Science-Tracking Fingerprint: Track the Tracker on the Example of Online Public Access Catalogs (OPAC)

1st Stefan Kiltz
Faculty of Computer Science
Otto-von-Guericke-University of Magdeburg
Magdeburg, Germany
stefan.kiltz@iti.cs.uni-magdeburg.de

2nd Nick Weiler
Faculty of Computer Science
Otto-von-Guericke-University of Magdeburg
Magdeburg, Germany
nick.weiler@st.ovgu.de

3rd Till-Frederik Riechard
Faculty of Computer Science
Otto-von-Guericke-University of Magdeburg
Magdeburg, Germany
riechard@ovgu.de

4th Robert Altschaffel
Faculty of Computer Science
Otto-von-Guericke-University of Magdeburg
Magdeburg, Germany
robert.altschaffel@iti.cs.uni-magdeburg.de

5th Jana Dittmann
Faculty of Computer Science
Otto-von-Guericke-University of Magdeburg
Magdeburg, Germany
jana.dittmann@iti.cs.uni-magdeburg.de

***Abstract*—We are motivated by the Science-Tracking Fingerprint (STF) from our companion conference article 'Science-Tracker Fingerprinting with Uncertainty: Selected Common Characteristics of Publishers from Network to Application Trackers on the Example of Web, App and Email' and apply this fingerprint concept to Online Public Access Catalogs (OPAC) provided by many libraries for literature research with the aim to track the tracker. We choose an approach rooted in digital forensics and using only open source, on-premises tools for comprehensibility and repeatability purposes. The goal of this article (together with its companion article) is not primarily to detect the amount of tracking that is taking place. Studies towards that goal have, indeed, been conducted both on the Science-tracking field and on the field of tracking in general. Our goal is to try and identify the publisher based on the employed first and third party tracking. In particular, for the application area of web we enhance the concept from the companion article with an automated acquisition, investigation and analysis process, including the calculation of the STF. Further, the single list of trackers from the companion article is extended and we provide 3 different lists of known trackers in order to increase the hit-ratio for known tracker domains. For the automation we introduce a toolset consisting of 6 self-created software tools and 4 automation scripts that are put into open source. The automation enables a substantially larger investigation on both the tracking habits of publishers and allows evaluations of the stability of the Science-Tracking Fingerprint. In total we fully evaluate 60 downloads from the 4 exemplary chosen individual publishers across 3 different test-series. Further, to detect any possible influence of the changes of the domains contained in the tracker lists, we use 3 different versions of each of the 3 tracking lists and apply it to each test series. The results of our in-depth study into Science-Trackers show that some publishers change their embedded trackers over individual papers and articles (intra-publisher diversity). For the duration of the tests, no changes on the content of the tracking lists relevant to the tests occurred. Results from 4 tested publishers show no difference in the observed tracking between open access and non-open access articles. Further, we show that using the exemplary chosen OPAC instance of our university library does not prevent Science-Tracking by the publishers, potentially contrary to the user's expectations. This article proposes a comprehensible, scientific process to support the identification of the tracking party (publisher) based on the trackers employed by the tracking party.**

***Keywords*-Security, trust and privacy metrics; IT forensics; Attribution.**

I. Introduction

Science-Tracking by publishers, as stated in [1] is in widespread use ([2], [3]). This often stealthy practice subjects users of literature information systems to unwanted data processing and impacts their privacy, sometimes with potentially grave consequences [2]. On a side note, data from scientists can also be obtained and sold through breaches in conference registration systems etc. (see e.g., [4]).

