ELEPHANT
IN THE LAB

OPINION

When your journal reads you – user tracking on science publisher platforms.

Short title	When your journal reads you
Long title	When your journal reads you – user tracking on science publisher platforms.
Authors	Renke Siems [1]
Author affiliation	[1] Ministry of Science and the Arts, Baden-Württemberg, Germany
Author bios	Dr. Renke Siems currently works at the Ministry of Science and the Arts in Baden-Württemberg. Previously, he worked in academic libraries for twenty years. He studied sociology, German Studies, and media studies. The article represents his personal views.
Author social links	LinkedIn
Date published	14 April 2021
DOI	10.5281/zenodo.4683778
Cite as (APA)	Siems, R. (2021). When your journal reads you – user tracking on science publisher platforms. *Elephant in the Lab*. DOI: https://doi.org/10.5281/zenodo.4683778

Introduction

In December 2018, a University of Minnesota web librarian, Cody Hanson, participated in a workshop hosted by the Coalition for Networked Information. The topic of this, and a number of other events to date, is the drive by major scholarly publishers to more fully integrate authentication systems for accessing electronic media into their platforms. Under various labels such as "Research Access 21 (RA21)", "Seamless Access", or "Get Full Text Research (GetFTR)", they want to replace the authentication options previously supported by libraries and academic institutions, such as IP range activation, VPN, proxy servers, or anonymous authentication to neutral third parties, as with the Shibboleth service, in favor of their own initiatives.[1] For years, librarians have countered these moves with their concerns that it will undermine the privacy of their users. Even at the event where Cody Hanson sat, the discussion raged until Todd Carpenter of the National Information Standards Organization (NISO) intervened, first correctly noting that RA21 does not require personally identifiable information (PII) to be sent to the publisher for authentication to occur. In fact, services like RA21 and Shibboleth share the same technical basis of a single sign-on via the Security Assertion Markup Language (SAML) - it's just that the technical realization is different. But then, to calm the discussion, Carpenter added, "that publishers don't need PII from RA21 to be able to identify library users." And that statement, of course, was absolutely designed to allay any concerns.

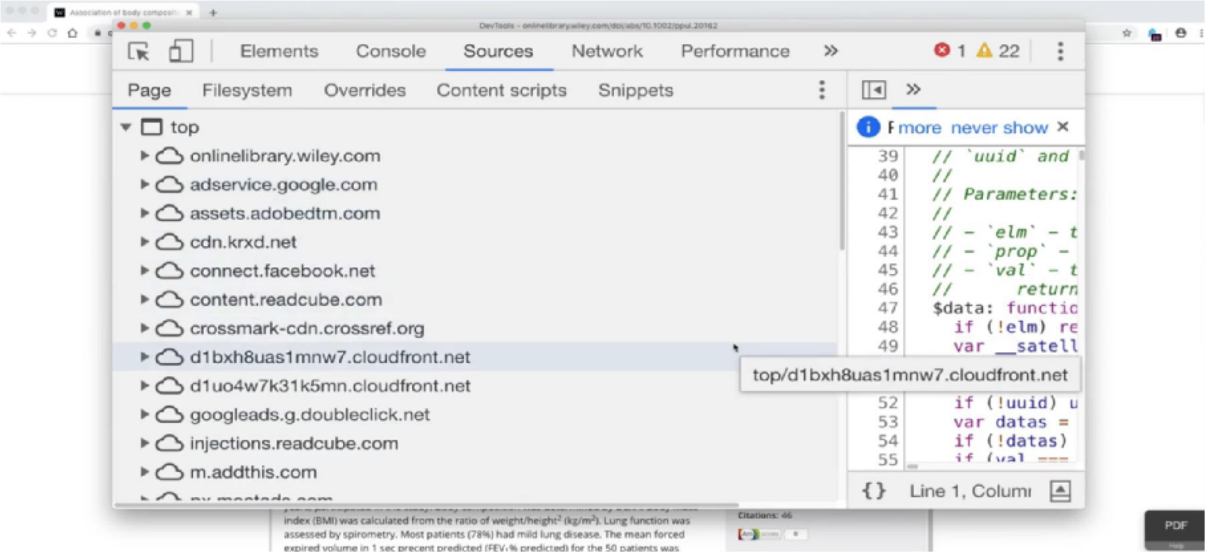
Cody Hanson, who spent a lot of time developing privacy-compliant access to electronic media for his users, began to wonder. Should this be true and can an analysis of the source code of publisher platform pages provide evidence of if and how publishers can identify library users? Cody Hanson undertook a testing exercise: he took the 100 most-demanded documents at his university and looked at them to see which platforms were represented in them. He picked one document from each of the fifteen platforms found, examined it with the Ghostery browser addon, and downloaded the document page to examine the source code. Several thousand lines of JavaScript later, Cody Hanson came to a simple answer: yes, it's true, Carpenter was right.

