO Internet Archive[29], which use scrapers to periodically collect and store snapshots of hundreds of thousands of websites. In this way, it enables public access to such copies via a specific interface called wayback machine[30]. There are two caveats to the Internet Archive's operations which are important to this case study. The first is depth limitation: because of the cost of storage, the Internet Archive stores only the first few levels of a website – akin to copying the cover of a book, the table of contents and perhaps one or two pages per chapter. Webpages beyond this surface level must be 'nominated' by users as important to be included in the scraping and then archiving. The second is content limitation: The Internet Archive stores static pages (i.e. pages which are stable and not served depending on user input) and basic web content (HTML code and images). Anything beyond this (particularly files linked in a webpage, such as reports or datasets) are either ignored or rendered incorrectly by the service.

Scraping-powered Internet archiving becomes politically relevant when sensitive

regularity of its actions, abnormal amount of information perused in the amount of time and so son). If it detects a scraper, it can react by slowing the connection to a crawl, shutting it off outright, blacklisting the scrapers' IP address or activating some sort of verification system such as CAPTCHA. Of course the server must balance those needs for security with the risk of false positive hindering the user experience of humans, which is a particularly critical aspect for private enterprises.

[28] For a review of the existing projects to preserve institutional digital information in the US, see: E. Johnson, A. Kubas, *Spotlight on Digital Government Information Preservation: Examining the Context, Outcomes, Limitations, and Successes of the Data Refuge Movement*, 2018; S.K. Lippincott, *Environmental Scan of Government Information and Data Preservation Efforts and Challenges*, Educopia Institute, 2018; L. Sare, "Providing Access to Government Information: A Survey of the Federal Depository Library Community", *Collaborative Librarianship*, 10, 3 (2018): 5.

[29] www.internearchive.com. Accessed on May 1, 2020.

[30] www.internetarchive.org. See E. Edwards, "Ephemeral to Enduring: The Internet Archive and Its Role in Preserving Digital Media", *Information Technology and Libraries*, 23, 1 (2004): 3-8 .

online public data starts being reduced by governments. This is the case of US governmental websites starting to reduce, remove or limit the visibility of information and datasets regarding climate change and other environmental issues starting from 2016, thus conforming to the shifting environmental priorities of the US Administration led by the 45th US President Donald J. Trump. In particular, the administration's skeptical response do climate change science was a source of great concern. Paul Edwards has called this shift as the last step in a decades-long siege on climate science.

> This new approach goes far beyond the questioning of knowledge *outputs* (conclusions, estimates of uncertainty, data analyses etc.). Instead, it seeks to delete important *inputs* to the knowledge-making process, such as key instruments required to monitor environmental change, and to remove climate change from the missions of agencies charged with environmental monitoring[31].

Amongst the responses to this situation, two data conservationist projects called Environmental Data & Governance Initiative (EDGI) and DataRefuge – originated in the US in November 2016. The initiatives are linked: DataRefuge emerged as a joint venture between EDGI, PPEH and Penn Libraries. set as their core mission data justice, and more specifically the preservation of 'vulnerable data' through the scraping and archiving of government institutions' pages and datasets, thus conserving information and documenting changes (change in terminology, page structure, etc.). Each of these websites and data repository tends to be idiosyncratic in terms of format and organization of data, thus requiring *ad-hoc* tools to be designed. To this end, both Data Refuge and EDGI established an alliance with the Internet Archive. On the one hand, it designed software (specifically, an extension to the Chrome browser) to semi-automate the nomination of pertinent institutional web pages for scraping, thus overcoming the Internet Archive's depth limitation and allowing the platform to scrape the majority or totality of institutional websites. On the other hand, it directly scraped elements that the Internet Archive could not scrape (such as dynamic tables) and transformed them into datasets. These scraping activities have been codified by DataRefuge into a workflow[32] with specific tasks and roles: "Seeders" identify datasets to be scraped ("uncrawable", in the organization's terms, by the Internet Archive); "Researchers" study of the websites where the targeted datasets are hosted, in order to inform the design of the scrapers; this information is fed to 'Harvesters' who materially create and run the scrapers; "Checkers" are tasked with quality control of the scraped data, which is then passed to "Describers" who write the metadata the dataset is archived with. EDGI then stored back the scraped data on the Internet Archive (in a separate folder and not integrated in the cached websites), while DataRefuge hosts its data on Amazon S3 servers and published a data catalog on its webpage. Both ways theoretically ensure continuing public access[33]. As of July 17, 2020, DataRefuge had successfully scraped 405 datasets, while EDGI's currently amount to 1072.