To get an overview of the extent of the tracking of scientists by publishers, an IT-forensic approach as motivated in [5], conforming with [1] can be a valid course of action. As already pointed out in the companion conference article [1], each forensic investigation method comes with the potential for error, loss and uncertainty, which can influence the resulting traces. Hence, using results from multiple, independent tools for the same forensic research goal is used to reduce these negative effects. Further, our goal is to gather hints/leads

The research from Stefan Kiltz, Robert Altschaffel and Jana Dittmann is partly funded by the EFRE project "CyberSecurity-Verbund LSA II – Prävention, Detektion und Reaktion mit Open Source-Perspektiven" https://forschung-sachsen-anhalt.de/project/cybersecurity-verbund-lsa-praevention-27322.

leading towards an individualization of a publisher based on the trackers employed (first and third party)

The ability to identify instances of tracking open the way to investigate interesting questions about the extent and practical use of tracking. This work aims to answer the following primary research questions (***RQ***), extending the companion conference article [1]:

- ***RQ1*** *Track the tracker:*
whether it is possible to find traces/hints that allow for an individualization (attribution) of the publisher that employs the tracking mechanisms
- ***RQ2*** *Intra-publisher diversity:*
how stable the traces are over time for a given publisher and within multiple documents from the same publisher
- ***RQ3*** *Countermeasures against tracking by using OPAC:*
whether the usage of library-supplied research gateways such as an Online Public Access Catalogues (OPAC) prevents the tracking techniques employed by the publishers
- ***RQ4*** *Effect of open-access on tracking:*
whether there is any noticeable difference in the tracking behaviour when accessing open-access and non-open access articles.

Based on [1] we trade broadness for detail in our research and focus on Science-Tracking on the example application area of web-based services accessed via browser. We choose a scenario that reflects the typical usage of scientific literature research using our university library and the Online Public Access Catalogue (OPAC) gateway [6] used therein. This is particularly interesting since users could expect to fetch the documents proxied by the university library and this could lead them to suspect that they are not tracked by the publishers in the same way as stated in [1]. We contribute a semi-automatic approach to calculate the Science-Tracking Fingerprint (STF) for the web application area. With the partially automated support, we can look into changes in detected tracking mechanisms per publisher using different articles and different points in time (intra-publisher diversity).

Addressing these research questions includes various steps, concepts, extensions and improvements over the companion publication [1] that might also be applied to other research questions in the future. These are:

- ***Extension E1*** - an inter-publisher comparison of intersected STFs for estimating the difference in tracking behaviour between publishers.
- ***Extension E2*** - the creation of a STF-deviation metric to show the difference between different STFs.

These extensions are necessary to identify the various tracking parties and hence address ***RQ1***.

- ***Extension E3*** - the concept of the evaluation of tracking across multiple documents and points in time (t_i, t_{i+1} see [1]) for an individual publisher (intra-publisher diversity) and its comparison using the STF.

The extension ***E3*** is necessary to investigate the diversity of tracking methods used by a specific publisher (***RQ2***).

Furthermore, the extension of the system landscape is necessary in order to investigate ***RQ3***, which is related to OPAC and hence requires its inclusion.

- ***Extension E4*** - the inclusion of an Online Public Access Catalog (OPAC) library gateway to the publisher's articles into the system landscape used for research, which adds a credible scenario of Science-Tracking in common literature research.

Additional noteworthy extensions to the work performed in [1] are provided in the following. They either extend previous work, simplify future forensic investigations or cover notable findings:

- ***Extension E5*** - a partially automated process consisting of 10 scripts that are put into Open Source and covering the acquisition, investigation and partly analysis according to the sets of investigation steps from [7] for calculating the STF.
- ***Extension E6*** - the usage of multiple lists of known trackers for the investigation, saved at different points of time, to have higher chances of detecting trackers and their analysing tracking detection behaviour.
- ***Extension E7*** - an additional analysis of the publisher Wiley to broaden our group of investigated publisher.
- ***Extension E8*** - the discovery of different tracking behaviour depending on the type of browser (interactive vs. headless with automated control flows.