Of the fifteen platforms examined, one was clean (InformPubsOnline); on the others, he found a total of 139 different third-party asset sources. AdTech's entire technical assortment was represented: simple trackers, audience tools like Neustar, AddThis, Adobe, and Oracle, and fingerprinters like Doubleclick. These finds were significant to Cody Hanson:

“The reason I was interested in third-party assets being loaded on these sites is that any JavaScript loaded on these pages has access to the entire DOM, or document object model, meaning it can read the address and contents of the page. It also has access to every user action that happens on that page, and can itself load additional scripts from additional sources. So when, for example, a publisher puts JavaScript from Google on its pages, Google can record any information from the page about the article being sought, or search terms from a library user in the publisher platform. Fourteen of the fifteen publisher platforms included Google code on the article page.“[2]

Facebook code was also represented in many cases, as were a number of other data collectors, which means that patron privacy is no longer a given. Personalized profiles of the information behavior of every scientist are created, and since the publishers involve both the large Internet corporations and the audience tools as large data collectors, the data does not remain with the publishers, but flows out and can be linked with the knowledge that already exists elsewhere about the person. A seamless and thus valuable and tradable online biography of every scientist is created; the previous special milieu of science communication has been incorporated into the general commercial (and governmental) surveillance of the digital space.

Researchers at risk



top
- onlinelibrary.wiley.com
- adservice.google.com
- assets.adobedtm.com
- cdn.krxd.net
- connect.facebook.net
- content.readcube.com
- crossmark-cdn.crossref.org
- d1bxh8uas1mnw7.cloudfront.net
- d1uo4w7k31k5mn.cloudfront.net
- googleads.g.doubleclick.net
- injections.readcube.com
- m.addthis.com