Both initiatives rely on and aggregate data coming from grassroots data conservation efforts, mobilizing academics, students and activists. All are engaged in workflows similar to the one described above. Between December 2016 and June 2017 DataRefuge

[31] P.N. Edwards, "Knowledge Infrastructures under Siege", in *Data Politics*, edited by D. Bigo, E. Isin, E. Ruppert, London-New York: Routledge, 2019.

[32] See https://datarefuge.github.io/workflow/, accessed on July 10, 2020.

[33] See EGDI's folder on the Internet Archive here: https://archive.org/details/EnvironmentalDataGovernanceInitiative and DataRefuge data catalog at https://www.datarefuge.org/dataset. Accessed May 1st, 2020.

organized 51 "Data rescue Events" in partnership with various institutions across the US. During such events, participants performed several data-related activities, including creating scraping programs[34], to collect data from government websites and databases. As stated by the press release of one of these events, "These harvested datasets will serve as a backup to the federal sites, ensuring the ongoing accessibility of the data through government sequestrations, government priority changes, and administration turnover"[35].

By writing *ad-hoc* scrapers tailored to collect systematically "at-risk" data, these initiatives shelter access to environmental information from the whims of political power. At the same time, by storing the points in time in which each item was accessible, they produce a material record of the shifts in discourse, priorities and objectives of the targeted institutions. This material record testifies to such political shifts in ways that go beyond statements, speeches or policy documents issued by regulators. Thus, in this instance, scraping reinforces the persistence of environmental data across time, and at the same time enables a materially focused historical perspective on sustainability politics. Furthermore, the case testifies to the necessity of huge manpower (voluntary, in this case) organized through an industrial logic to overcome costs of collection under extreme datascape fragmentation with information disclosed in unpredictable formats and within equally unpredictable website structures. This in turn raises the question of how to raise the resources to perform these operations. The two projects under examination are supported by wealthy American private and public entities[36] and were driven by the prestige of powerful institutions: in contexts with less available resources they might have never been able to take off.

2.2. *Scraping for accessibility and comparability: the World Air Quality Index*

In other cases, scraping is used to collect data from public or private sources which are scattered and disaggregated and recompile them into aggregated, homogeneous databases. This allows a general reduction of accessibility costs, as well as comparability across space (typically across localities) and/or time (typically tracking the performances of an entity over time).

The air quality datascape has always been particularly fragmented. In 1974, United Nations Environment (then UNEP) originated the GEMS/Air project precisely to assess air quality on a global scale; thirty-two years later, in 2006 in the GEMS assembly there was still a recognition of "large regional differences" and "data quality and intercomparability [being] a major issue"[37].

Amending this situation is the objective of projects such as the World Air Quality Index (WAQI). Originated in Beijing in 2007, the initiative continuously scrapes data

[34] See for example the announcements for the first event (December 4, 2016). https://technoscienceunit.org/2016/12/04/guerrilla-archiving-event-saving-environmental-data-from-trump/. Accessed on June 21, 2020.

[35] https://library.wustl.edu/participants-help-preserve-information-datarescuewu-event/. Accessed on May 12, 2020.

[36] DataRefuge is supported by Penn State University. EDGI as of 2020 receives financial support by the David & Lucile Packard Foundation and the Doris Duke Foundation, and technical supports from various tech corporations including Google Cloud and Amazon.

[37] T. Kjetil, V. Aasmund Fahre and L. Tarrasón, "GEMS air quality database: Potential sources of observational data and the efforts needed to make them accessible", paper presented at GEMS Annual Assembly, Reading, UK, 2006.

from 12,000 air quality monitoring stations (as of May 2020) across the world, and re-aggregates information into a single database, presenting a global, real-time map of air quality. The project started as AirQualityCn and was initially focused on aggregating information on Chinese air quality. Since 2013 it started publishing a map with air quality information from US consulates across China[38], subsequently evolving into an aggregator of air quality data at the global scale. In addition, the WAQI project offers a free Application Processing Interface which allows its aggregated data to be easily used by other projects – for example, weather apps – to show air quality information for a specific city. This enables citizens and institutions to perform quick and easy comparisons across different localities. In turn, comparability enables actors to put pressure on local institutions, companies or organizations to foster change.