With both implementing an automated process within the investigation (***E5***) and with the concept of the evaluation of tracking across multiple documents and points in time for an individual publisher (***RQ2***) and its comparison using the STF (***RQ1***) we are addressing the future work suggested in the companion conference article [1].
This article is structured as follows: In Section II aspects of the relevant state of the art are outlined briefly. Section III describes the necessary fundamentals for understanding the concept, implementation and evaluation of this article. In Section IV we discuss our conceptual approach centred around a model of the forensic process and introduce the STF-deviation as a metric to describe differences between STFs. In Section V we describe the implementation of the concept using pseudo-code to illustrate the workings of the 6 self-created software tools. In Section VI the concept and its implementation is evaluated, forming the contributions outlined in Section I. This article closes with a conclusion and an outlook regarding future work in Section VII.

II. State of the Art

As already stated in [1] a number of studies exist that look into data tracking in general. For instance, Wolfie Christl in [8] investigates digital tracking and profiling by corporate networks and their implications for the user ranging from individuals to society at large. On the technical side the research covers the practices of recording, combining, sharing, and trading of personal data. The main effort is directed at mapping of today's personal data ecosystem and determining its scope. The study from Mildebrath ([9]) takes a detailed look on the tracking mechanisms and practices employed by Google, Facebook and Amazon, both on the web and using mobile app infrastructures. The focus of the study from Samarasinghe et al. [10] is put on the influence of the geolocation of a tracked user by differentiating the tracking results from 56 countries based on a selection of frequently accessed websites. The study from Sim et al. [11] primarily focuses on existing tools and measures to detect (tracking-measurement) and prevent various types of web-based tracking and also glances into app-based tracking. Also addressing prevention of tracking, the study from Pan et al. [12] looks at the success of the attempt of browser manufacturers to block tracking mechanisms. The measurement of the success is performed using available privacy scanner and its conclusion is a slight reduction of tracking by modern browsers on the example of Google Chrome. Geared towards the field of mobile devices, the study from Krupp et al. [13] focuses on mobile devices, which offer lesser tracking protection based on the fact that privacy enhancing browser add-ons and extensions are typically unavailable for the apps. The research focuses on iOS devices and reveals a substantive amount of tracking in the apps chosen for the research by the authors.

Science-Tracking, which is the subject of this article and its companion conference article [1], can be looked upon from very different angles, e.g., primarily from a legal perspective as conducted in the article from Altschaffel et al. [14]. The study done by Hanson [15] looks into the extent of Science-Tracking from a technical perspective. Key findings also include the huge amount of third party tracking by third-party code being loaded whilst accessing an article's page provided by a publisher. The tracking mechanism provided by the third parties employed by the publishers identified by [15] seem to primarily consist of the generic third party tracking solutions also employed in general tracking as outlined in the above paragraph, which also mirrors our findings from [1].

All reviewed studies share the fact that they try to determine to what extent tracking exist on various application fields and elaborate on the consequences of user tracking. The study [5] already employs forensic techniques for the detection of tracking. According to our knowledge, [1] is the first attempt at a study with forensically motivated systematic means to give hints/leads to individualize (attribute) tracking to identify an originator. Hence, this publication is used a foundation for our work. In this article at hand the approach outlined in Section IV is a refined attempt at fingerprinting originators of Science-Tracking (organizations such as publishers) on the basis of their employed first and third party tracking mechanisms also for the task of comparing different originators. The authors are fully aware that the suggested approach alone will not suffice for individualization and thus attribution but believe that it can give hints/leads towards further investigation.

The topic of data tracking is also of interest outside the field of academia. For instance, the European Union *Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers* [16] investigates the tracking of users a foundation for targeted digital advertisement. The study investigates the data reported and the means employed by publishers to do so. As such, it provides a broad understanding of data tracking but does not provide any means to measure the occurrence of data tracking.

III. Fundamentals

This section describes the necessary fundamentals for the research presented in this article. It relies heavily on the findings, fundamentals and findings from our companion conference article [1].