top/d1bxh8uas1mnw7.cloudfront.net

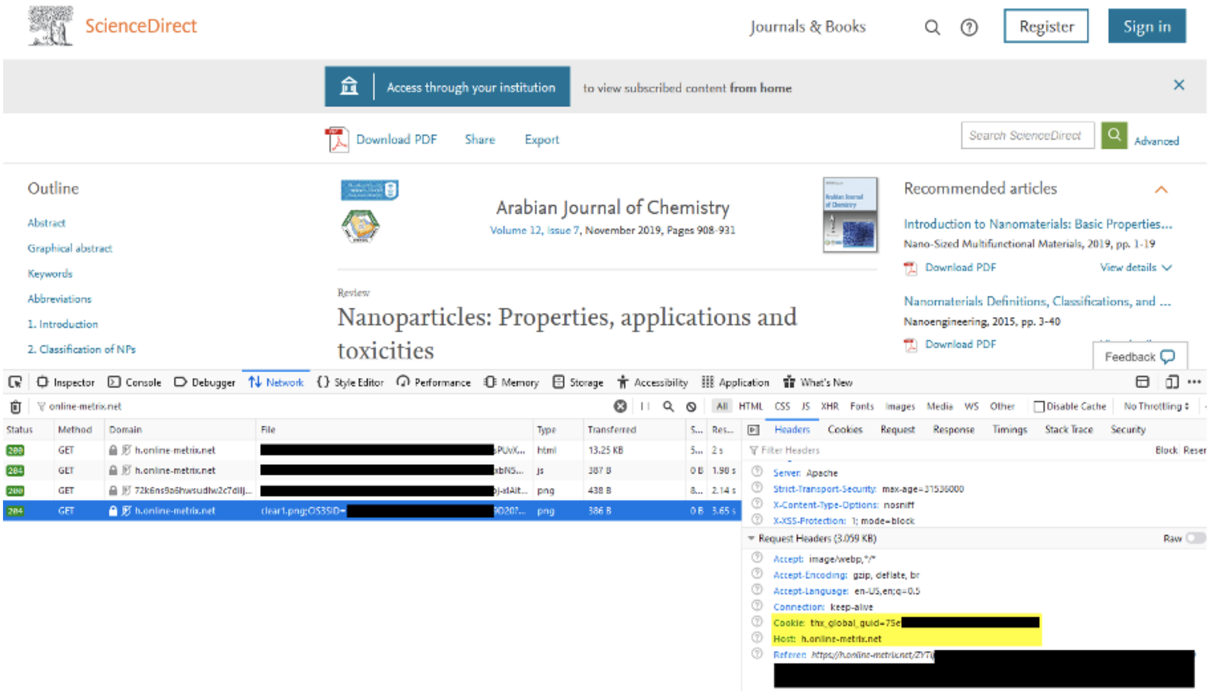
```
// `uuid` and
//
// Parameters:
//
// - `elm` - t
// - `prop` -
// - `val` - t
//      return
$data: functio
  if (!elm) re
  var __satell
  if (!uuid) u
  var datas =
  if (!datas)
```

Figure 1: Screenshot of Cody Hanson's talk

What are the consequences of this? First, it must be noted that this is a violation of fundamental rights: liberal societies have freedom of research and teaching enshrined in their constitutions, but a monitored freedom is not. Thus, academic freedom as we knew it has come to its end. But this is not merely an abstract shortcoming, because tracking can put scientists at concrete and grave risk. After all, in recent years we have seen that, as a researcher, you can get into trouble even in countries where you would not have suspected it in the past: in one country you don't make friends with research on climate change,[3] in another not with gender studies.[4] And in Hungary, the Central European University was chased out of the country.[5] Putting pressure on scientists is, of course, much easier when you know more about them, and since their behavioral profiles are tradable, they are of course open to all interested actors, as long as they want to spend the necessary money.

Economic consequences are also foreseeable, be it either in the form of a distortion of competition to the detriment of publishers who act in a trustworthy manner and therefore do not track, or, above all, in the destruction of the value of public research investments. In many future fields such as AI, personalized medicine, and material sciences, there is strong competition between public and commercial research, whereby the commercial players are in part also the ones who invented and disseminated the tracking technologies. Now, the same power of metadata makes itself felt in tracking as it does in telecommunication surveillance: if you know who contacted each other, when, how often, and in what sequence, you don't have to listen to the phone calls anymore, it's all there in plain sight anyway. Likewise, if researchers are being tracked, you no longer need to break into their lab, because you know what they are working on and how far they have gotten with it. And since it's not just one researcher who is being tracked and all the data can be analyzed together in business intelligence - who will be faster in filing a patent application? And will research results still benefit the public that paid for them?

So is it time to update the old joke "Come to the dark side, we have cookies"? Absolutely. And when we hear that Google, Apple and others are going to put a stop to third-party tracking, can we sit back because it will all have been an ugly transitional phase? Not at all. All the interested players are either busy monopolizing data access for themselves, like Google, or continuing to find ways for themselves, like others. If you take a closer look at your favorite journal, chances are you'll find tools for collecting Real Time Bidding Data, meaning your information behavior is auctioned off in real time without the need to set a cookie but e.g. for the benefit of intelligence services.[6] Elsevier, as part of RELX, benefits from technologies that reside in the group's Risk Solution division and has installed ThreatMetrix on its ScienceDirect platform, a technology that boasts of being able to individually identify billions of devices and, if you recall the discussion about ebay, does not shy away from port scanning.[7]



online-metrix.net

Status	Method	Domain	File	Type	Transferred	S...	Res...
200	GET	h.online-metrix.net	[redacted]sPUvK...	html	13.25 KB	5...	2 s
204	GET	h.online-metrix.net	[redacted]xbN5...	js	387 B	0 B	1.98 s
200	GET	72k6ns9a6hwsudlw2c7dilj...	[redacted]oj-xlAlt...	png	438 B	8...	2.14 s
204	GET	h.online-metrix.net	clear1.png;OS3SID=[redacted]9D207...	png	386 B	0 B	3.65 s