This aggregation of data is not neutral. Air quality indexes (AQIs) are aggregate indicators which recompile discrete readings of single pollutants proceeding from sensors, summarizing disparate pollutants into a single number expressing the severity of air pollution. The recompilation formulas and breakpoints are decided at the national or sub-national level, and are thus unique for each country[39]. Moreover, different countries can monitor and release information of a different range of pollutants: for instance, Italian city Sesto San Giovanni does not release readings about PM_{10} but only about $PM_{2.5}$, whereas the neighboring city of Monza releases both. Therefore, air quality indexes are in themselves difficult to compare. However, most platforms share, along with aggregate indexes, discrete information about individual pollutants (usually Sulfur Dioxide, Nitrogen Dioxide, Ozone, CO_2, as well as particulate matter PM_{10}, and fine particulate matter $PM_{2.5}$) composing the index. To obtain comparability, the platform scrapes this discrete information from the web pages of local environmental protection agencies, and normalizes it, recalculating all air quality indexes using the American Environmental Protection Agency standard[40]. This choice is obviously not politically neutral and would not be feasible by supra-national organizations such as the UN[41]; but

[38] Air quality information in China and how to measure has represented a contentious issue, in particular between 2012 and 2015. US consulates started publishing information about air quality following a controversy with the Beijing's Environmental Protection Bureau, whose AQI readings had been consistently more positive than the US. Embassy ones, due to differences in pollutants measured and breakpoints in the national standards. The controversy is detailed in M. Tarantino, "In the Air Tonight: The Struggles of Communicating about Urban Environmental Quality", in *The Routledge Companion to Urban Media and Communication*, edited by Z. Krajina and D. Stevenson, Routledge, 2019.

[39] On the topic of air quality indexes see also J. Longhurst, "1 to 100: Creating an Air Quality Index in Pittsburgh", *Environ Monit Assess*, 106, 1-3 (2005): 27-42; Tarantino, "In the Air Tonight: The Struggles of Communicating about Urban Environmental Quality" and Id., "The Multiple Airs: Pollution, Competing Digital Information Flows and Mobile App Design in China", in *The Local and the Digital in Environmental Communication*, edited by J. Díaz-Pont, P. Maeseele, A. Egan Sjolander, M. Mishra, K. Foxwell-Norton, Springer, 2020.

[40] See the project's article about AQI comparisons at https://aqicn.org/faq/2015-03-20/a-comparison-of-worldwide-air-quality-scales-part-1/ (accessed May 14, 2020).

[41] The UN could perform the same operation using WHO's air quality guidelines for recalculation. However, no such project currently exists. UN-Environment is supporting WAQI from a technical standpoint, hosting the database on their servers. Another partnership that UN Environment has is Swiss company IQAir, which publishes the AirVisual project (https://www.iqair.com/world-air-quality), which is similar in scope with WAQI but also includes data from IQAir's own sensors and other similar bottom-up initiatives. IQAQI and UN-Environment are developing together the Urban Air Action Platform, which would represent the "largest real-time air quality databank, bundling real-time air quality data for particulate matter ($PM_{2.5}$) from thousands of initiatives run by citizens, communities, governments and the private sector." The project was launched on February 10, 2020. However, as of July 23, 2020, its website is no longer online.

it is feasible by a non-governmental organization. Effectively, WAQI reports air quality indexes which differ from those citizens can find in their own countries, thus effectively working as an alternative data source. In countries with laxer air quality standards than the American one, for instance, citizens can see worse readings on the platform than the one they obtain from official sources. Conversely, in countries with stricter standards, readings on the platforms may be better than the ones they see in official ones[42].

At any rate, through the aggregation the different cities can be effectively compared in terms of their air quality. The website opens on a world map showing the air quality of all monitored cities. Moreover, the platform offers cross-national comparison as well, ranking daily air quality averages for all the included countries.

However, air quality data retains a more difficult element to standardize: the different timeframes of sampling and disclosure. For example, Italian cities tend to release data taken 48 hours previous, whereas most Chinese cities release real-time hourly data. Things may vary even within countries, particularly federal ones: in Spain, Catalonian cities releases hourly data, whereas Andalusian cities release data once per day. This variance is not reconciled by the platform, which at first sight gives the mistaken impression of providing a snapshot of air quality. However, upon clicking on individual cities the timing of update appears.