A. The Data-Centric Examination Approach (DCEA) forensic process model

A comprehensive, model-based approach (as also used in [1] supports the forensic soundness. The Data-Centric Examination Approach (DCEA) [7] uses data streams and forensic data types, which together with forensic methods (represented by capabilities of forensic tools) supports a detailed description of the provenance of the data from the beginning of the examination to its end. This is seen by the authors as an aid to attribution. The model from [7] distinguishes three data streams:

- Mass storage data stream DS_T (time-discrete, low volatility, long-term data retention),
- Main memory data stream DS_M (time-discrete, high volatility, short-term data retention),
- Network data stream DS_N (time-continuous, high volatility, short-term data retention).

Throughout this article (as in [1]) we will use DS_T and DS_N during our examinations. Those data streams can be further divided into 8 forensic data types with the assumption that data of a specific data type is created, processed, stored and used similarly by a given IT system and thus can be acquired, investigated, analyzed and documented similarly in a forensic examination [7]. For our article (as in [1]) we use DT_3 (details about data) and DT_5 (communication protocol data) in the context of the network data stream and its representation in mass storage.

The collection of main memory data from DS_M and its examination, whilst being available in theory, is omitted due to the extra effort weighted against the additional information gained. It would involve halting the VM used for the examination to capture the RAM content for each point of interest during the examination and creating a dwarf specifically for the examination environment with the Volatility framework (see e.g., [17]) and browsing the processes for relevant data. The authors believe that capturing highly volatile data in the shape of DS_N and DS_T for data with low volatility represents a measured approach and a good balance between effort and the gain with regards to data containing relevant information for the research.

The system landscape analysis is, according to [7], part of a forensic examination. The spatial and temporal intricacies

of tool placement and operation define what can be obtained and analyzed. As stated in [5], the usage of on-premises tools allows for finer control over the tool operation and external data (e.g., lists used for comparison against known tracker URLs) and better data access (e.g., regarding intermediate results). Opposed to the original work in [1] we will use exclusively on-premises tools and rely on corroboration of the tool results of the different on-premises tools. This enables a finer control over the tool configurations and external data used. In Section IV-D we discuss the properties of both approaches with our system landscape analysis.
The existing model-based approach of the forensic examination as described in [7] alone is not sufficient for the individualization (attribution). However, it provides us with the elementary building blocks for the fingerprint (e.g., data streams, forensic data types).

B. Selected tools and data sources for URL and Tracker examination

We select existing tools based on their proven functionality (analogous to the companion conference article [1]) and combine them in scripts (bash- and python-based) that cover different tasks of the investigation process. The choice of tools is based on the following requirements:

- Open Source: the tool must be comprehensible and potential changes on the source code must be possible
- Maintenance: the code must be maintained and updated by the tool authors
- On-premises installation: access to the data collected (including intermediate data and examined must be strictly local
- Forensic operation: the tool must not alter the immediate data nor alter the behaviour of the client or server software

Frameworks such as OpenWPM [18], whilst being generally suited for privacy measurements, can violate some of the requirements (e.g., due to using the Firefox engine, which can automatically start connections unrelated to the measurements such as software and certificate updates, contacting safebrowsing service providers etc., interfering with the data in the network data stream DS_N). We further select tools such as Webbkoll [19], although they only collect a subset of data of privacy measurement tools, based on the goal of our research regarding individualization of publishers based on their employed first and third party tracking mechanisms. Using the terminology from [1] we use tools that operate both statically and dynamically. The forensic data types and data streams (see Section III-A) are used from [7]. Contrary to [1] we only use on-premises tools for full source-level control over their functionality and parameterization. The existing tools used in the scripts combined are:

- Webbkoll [19]: on-premises, operating on the network data stream DS_N on Raw Data DT_1 and yielding tracker output DT_3 (in conjunction with external data, i.e., tracker list data) as results as well as URL and IP data DT_5 as results, both output to the mass storage data stream DS_T
- TShark [20]: on-premises, operating on the network data stream DS_N on Raw Data DT_1 and yielding URL and IP data DT_5 as results on the mass storage data stream DS_T
- Website evidence collector [21]: on-premises, operating on the network data stream DS_N on Raw Data DT_1 and yielding URL data DT_5 as results on the mass storage data stream DS_T