Server: Apache
Strict-Transport-Security: max-age=31536000
X-Content-Type-Options: nosniff
X-XSS-Protection: 1; mode=block

Request Headers (3.059 KB)

Accept: image/webp,*/*
Accept-Encoding: gzip, deflate, br
Accept-Language: en-US,en;q=0.5
Connection: keep-alive
Cookie: thx_global_guid=75e[redacted]
Host: h.online-metrix.net
Referer: https://h.online-metrix.net/ZYT[redacted]

Figure 2: Traces of ThreatMetrix on the ScienceDirect-platform

The argument about authentication is also far from over[8], and publishers sometimes use sleazy methods to enforce their wishes. In 2020, for example, libraries were given opt-out deadlines for switching to seamless access in the middle of the first lockdown, and those who didn't keep a constant eye on their email in this situation woke up to new access rules.[9] Using the same instrument over the Christmas holidays is also very popular, for example, to enforce the installation of Google CASA (Campus Activated Subscriber Access), which allows information behavior to be synchronized with the Google account. The goal is to always keep a hand on the valuable first-party data, because this is the best way to secure direct personalized access.

Borders are hardly accepted there; currently Elsevier and SpringerNature are considering infiltrating Trojans into university networks via the libraries in order to be able to collect biometric data such as typing speed or types of mouse movement in order to be able to identify users in other areas of the Internet as well.[10]

So what is the strategic goal of these publishers? First of all, they want to expand their unassailable position in information access to the entire research life cycle. This is achieved by contracts such as those concluded by Elsevier in the Netherlands, under which researchers can publish open access at no extra cost if the universities license Elsevier products in the area of research information systems in return.[11] Such contracts are disastrous, because it will be immediately clear to all researchers that their research will only be properly noted in the next evaluation if it also ends up in the appropriate journals. Consequently, the market for both publication and research information systems will quickly clear up as a result of this linkage business, and then the platform will have moved a step closer to its goal again: namely, that it is not science, universities or research funders who determine which research is the "right" one and how much it "counts"[12] but rather the platform who makes the determination.

Such deals seem like revenants from the nineties, when, in the browser wars Microsoft knocked Netscape out of the market by tying Internet Explorer to Windows. Another revenant is the behavior of libraries toward OA: if libraries had dried up the green road of OA in the nineties by first allowing publishers to transform their products and, most importantly, making self archiving worthless by signing up for package licenses[13] (which is why the head of SpringerNature called package licenses "the best invention since sliced bread"[14]), now they are once again holding out for publishers, where publishers are having their Napster moment through SciHub. By partnering with publishers to commercialize open access through publication fees, libraries are trying to crisis-proof their business: as long as science governance doesn't change, scientists will have to publish in the appropriate journals, and libraries will prevail precisely in the area of accounting and managing publication costs so they can continue to justify their staff and budgets. As a result, publication fees are already 30 times the actual expenses in some cases[15] - so there is always enough money with the publishers to continue to outspend, out-buy, and out-develop big data analytics any alternatives that are developed from within the scientific community. And, moreover, DEAL contracts in Germany are already beginning to distort competition from publishers in favor of the big players on whom the negotiations are focused – adding even more more money to even fewer players.[16]

