J Public Health Pol (2018) 39:4–11
https://doi.org/10.1057/s41271-017-0106-8

GUEST EDITORIAL

ToxicDocs (www.ToxicDocs.org): from history buried in stacks of paper to open, searchable archives online

David Rosner[1,2] · Gerald Markowitz[1,3] · Merlin Chowkwanyun[1]

Abstract As a result of a legal mechanism called *discovery*, the authors accumulated millions of internal corporate and trade association documents related to the introduction of new products and chemicals into workplaces and commerce. What did these private entities discuss among themselves and with their experts? The plethora of documents, both a blessing and a curse, opened new sources and interesting questions about corporate and regulatory histories. But they also posed an almost insurmountable challenge to historians. Thus emerged ToxicDocs, possible only with a technological innovation known as "Big Data." That refers to the sheer volume of new digital data and to the computational power to analyze them. Users will be able to identify what firms knew (or did not know) about the dangers of toxic substances in their products—and when. The database opens many areas to inquiry including environmental studies, business history, government regulation, and public policy. ToxicDocs will remain a resource free and open to all, anywhere in the world.

Keywords ToxicDocs · Environmental health · Occupational health · Public health

Our interest in the documentary history of toxic agents began in the 1980s when we (at that stage Markowitz and Rosner) discovered a collection of documents at the United States (US) National Archives that detailed the history of the fuel additive tetraethyl lead and the controversy that surrounded its introduction into gasoline in

✉ Gerald Markowitz
gmarkowitz@jjay.cuny.edu

[1] Mailman School of Public Health, Columbia University, New York, NY, USA

[2] Department of History, Columbia University, New York, NY, USA

[3] John Jay College and Graduate Center, City University of New York, 524 West 59th Street, Room 6.65, New York, NY 10019, USA

the 1920s. Rosner and Markowitz described their findings in an 1985 article published in the *American Journal of Public Health* [1].

Discovering discovery

Soon, Neil Leifer, a Boston lawyer pursuing environmental lawsuits, invited the two of us to visit his office. He had collected some documents about lead poisoning in children who had ingested or breathed lead-based paint in their homes. We traveled from New York to Boston thinking we would spend the afternoon looking for historically interesting materials in the "few" boxes mentioned. Entering the office suite quickly put to rest any thoughts about a quick research trip. Lining the walls and hallways of the law firm, Thornton Naumes, sat literally scores of banker boxes filled with corporate documents from paint and pigment companies, and their trade associations.

Thus, we two public health historians found ourselves drawn into the legal wrangling surrounding childhood lead poisoning. Historical research became critical to debates over responsibility for harm. For historians, a legal mechanism called *discovery*, where plaintiffs and defendants are required to exchange information and documents relevant to a lawsuit, made all the difference. While historians had always depended upon what was available through public sources such as newspapers, personal and corporate archives, and government records—now they could gain access to the inner workings of corporations. When corporations introduced new products into workplaces and commerce, often incorporating toxic agents, what did they discuss among themselves and with their experts?

Over the next two decades, millions of pages of internal corporate documents overwhelmed our academic offices in New York. In about 1998, a group of lawyers asked us to look at an enormous store of chemical industry documents that companies (defendants in a lawsuit) had turned over to them because they represented a vinyl chloride worker dying from an angiosarcoma of the liver, an extremely rare cancer. They, the plaintiff's lawyers, wanted help to evaluate, based on the accumulated documents, whether or not there was reason to believe that the chemical industry had acted in bad faith with regard to their workers' safety.

We made our way to Lake Charles, Louisiana, to review the millions of pages of corporate and trade association records stored in a split-level house in a desolate part of town. So began a process that resulted in a 300-page timeline of the Manufacturing Chemists Association (MCA) knowledge and activities. (Now renamed the American Chemistry Council.) The work was literally 'done by hand'—leafing through boxes one document at a time, photocopying, and sending the copies to our offices in New York. The documents we catalogued contributed to many lawsuits against the chemical industry.

Millions of documents

Over the years, we researched and wrote about occupational and environmental disease in *Deadly Dust* [2], in *Deceit and Denial* [3], and in *Lead Wars* [4]. Our analyses and documents helped shape many lawsuits. New stores of documents have

been sent our way, sometimes in paper but now more often in digital form. In addition to lead, we received thousands upon thousands of documents related to silicosis, polychlorinated biphenyls (PCBs), and asbestos. Minutes of trade associations' meetings contained detailed discussions of technical and political problems associated with products. How should industry deal with regulatory agencies at the state and federal levels?