The tools used in our research (see Section III-B) utilize a headless version (without graphical interface) of the chromium browser [22]. For this the library Puppeteer [23] is employed to provide an easy as well as time and resource efficient way of implementing the forensic tools in a headless environment. The data sources for the 60 papers originate from our university's library OPAC gateway that are redirected to the 4 selected publishers:

- Association for Computing Machinery ACM Inc.
- Elsevier
- Institute of Electrical and Electronics Engineers IEEE
- Springer Nature

All recordings are conducted at the dates of:

- 20/02/2024
- 12/03/2024
- 25/03/2024

For the external data we use the sources of the lists of:

- Disconnect [24]
- Easy Privacy [25]
- Fanboy Annoyance [25]

These lists provide the classification of a given domain as a tracker. They are used for the dynamic examination (see also [1]) of the recordings created by TShark. A decision is reached whether a given domain is a tracker by comparing them against the lists. Different lists are used to render the results more plausible. We acquire the list data at the dates of:

- 20/02/2024
- 25/02/2024
- 13/03/2024

With those differently timed versions of the lists, we can conduct experiments regarding changes in detection depending on the changing content of the 3 lists over time. With this setup we can address the point raised in [1], which at a minimum asked for the dates to be recorded alongside with the result for comparability. Our setup allows for retrospective runs of the tests on the data with arbitrary dated lists.

C. Uncertainty in forensic examinations

Uncertainty is a property that should be factored in for all forensic examinations [1]. This is laid out in detail in [26]. For the approach in [1], which is adapted for usage in the research described in this article, the certainty category therein is also employed. This certainty category from [1] weighs the results of different forensic tools capturing URL-data and tracker detection data as matches of the results being plausible, uncertain or non-existent, depending on whether all tools agree with the results, at least one tool returns a diverging result or no matches exist at all.

D. Semantics and syntax of the Science-Tracking Fingerprint (STF)

In the companion conference article [1] the semantics and syntax of the Science-Tracking Fingerprint (STF) are introduced. Here, we provide a brief summary of the concept as a basis for this article.

One goal of the STF is the support for individualization [27] and attribution of the publisher employing the tracking techniques (track the tracker). The general idea, with regards to the semantics of the STF, is to employ more than one forensic method to acquire, investigate and analyse the data in the absence of a ground truth when accessing the articles supplied by the publisher. We record the agreement (matches) of the respective tool results according to the certainty categories (see Section III-C) of:

- plausible (pl): all tools return the same or comparable result,
- uncertain (unc): at least one tool returns a diverging result,
- none (-): no tool returns a meaningful result.

Semantically, the Science-Tracking Fingerprint can be described as a matrix of A-Records for first and third party as well as CNAME domain names for first and third party on one axis and Web, App and Email on the axis. Each cell contains a structured description covering the following elements:

- Counter: Number of occurrences,
- Certainty: plausible, uncertain or none,
- Data stream: Mass storage (T) or Network (N),
- Data type: DT_5 (URL) or DT_3 (Tracker),
- Discovery mode: list-based (L) and/or manual (M).

A fixed structure for the notation of these elements is necessary to support comparisons between the findings obtained with different forensic methods. The structured description is summarized:

```
<CELL> ::= <Counter> <EXPR>
<EXPR> ::= <EXPR1> | <EXPR>,<EXPR1>
<EXPR1> ::= <Certainty>,<Data stream>,<Data type> |
<Certainty>,<Data stream>,<Data type>,<Discovery
    mode>
```

Listing 1. Structured description for the cell contents formed from relevant elements.

The semantics of the STF describe quantifiable and qualitative differences between the Science-tracking employed by the publishers, with changes over time to be expected, which is why the STF is treated as a similarity measure [1].

According to [1], the syntax of the STF can be described a concatenation of vectors consisting of element value pairs, which form the matrix shown in Figure 1.