New Business Models

Figure 3: ICE-Officer detaining a suspect

But for some players in scholarly communication, this is no longer enough. Thomson Reuters and RELX have discovered the global security industry as a buyer of their data analytics products, and have signed big data policing contracts with law enforcement organizations like U.S. Immigration and Customs Enforcement (ICE) that use these tools to track immigrants.[17] Since both publishers are central providers of specialized legal information, it is therefore not at all clear at the moment whether, for example, an attorney who consults Westlaw or LexisNexis for guidance on immigration law issues is thereby contributing to his or her clients being found and deported.[18] The publishers, of course, dismiss it all - and at the same time continue to escalate the path taken: RELX was also an early investor in Palantir, and so the information provided by LexisNexis can also be imported and analyzed in Palantir's analytics tools. And since these platforms and tools are sold globally just as the data they feed is collected globally, illegal immigrants don't have to be the only ones they're used against - the respective interested parties can surely think of something – and library money is supporting these systems of surveillance.[19] A scene of revolving-door deals between (semi-)state and commercial actors is emerging, where security institutions can get whatever they want to know on the market without having to rely any longer on onerous rule-of-law tools like obtaining a judicial search warrant. The motto of a data broker applies, who, when asked who was actually exchanging data with whom, replied: "Everyone with everyone. It's one[2] big whorehouse."[20]

So where do we stand? We have seen in the past that, due to the special situation of the Cold War, the major scientific publishers grew into an early form of platform economy, long before the large Internet corporations adapted this principle and were thus able to achieve their current position.[21] With their work on the Trojans, the publishers are also revealing geopolitical interests, namely to cut off entire states from the flow of scientific information, and are thus part of a Cold War 4.0.[22] All of this no longer has anything to do with scientific progress in knowledge, and so for years we have also had to recognize that the publication infrastructures are serving their actual purpose less and less. The attention economy of the journals has been analyzed time and again in the past[23], but only now, through the replication crisis and every glance at Retraction Watch, is it becoming clear how the compulsion for storytelling, the fudging of data, and the fight for the smallest publishable unit is destroying and corrupting science from within. Leading players in science communication will have no problem with this, as Elsevier itself has had half a dozen journals in the past that claimed to be peer reviewed but were mere shells rented out to pharmaceutical companies to be filled.[24] And the parent company RELX engages in political landscaping with donations to climate change deniers and to actors who drummed for U.S. withdrawal from WHO.[25] All this can only happen almost without contradiction because infrastructure is still a largely blind spot in science governance. More and more voices are calling for an "infrastructural turn",[26] but even in the welcome European initiatives the old mistakes are being made, because in the European Open Science Cloud Elsevier has nested[27] just as ORE is run by F1000 Research - which has been bought up by Taylor & Francis.[28]

So parts of the scientific information infrastructures are increasingly moving toward the dark money complex, as characterized by the Koch brothers, or even by Bob Mercer, whose investment made Cambridge Analytica possible. If you read Christopher Wylie's report on this[29] and compare it with Shoshana Zuboff's analysis of surveillance capitalism[30], you will see that science and antiscience[31] are increasingly amalgamating into a confusing tangle, because in the same way that this system could not be operated at all without a developed scientific education, it is at the same time laying the axe to the foundations of truth, enlightenment and social liberties. In discourse, this becomes clear in the meteoric rise of half-truths, which are used by conspiracy theorists, politicians, economic players, as well as the media - and more and more scientific authors. Jörg Peters recently developed this in his discussion of Stuart Ritchie's "Science Fictions" as a structural problem of an information infrastructure that rewards unreplicable research results and flushes them to the forefront of attention.[32] Not from the perspective of an economist, but of a literary scholar, Nicola Gess analyzes half-truths as a technique that no longer operates according to the scheme true/false, but according to schemes such as credible/uncredible, affective/sober, and connective/closed. What is central, she says, is internal coherence and no longer correspondence with external facts. Half-truths, therefore, could no longer be resolved with a fact check, but only with a fiction check, which highlights how the half-truth builds a narrative around a crystallization core of truth while pretending to report facts.[33] We find much of this in the corrupted information infrastructures, such as the impossibility for the reviewer to follow documents of a 21st century data-driven science to the ground of fact within a 17th century publication system with reasonable effort. He also has to rely on credibility and internal coherence, and as a consequence, subsequently exposed falsifiers like Diederik Stapel resort to the same justification phrases as Nicola Gess highlights on the fallen star journalist Claas Relotius.[34]