In these records, one could detect the role of public relations firms hired by corporations to promote their products and to protect the industry. Every uncertainty in the science—the predictive value of animal studies, the epidemiology, what constitutes a safe level—was used to cast doubt on dangers. Trade associations and corporations tried to influence scientific research, sometimes trying to intimidate researchers whose work was deemed threatening to their products [5].

Having so many millions of documents was both a blessing and a curse. While they opened up new sources and interesting questions about corporate history, they also posed a very practical and almost insurmountable challenge to our generation of historians. Throughout most of our careers, documents were sorted by hand and read page by page, as we had done for our lead studies and for the Manufacturing Chemists Association documents. Our exhaustion from the seemingly endless search through the papers limited what we could accomplish.

But in the early 2000s that was about to change. A college student at Columbia, Merlin Chowkwanyun (a co-author of this piece and now a professor at Columbia), happened upon our work. In it, we had relied on very large collections of documents. He explored a modern approach to overcome historians' previous limits: how might computer search engines aid scholars facing the vast amounts of data that were becoming available through discovery? Could one create a website to which one could post a cache of documents related to our book, *Deceit and Denial?* The book had come under attack by the chemical industry, specifically by an historian who questioned our sources. The idea was simple: make the sources we used public so that historians and other scholars could evaluate the accuracy and truthfulness of our analysis. We named that first effort *deceitanddenial.org.* (It is no longer posted.)

The possibilities for research and for open access expanded dramatically with new technologies. We found the idea of sharing our primary sources and millions of original pages with students, scholars, and others interested in environmental and occupational health issues particularly exciting. And from this excitement emerged our new online database: www.ToxicDocs.org.

Big data technologies

The ToxicDocs project makes available at a dedicated website (www.ToxicDocs.org) information from a trove of documents unearthed from corporate vaults. They came from small brake manufacturers to multinational giants like Monsanto. ToxicDocs would not be possible without technological innovation commonly referred to as 'Big Data.' By Big Data, we refer to two major trends:

- the sheer volume of new digital data produced, and
- the computational power available to analyze them.

Onerous tasks that would have taken months to complete on a personal computer can now be broken up into smaller, parallel jobs and distributed across thousands of machines. Metaphorically, this is the difference between building automobiles without an assembly line versus with one. New types of data have been digitized and made available for researchers, including enormous collections of text we assembled. Data from these collections have been used, for example, in studies of narrative structures in novels or studies of changing word use over time. The so-called 'high-performance computing' makes the prospect of analyzing huge amounts of data far less intimidating.

ToxicDocs leverages these developments. One product is a novel document processing infrastructure that underlies and solves many of the problems intrinsic to working with enormous numbers of documents. Documents often came to us in a format that is not machine readable. Digitized documents contained *images* of alphanumeric characters but not ones recognized as such by a computer. This prohibited full-text searches and more complex computer-assisted analyses of the text.

We were able to solve this problem using optical character recognition (OCR), the process of turning scanned text from images of text to actual text (so that the image "A" is actually recognized as the letter "A"). Unfortunately, OCR is an onerous process. It might take months using an off-the-shelf OCR program on a personal computer to convert a few million pages. High-performance computing has permitted us to cut OCR times down to just a few days. We learned about and applied techniques developed at the University of Wisconsin-Madison Center for High Throughput Computing (CHTC). These techniques have made our documents full-text searchable easily and quickly.

ToxicDocs also contains a simple but sophisticated user interface for navigating through the site. It allows visitors to retrieve documents quickly, sorting through material using full-text searches with category filters (such as names of organizations, countries, or toxic substance type). When a user types a query, he or she sees thumbnail images that provide a birds-eye snapshot of a whole document. The user can then expand a particular document of interest, page through it online, download it, or bookmark it for later use. Due to our dissatisfaction with other document retrieval interfaces that lacked simplicity, contained bloated features, and worse, operated at sluggish speeds, we decided early to create our own user interface from the ground up. We built the ToxicDocs user interface with speed in mind; a typical search result takes less than one second to appear.

What are some examples of documents that a user might find? In our collection of polyvinyl chloride documents is a 1973 memo from the Vice President of the Manufacturing Chemists Association to the companies sponsoring the MCA's Vinyl Chloride Research Program. The memo discussed an impending meeting with the US National Institute for Occupational Health and Safety (NIOSH). He wrote that withholding the results of a European toxicological study on the carcinogenic

properties of vinyl chloride monomer might be interpreted as "evidence of an illegal conspiracy by industry."