While spatial comparability remains imperfect, through scraping this data acquires also a dimension of temporal comparability. The accumulation of scraped data and its normalization allows the platform to offer time series for each city (see Figure 2). Some of these series go back several years (for instance, data from stations in Beijing can go back as far as 2014; further data is available on request), thus enabling comparison of long-term historical trends within and across cities – comparisons which would otherwise be quite costly to perform.

Finally, WAQI complex aggregation processes also has implication on the reliability of scraped data. The WAQI project has hundreds of scrapers collecting data from each station, all running independently. Glitches and errors are possible. For instance, for the entire week of July 13, 2020, the discrete pollutants' data shown on the website for the station of Milan Senato did not match official sources. Effectively, a user (or a service) relying on WAQI to take everyday decisions based on air quality, would have been misled. Catching up to these errors can be slow and cumbersome for the platform. Since data does not proceed directly from sources but are effectively manipulated by a third party, the responsibility rests solely on the scraping institution. The legal boundaries of this responsibility, especially at the transnational level, are however not clear: glitches and malfunctions can always be invoked to justify mismatches in data. To this end, the website shows the following disclaimer:

> The World Air Quality Index project has exercised all reasonable skill and care in compiling the contents of this information and under no circumstances will the World Air Quality Index project team or its agents be liable in contract, tort or otherwise for any loss, injury or damage arising directly or indirectly from the supply of this data[43].

[42] For a discussion on how such differences can impact trust in this data see M. Tarantino, "The Multiple Airs: Pollution, Competing Digital Information Flows and Mobile App Design in China", in *The Local and the Digital in Environmental Communication*, edited by J. Díaz-Pont, P. Maeseele, A. Egan Sjolander, M. Mishra and K. Foxwell-Norton, Springer, 2020, which describes exactly this kind of situation with respect to the US and the People's Republic of China.

[43] https://waqi.info, "Usage Notice" Section, accessed June 1, 2020.

While the service these private entities provide to the environmental community is clear, a closer scrutiny of the power they wield in describing reality by aggregating data appears therefore desirable.

3. DISCUSSION

We have seen two cases in which a fragmented environmental datascape is subject to attempts of reordering through various combinations of human and technological agency. We can derive three main observations from this corpus.

The first is about the *ontology* of scrapers, and builds on our previous observations on the hybrid human/nonhuman nature of scraping labor[44]. As all our cases show, scraping cannot be simply equated with the coding of information-retrieval software automata. Rather, it entails a complex assemblage of human and nonhuman labor. This assemblage tends to become more complex as the target datascape complexifies. Environmental data, because of the features of its datascape in many geographical contexts, appears inclined to require such complex webs. From this basis, we propose an extension of the definition of scraping as a *formalized protocol for unauthorized acquisition of online data*. The key steps of such protocols appear to be the mapping of data (that is the identification of its location online); the design of continuous acquisition practices; the collation of data into machine-readable datasets; the verification of data quality. Any of these key steps can be performed by a combination of human and nonhumans, depending on the resources, the target datascape and the context. To reduce scrapers to mere software is to miss the complex economies, knowledges and communities that they can entail, as the DataRefuge/EDGI case has shown.

The second observation regards the *partiality* of the reordering brought about by the scrapers. In each of the cases compromises have been necessary to arrive to a finished dataset. In the case of tap water quality in Italy, those compromises have so far overwhelmed any attempt to aggregate information, which has been limited to mapping data. In the case of DataRefuge and EDGI, the collection of data has been made possible only through the aggregation of significant voluntary manpower and an alliance with a well-established organization like the Internet Archive. In the case of the World Air Quality Index, the aggregated data feature significant temporal, spatial and ontological disparities. In all these cases, the partiality emerges from the tension between the labor-intensive nature of scraping and the high levels of fragmentation of the datascape. Scraping for the environment may be necessary, but it can require massive resources.