	A-Record 1st Party	CNAME 1st Party	A-Record 3rd Party	CNAME 3rd Party
Web	<CELL>	<CELL>	<CELL>	<CELL>
	<CELL>	<CELL>	<CELL>	<CELL>
App	<CELL>	<CELL>	<CELL>	<CELL>
	<CELL>	<CELL>	<CELL>	<CELL>
Email	<CELL>	<CELL>	<CELL>	<CELL>
	<CELL>	<CELL>	<CELL>	<CELL>

Figure 1. Syntactical matrix representation of the STF according to [1].

Each row (according to [1]) consists of a set of cells that are ordered according to the URL specifics (DT_5), namely the A-Record and CNAME domain name entries for both first and third party, respectively. Those cells can also be empty (represented by a 0), if there are no domains in the investigated recording. The part of the counter in the cell describes numbers of occurrences according to the following conditions:

- matching certainty per cell,
- tracker certainty is either plausible or uncertain.

A special case is met when a row contains entries where the DNS response provided URL information containing CNAMEs for the first and/or third party. In [1] it is described to duplicate the cell entries from the A-Record to the CNAME without increasing the counter value, as this case with CNAMEs first and/or third party in one row technically describes the same examination step.

As stated in Section I, in this article we are only using the Web application area part of the syntactical representation of the STF.

IV. Conceptual Approach

This section describes the conceptual approach to the web-based investigation performed in this article. The approach focuses on collecting DT_5 and DT_3 data from as many publications as possible (to reduce the potential error, loss and uncertainty, see Section I) by intersecting the sets of gathered trackers from different publications of a publisher. The investigation is performed on a test series. It uses the set of examination steps from [7] (see Section III-A). We discuss in detail the three steps of:

- Data gathering
- Data investigation
  - Generation of result tables and STFs
  - Aggregation of STFs
- Data analysis

The complete analysis process is shown in Figure 2. It outlines the three main steps (data gathering, data investigation, data analysis), the input (test set, external list data, see Section III-B) and intermediate results and the analysis questions to be answered.

Data gathering in essence marks the acquisition of data. It only gains raw data DT_1 for further investigation and analysis in the following steps.

By including the Online Public Access Catalog (OPAC) gateway provided by our universities library this puts restrictions on the location of the acquisition device (see Section IV-D) but allows us to see the perspective of the researchers using the library services (see ***Research Question RQ3*** by employing ***Extension E4*** in Section I).

By using different types of browsers (interactive vs. headless) during data gathering we extend the research from the companion article [1] and provide the ***Extension E8*** (see Section I).

The selection of the types of documents (open-access vs. non-open-access) to be queried during data gathering addresses the ***Research Question RQ4*** (see Section I).

The process of the generation of result tables and STFs as part of the data investigation step allows for multiple comparisons against different versions of the tracker lists from the documents already gathered enhances the findings from the companion article [1] as the ***Extension E6*** (see Section I).

Intersecting the generated result tables and STFs provides further insight into intra-publisher diversity and inter-publisher differences addresses the ***Research Question RQ2*** and enhances the findings from the companion article [1] as the ***Extension E3*** (see Section I).

The general design of the examination process with a focus on automation enhances the findings from [1] as the ***Extension E5*** (see Section I) while adhering to the model from [7]. It thus ensures a correct re-iteration of each step, which enhances the findings from [1] as the ***Extension E3*** (see Section I).

In the following, we will describe details regarding each of those selected examination steps.

A. Data gathering

During the acquisition, the necessary data is collected via the described tools in Section III-B and saved to the mass storage. While Webbkoll [28] and Website Evidence Collector initially gather raw data DT_1 internally for later investigation of the website for possible third party hosts, TShark records the network traffic as raw data DT_1 (for later external investigation and analysis) whilst querying the publisher website for the literature. The acquisition must be performed within the network of the university ; without an explicit login access to the papers provided by the OPAC of the university is impossible. Further processing of the gathered data may be performed elsewhere.