In our asbestos collection, there is an October 1972 memo to the Executive Committee of the Asbestos Information Association (AIA), a major trade group. Executive Secretary Matthew Swetonic described a proposed AIA "Employee Safety & Health Guide" on asbestos. It reflects an attempt on the part of AIA to shape the discussion on asbestos safety before non-industry entities did so. "The point is simple," he wrote. "If we don't inform our employees, somebody else will!" Like many documents at www.ToxicDocs.org, this memo displays industry response to impending regulation and growing attention to the danger of their products.

The list below suggests the breadth of the current collection:

- An extensive collection of documents on PCBs, from Monsanto's own corporate archives.
- Documents on asbestos, from dozens of firms, and from the records of the Asbestos Industries Association (AIA) and the Friction Materials Standards Institute (FMSI), two industry interest groups.
- Documents on polyvinyl chloride and petrochemicals, from firms that include Texaco, Union Carbide, British Petroleum, 3 M, Dow Chemical, and Du Pont, as well as from chemical industry trade associations.
- Hundreds of thousands of documents on lead, especially lead paint, and the records of the Lead Industries Association (LIA), the industry's leading trade association.
- Documents from the United States, including US researchers doing studies in China, from Shell and Texaco on benzene as well as documents on a variety of toxic substances from France, Russia, Brazil, Chile, Argentina, Mexico, and England.
- Documents on silica and silicosis in both the United States and in South Africa.

ToxicDocs' consequences for public health

The information available in an organized and searchable home, ToxicDocs, has implications for public health. 'Analog data dumps'—endless stacks of cardboard boxes, folders, and dusty papers—now seem much less formidable. Rapid digitization and OCR conversion make text documents searchable and accessible to anybody with a keyboard.

ToxicDocs is growing and will soon add inductive search tools to make document location all the more intuitive and easy:

- A document classification tool will automatically sort documents into discrete types (such as an e-mail, an internal memo, a published scientific article, among others).
- A named entity recognition tool will allow users to pry out frequently occurring names, organizations, and countries that show up within user-defined sets of

documents, allowing users to tease out network relationships that are hard to spot when simply reading documents one by one.
- A text time-series analyzer allows one to see the rise and fall of certain phrases across time.

(See https://www.ToxicDocs.org/blog/welcome-to-toxic-docs-alpha/ for more detail.)

Taken together, ToxicDocs permits users to identify what firms knew (or did not know) about the dangers of toxic substances in their products—and when. Users can explore how firms responded publicly and to regulators at the time of major benchmarks in toxic substances regulation, such as the creation of the United States Environmental Protection Agency (1970). One can also reconstruct the status of scientific knowledge about various toxic substances' risk to human health. These are just some of dozens of broad questions one can ask of our dataset.

ToxicDocs comes at a critical time for environmental health policy. The United States updated the Toxic Substances Control Act of 1976 by enacting in 2016 the Chemical Safety for the 21st Century Act. Other industrial countries are developing ways to regulate chemical exposure such as the European Union's REACH regulatory program promulgated in 2006.

And industrializing countries are trying to avoid the mistakes of others. Everywhere the question remains whether government and public health regulators should be asked to ferret out safety risks or whether industry itself should be obligated to study and make public what they know about their processes and products, and to show that those processes and products do not present a risk to workers or the public. This is part of a larger debate over the 'precautionary principle' and how much of it to realize in regulatory practice [6].

Away from rarefied policymaking bodies and closer to the ground, we hope ToxicDocs will become useful to community health advocates who will now have a strong evidentiary base for raising questions about industrial firms' behavior in their communities. The United States' environmental justice movement began as a response to regional-level ecological threats—protests against sites selected for industrial manufacturing plants in Louisiana and agitation against diesel-fueled public buses in New York City. (See for example, *Dumping in Dixie: Race, Class, and Environmental Quality* [7], *Uneasy Alchemy: Citizens and Experts in Louisiana's Chemical Corridor Disputes* [8], and *Noxious New York: The Racial Politics of Urban Health and Environmental Justice* [9].).

Much of the ToxicDocs collection is about the US scene. It also includes a wide range of documents that speak of activities by the same companies in other countries as well as by companies and trade associations in those countries. One extensive collection, generously provided by Jock McCulloch, is from South Africa, specifically focused on the mining industry, asbestosis, and silicosis. (He has written one of the Commentaries included in the collection accompanying this Guest Editorial [10].). Anybody entering geographic terms is bound to come up with interesting information. We hope, too, that the site helps dilute unsubstantiated claims that occasionally pop up in environmental health circles, for example, over genetically modified foods, vaccines, or "chemtrails" in air. The website www.