The final observation regards the *political* dimensions of web scraping. Comparability of data in space and time, as provided by projects such as the World Air Quality Index, tends to put pressure on the social actors (both public and private) responsible for the situation the data captured. Conversely, choosing not to address at the institutional level the fragmentation of environmental datascapes can, in some contexts, provide cover to noncompliant actors. Specifically, the fragmentation of environmental datascapes can, in some instances, be beneficial (if, in most instances, not directly pursued by) to public and private actors trying to eschew environmental performance monitoring. Shielded by the broader positive discourse on decentralization and local autonomy, by enacting disclosure

[44] M. Tarantino, "Navigating a Datascape: Challenges in Automating Environmental Data Disclosure in China", *Journal of Environmental Planning and Management*, 63, 1 (2020): 67-86.

through their own schemes and patterns these actors manage to raise the cost of aggregation and thus discourage comparison, while formally fulfilling their obligations. The transient nature of web information further helps such actors, as data can disappear easily from the Internet – and in most cases, it is expected to. Hence, scraping can be seen as a societal response to an unjust state of things: the lack of spatial and temporal comparability in publicly disclosed data can disempower society with respect to extremely crucial environmental challenges. Aside from the action in itself (the political nature of which is explicit in the EDGI/DataRefuge case), also the specific rules of this act of reordering can be in themselves political. Choices in fields and records (what to scrape) testify to priorities; even more political are choices in collation and reformatting, as in the case of the WAQI project using US EPA standards (instead of any other standard) to recalculate all air quality information globally. Any of these choices impact the shape of the aggregated data, and as such can impact the range of narratives data itself can support.

These three dimensions can shape situations in which scraping environmental data is conceivably possible, but does not happen. For instance, aggregated tap water quality information tends to be rare in many countries. In Italy, for example, it is available only in highly disaggregated form. It is collected by local Health Protection Agencies (*Agenzia di Tutela della Salute*, ATS)[45], and by local water suppliers. Some of these actors release no data at all; some allow the consultation of a single sampling point at a time through a web interface; others allow the downloading of periodic PDF reports; others allow the download of datasets in .csv format; yet others provide data only upon official request. As of 2020 there is no public aggregated national water quality database. Partial attempts to aggregate the data include the website cheacquabeviamo.it, which has been hosting since 2007 a list of links to 169 water suppliers' pages containing water quality data covering 354 townships (over 7914 total) across all 20 Italian regions[46]. Such attempts reduce the cost of locating the data, but still leaves all costs related to collation and integration to the end-user[47].

This example represents an effective illustration of the consequences of a highly fragmented datascape: due to uncoordinated data release, the operational cost of collecting and comparing water quality across cities or regions in Italy remains high, thus disempowering any emerging initiative regarding water and depriving citizens of a comprehensive, granular perspective on drinkable water quality.

4. CONCLUSIONS

In this article we examined the practice of web scraping of environmental data through three case studies characterized by different scopes, organization and outcomes. Scrap-

[45] See for example the latest available report (2018) by the Milan ATS at https://www.ats-milano.it/Portale/Portals/0/AtsMilano_Documenti/A173-MS002%20Relazione%20acqua%20potabile%202018%20rev01_22ccc2f0-1d14-4202-818c-35cf0dbbc68c.pdf. Accessed on June 1, 2020. For comparison, the ATS of the neighboring city of Bergamo offers no comparable report.

[46] www.cheacquabeviamo.it. Accessed on June 1, 2020. The website is cited in the *Synthesis Report on the Quality of Drinking Water in the Union Examining Member States' Reports for the 2011-2013 Period, Foreseen under Article 13(5) of Directive 98/83/EC*, published by the European Commission in 2016, as the reference for drinkable water quality information in Italy. Yet public interest appears limited: as of July 15, 2020, the website reported a little over 1.2 million total views (1,262,789 to be precise), thus averaging less than 100,000 views per year.

[47] Attempts to collate the data by the website have been discontinuous and very limited in scope: for instance, the website published aggregated water quality data for the ten biggest Italian cities only until 2013.

ing of environmental data has emerged as a non-trivial, complex and labor-intensive operation, possessing considerable political implications. It emerges as an important practice in a world in which datascapes are more and more fragmented but environmental challenges are more and more transboundary. If aggregation of data through institutional channels – i.e. institutions formally agreeing to pool data and producing shared databases – can be slow and cumbersome, scraping emerges as a viable solution, bridging areas of the datascape which would be otherwise disconnected. These bridges can empower citizens and other entities in the pursuit of greater sustainability. At the same time, the power dynamics involved in the scraping of environmental data (and in general of public interest data) are still in need of closer examination. Moreover, it would be worthwhile to examine the conditions under which environmental data scraping should be acknowledged as a public interest practice – at the national and supranational level – and thus should be financed by public funds.