Interestingly, the choice of the type of browser, headless or graphical browser, influences the recording of the network data (see Section VI-A) and forms our ***Extension E8*** in Section I). Although at first counter-intuitive since researchers use a graphical browser in their daily research, we choose to use headless browsers on the grounds that:

a) this is also used in commonly accepted forensic tools such as Website Evidence Collector [21],
b) because it allows for automation and thus enables an examination for a much larger figure of documents.

To get more insight into the influence of the used type of browser on tracking behaviour, sample recordings with a graphical browser are conducted to compare the amounts of gathered data in both cases. The findings of this sample to our research data are detailed in Section VI-A.

B. Data investigation

We highlight the two steps that are performed during data investigation step that is following the data gathering step. The data investigation is partial automated by using self-created scripts.

1) Generation of result tables and STFs: With the collected data from the publisher websites (in our case of our research totalling 2.1GB, see also Section V-A for technical data on the devices used), a result table listing all discovered third party hosts is generated based on [1].

For this, the relevant DT_5 data:

- host name,
- ip address,
- whether host is third party,
- host is A Record or CNAME (see [1] and Section VI-A),

is gathered from the output data and combined to a structure. This structure is checked, by identifying whether a host is known in a list or not, gaining DT_3 data. This check is performed with every list and the result of each check is kept separately, since there could be differences within the lists. Once all hosts were checked on each list, a DT_3 and DT_5 match will be performed to grade the plausibility of the detected tracker. If on either DT_3 or DT_5 match at least one result of ”uncertain” was achieved, the host is classified as a potential tracker [1]. After all checks have been performed on each gathered host, the result table and STF are generated based on the information gained. To also cover the possibility of change in detection of trackers over time, each tracker list has a version related to the date of data acquisition. In Listings 3 and 4 from Section VI-A the pseudo-algorithmic approach of the evaluation and the update of the STF for every host is shown.

2) Aggregation of STFs: When the test series is processed completely, an aggregation based on the DT_3 and DT_5 of the results and the STFs calculated thereof is performed to get a more general view. The papers are divided in groups depending on their publisher and open access status. A comparison between open access and non-open access literature of a publisher provides further insights into differences in observable tracking behaviour between the aforementioned groups. A result table and STF, which represent the intersection of all detected hosts in each paper, are generated from the groups.