ToxicDocs.org shows that more than enough empirically sustainable claims exist so that advocates can avoid more speculative proclamations.

Most importantly, ToxicDocs will remain a resource free and open to all, anywhere in the world: investigative journalists, toxicologists, policymakers, historians of public health like ourselves, environmental justice advocates, and the general public.

References

1. Rosner D, Markowitz G. A 'gift of God'?: The public health controversy over leaded gasoline during the 1920s. Am J Public Health. 1985;75:344–52.
2. Rosner D, Markowitz G. Deadly dust: silicosis and the on-going struggle to protect workers' health. Ann Arbor: University of Michigan Press; 2006.
3. Markowitz G, Rosner D. Deceit and denial: the deadly politics of industrial pollution. Berkeley: University of California Press; 2002.
4. Markowitz G, Rosner D. Lead wars: the politics of science and the fate of America's children. Berkeley: University of California Press; 2013.
5. Rosner D, Markowitz G. Standing up to the lead industry: an interview with Herbert Needleman. Public Health Rep. 2005;120:330–7.
6. Rosner D, Markowitz G. Industry challenges to the principle of prevention in public health: the precautionary principle in historical perspective. Public Health Rep. 2002;117:5012–512.
7. Bullard RD. Dumping in dixie: race, class, and environmental quality. Boulder: Westview Press; 1990.
8. Allen BL. Uneasy alchemy: citizens and experts in Louisiana's chemical corridor disputes. Cambridge: MIT Press; 2003.
9. Sze J. Noxious New York: the racial politics of urban health and environmental justice. Cambridge: MIT Press; 2007.
10. McCulloch J. Archival sources on asbestos and silicosis in Southern Africa and Australia. J Public Health Policy [special section] "ToxicDocs: Opening a new era of evidence for policies to protect public health" (Guest Eds. Rosner D, Markowitz G, Chowkwanyun M). 2018;39(1). https://doi.org/10.1057/s41271-017-0110-z.

David Rosner, M.S. (Public Health), Ph.D., is Ronald Lauterstein Professor of Public Health and Professor of History at Columbia University and Co-director of the Center for the History and Ethics of Public Health at Columbia's Mailman School of Public Health. He is the 2014 recipient of Sigma Xi's John P. McGovern Science and Society award and of a John Simon Guggenheim Fellowship and recipient of numerous grants from private and federal agencies, including the Milbank Memorial Fund, National Endowment for the Humanities, and the National Science Foundation. In addition to his works co-authored with Gerald Markowitz, he has authored and edited *"Hives of Sickness": Epidemics and Disease in New York*, New Brunswick, NJ: Rutgers University Press, 1995 and *A Once Charitable Enterprise: Hospitals and Health Care in Brooklyn and New York*, New York: Cambridge University Press, 1982. He is a member of the US National Academy of Medicine and a member of the *JPHP* Editorial Board.

Gerald Markowitz, Ph.D., is Distinguished Professor of History at John Jay College of Criminal Justice and the Graduate Center, City University of New York, and Adjunct Professor, Socio-Medical Sciences, Mailman School of Public Health, Columbia University. He is the recipient of numerous grants from private and federal agencies, including the Milbank Memorial Fund, National Endowment for the Humanities, and the National Science Foundation. Together with David Rosner, he has authored and edited books and articles on the history of public health, environmental health, and occupational safety and health, including those cited in the text and *The Contested Boundaries of American Public Health*,

New Brunswick: Rutgers University Press, 2008, with James Colgrove and David Rosner; *Are We Ready? The Public Health Response to 9/11*, Berkeley: University of California Press, 2006; *Dying for Work*, Bloomington: Indiana University Press, 1987; and *Slaves of the Depression: Workers' Letters about Life on the Job*, Ithaca: Cornell University Press, 1987. He is a member of the US National Academy of Medicine and a member of the *JPHP* Editorial Board.

Merlin Chowkwanyun, M.P.H., Ph.D., is an Assistant Professor of Sociomedical Sciences at the Mailman School of Public Health, Columbia University. He was previously a Robert Wood Johnson Foundation Health & Society Scholar at the University of Wisconsin-Madison. He works on the history of public health, race, and social movements around health.