Conclusion

After all - when we talk about power and power abuse in academia, the information infrastructures must not be forgotten. They are in your lab, they are in your life, they're corrupting the truth, they're selling your ass and there is little you can do about it.[35] Nevertheless, it remains the responsibility of each individual to reflect on this for their actions in the science system and not to blindly accept it, as Tal Yarkoni states in his great rant:

"When people start routinely accepting that The System is Broken and The Incentives Are Fucking Us Over, bad things tend to happen. It's very hard to have a stable, smoothly functioning society once everyone believes (rightly or wrongly) that gaming the system is the only way to get by."[36]

Gaming the system means, as in any game of chance: the bank always wins. Acting responsibly, on the other hand, means supporting and improving existing European initiatives. In Germany, this is currently being done, among other things, through the National Research Data Infrastructure (NFDI), which is cooperative and public. This path must be expanded, and the mistakes from the literature supply must be avoided at all costs. Science will never be happy again as long as the big players remain in the game. The success of the NFDI is all the more crucial because in the data-driven sciences, data, software and text belong in a common context so that results can be verified and immediately developed further. For scientists, the current situation is like when a programmer first writes his code, then publishes an article about it, and the next programmer tries to guess the code from the article and then creates an improved version. Sounds highly efficient? And that's exactly why GitHub exists - and a trusted, powerful and non-buyable version of it would be exactly what science needs.[37]

And before I forget - there is one more thing you can easily do: join the community of scientists who don't want everything done to them and sign the Stop Tracking Science petition![38]

References

[1] Recommendations on methods for controlling access to scientific information resources. A joint paper by Deutscher Bibliotheksverband e.V. (dbv) and the Digital Information Priority Initiative of the Alliance of Science Organisations in Germany, 29.11.2019; https://gfzpublic.gfz-potsdam.de/pubman/item/item_4906895

[2] Cody Hanson: User Tracking on Academic Publisher Platforms. Prepared for the Coalition for Networked Information Spring 2019 Member Meeting, April 8-9, 2019, St. Louis, Missouri. https://www.codyh.com/writing/tracking.html

[3] Kari De Pryck, Francois Gemenne: The Denier-in-Chief: Climate Change, Science and the Election of Donald J. Trump. In: Law Critique. Band 28, Nr. 2, 2017, S. 119–126, doi:10.1007/s10978-017-9207-6

[4] https://www.dw.com/en/hungarys-university-ban-on-gender-studies-heats-up-culture-war/a-45944422

[5] https://www.derstandard.at/consent/tcf/2000093054082/top-uni-zum-abzug-aus-orbans-budapest-gezwungen

[6] https://www.vice.com/en/article/k78ewy/bidstream-data-google-twitter-att-verizon-foreign

[7] https://twitter.com/wolfiechristl/status/1286341387718397952

[8] Samuel Moore: Individuation through infrastructure. Get full text research, data extraction and the academic publishing oligopoly; DOI: 10.1108/JD-06-2020-0090

[9] https://twitter.com/codyh/status/1243250490403483648

[10] Gautama Mehta: Proposal to install spyware in universities libraries to protect copyrights shocks academics, Coda (13. November 2020), https://www.codastory.com/authoritarian-tech/spyware-in-libraries/

[11]https://www.scienceguide.nl/2019/11/leaked-document-on-elsevier-negotiations-sparks-controversy/

[12] Rupert Gatti: Business Models and Market Structure within the Scholarly Communications Sector; DOI: 10.24948/2020.04

[13] Heidemarie Hanekop, Volker Wittke: Der Wandel des wissenschaftlichen Publikationssystems durch das Internet. Sektorale Transformation im Kontext institutioneller Rekonfiguration. In: Dolata, Ulrich; Schrape, Jan-Felix, (Hrsg.): Internet, Mobile Devices und die Transformation der Medien. Radikaler Wandel als schrittweise Rekonfiguration. Berlin 2013, S. 147 – 172.

[14] Cited in SPARC Landscape Analysis, p. 22 (https://infrastructure.sparcopen.org/landscape-analysis).

[15] The max planck digital library concluded a contract with SpringerNature that provided for Article Processing Charges for "Nature" in the amount of 9,500 euros plus VAT. The actual costs for the publication are often around 300 euros or even less. So the VAT alone in the Nature deal is higher than the actual costs. Alexander Grossmann, Björn Brembs: Current market rates for scholarly publishing services; https://doi.org/10.12688/f1000research.27468.1

[16] Justus Haucap, Nima Moshgbar, Wolfgang Benedikt Schmal: The Impact of the German “DEAL” on Competition in the Academic Publishing Market. DICE Discussion Paper No 360. Düsseldorf 2021.

[17] Sam Biddle: LexisNexis to Provide Giant Database of Personal Information to ICE. The company signed a contract with an ICE division that plays a key role in deportations. The Intercept, April 2 2021; https://theintercept.com/2021/04/02/ice-database-surveillance-lexisnexis/

[18] Sarah Lamdan: Librarianship at the crossroads of ICE-surveillance; http://www.inthelibrarywiththeleadpipe.org/2019/ice-surveillance/

[19] SPARC: Addressing the alarming systems of surveillance built by library vendors; https://sparcopen.org/news/2021/addressing-the-alarming-systems-of-surveillance-built-by-library-vendors/

[20] Felix Ebert, Hannes Munzinger: Auf Sendung. Süddeutsche Zeitung 14./15.12.2019.

[21] Stephen Buranyi: Is the staggeringly profitable business of scientific publishing bad for science? https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

[22] See the attached documents in https://netzpolitik.org/2020/news-from-elsevier-no-open-access-deal-but-spyware-against-shadow-libraries/

[23] Martina Franzen: Breaking News. Wissenschaftliche Zeitschriften im Kampf um Aufmerksamkeit. Baden-Baden 2011.

[24] https://www.the-scientist.com/the-nutshell/elsevier-published-6-fake-journals-44160

[25] https://twitter.com/richardhorton1/status/1295818198059814921

[26] Benedikt Fecher, Gert G. Wagner: Open Access, Innovation and Research Infrastructure; https://doi.org/10.3390/publications4020017, Björn Brembs et.al.: Plan I - Towards a sustainable research information infrastructure; https://zenodo.org/record/4454640

[27] Jon Tennant: Elsevier are corrupting open science in Europe; https://www.theguardian.com/science/political-science/2018/jun/29/elsevier-are-corrupting-open-science-in-europe

[28] https://www.tandfonline.com/openaccess/f1000

[29] Christopher Wylie: Mindf*ck. Cambridge Analytica and the Plot to break America. New York 2019.

[30] Shoshana Zuboff: The Age of Surveillance Capitalism. New York 2019.

[31] Peter J. Hotez: The antiscience movement is escalating, going global and killing thousands; https://www.scientificamerican.com/article/the-antiscience-movement-is-escalating-going-global-and-killing-thousands/

[32] Jörg Peters: Empirical economics: A replication crisis in the making?. Elephant in the Lab. https://doi.org/10.5281/zenodo.46234

[33] Nicola Gess: Halbwahrheiten. Zur Manipulation von Wirklichkeit. Berlin 2021.

[34] Yudhijit Bhattacharjee: The Mind of a Con Man, New York Times 26.4.2013; http://archive.nytimes.com/www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html

[35] See the struggle of a librarian with Thomson Reuters: https://twitter.com/SheaSwauger/status/1205587676172144641

[36] https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/

[37] Björn Brembs: Scholarship has bigger fish to fry then access; http://bjoern.brembs.net/2019/10/scholarship-has-bigger-fish-to-fry-than-access/ And because digital teaching is the next focus of EIFL: comparable initiatives also exist in the field of teaching, which EdTech combats: Justus Lentsch: Unsere Bildungsdaten gehören uns! https://www.jmwiarda.de/2021/02/16/unsere-bildungsdaten-geh%C3%B6ren-uns/

[38] https://stoptrackingscience.eu/