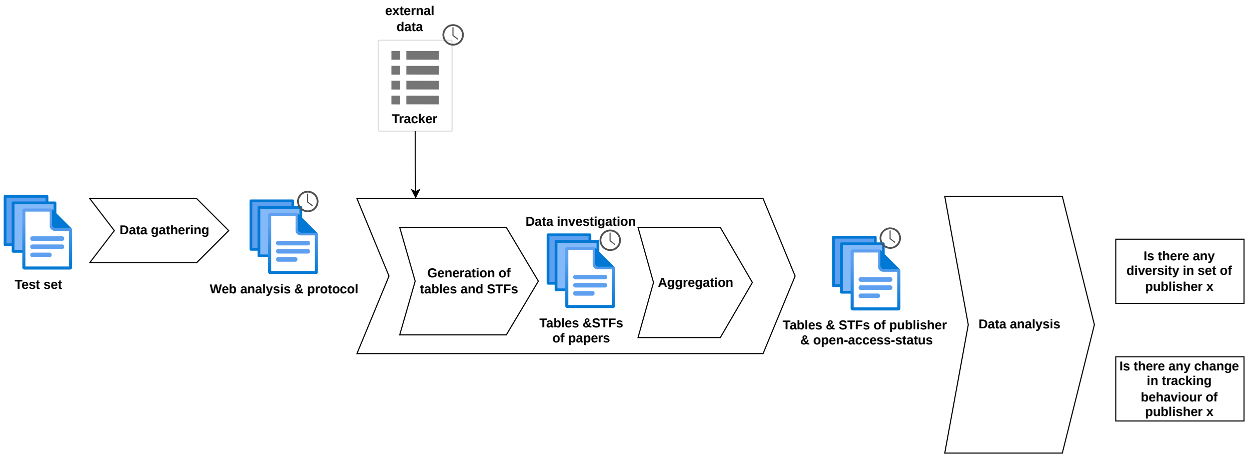

external data
Tracker
Test set
Data gathering
Web analysis & protocol
Data investigation
Generation of tables and STFs
Tables &STFs of papers
Aggregation
Tables & STFs of publisher & open-access-status
Data analysis
Is there any diversity in set of publisher x
Is there any change in tracking behaviour of publisher x

Figure 2. Visualization of the complete conceptual approach.

The Listing 5 in Section VI-A visualizes the aggregation as a pseudo-code algorithm.

C. Data analysis

With the generated result tables and STFs based on the DT_3 and DT_5 data from the investigation step, further examination is performed during the data analysis step, which also entails a detailed evaluation. The trackers detected in the emerging groups of papers are checked for intra-publisher diversity (a deviation from the intersected STF of the publisher; not to be confused with the statistical deviation) within the set of detected trackers (see also Listing 8 in Section V-B3).
Additionally to the intra-publisher analysis, results from different test series are compared by checking the results of the same paper for differences between the test series.
Furthermore, groups of the same publisher but with different open access status are checked for a difference in the set of detected trackers. Last but not least, a comparison between STFs of different versions of the tracker lists is executed for each paper.

Table I from the companion conference article [1] shows an exemplary result table containing the STFs of the publisher ACM and is used to outline the procedure. (Note that in this article, we are focusing exclusively on the web-based retrieval of papers and thus only on web-based Science-Tracking.)

	A-Record 1st party	CNAME 1st Party	A-Record 3rd Party	CNAME 3rd Party
Web	0	0	$3_{PL,N,DT5;PL,N,DT3,L}$	0
	0	0	$2_{PL,N,DT5;PL,N,DT3,L}$	$2_{PL,N,DT5;PL,N,DT3,L}$
	0	0	$6_{PL,N,DT5;UNC,N,DT3,L}$	0
	0	0	$2_{PL,N,DT5;UNC,N,DT3,L}$	$2_{PL,N,DT5;UNC,N,DT3,L}$
	0	0	$1_{UNC,N,DT5;UNC,N,DT3,L}$	$1_{UNC,N,DT5;UNC,N,DT3,L}$
App	0	0	$1_{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}$	$1_{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}$
	0	0	$1_{UNC,T,DT5;PL,T,DT3,L; UNC,N, DT5; PL, S, DT3, L}$	0
Email	0	0	$1_{PL,T,DT5;PL,T,DT3,M, 1PL,N,DT5;PL,N,DT3,M}$	0

TABLE I. Exemplary Science-Tracking Fingerprint (STF) from the companion conference article [1] of the ACM publisher using the structured semantic description and the syntactical vector formed by element-value pairs.

As a means of evaluating the difference between two STFs, we introduce the STF-deviation as a metric, forming ***Extension E2*** in Section I.
The STF-deviation serves as an estimate to the degree of difference between two STFs. The intention is to generate a value that can be compared to a percentage difference where 0.0 means no difference and 1.0 and above means total difference in tracking behaviour. The STFs in question can be either derived from a publication or an intersection of a group of papers from a publisher. This enables a comparison of paper websites to the collected information of a group. For the following equations we will use $a_{i,j}$ as the value of the cell of a STF A and $b_{i,j}$ is the value of the cell of a reference STF B. The STF-deviation is formed by row-wise comparison of the STFs and summing the relative differences with respect to the size of the respective row difference and the total size of STF B. The latter results in the STF-deviation to be a weighted sum due to the ratio, which is intended to put the row difference in perspective to the total size of STF B.

$$\Delta_{row}(i) = \sum_{j \in StfCol} |a_{i,j} - b_{i,j}|, i \in StfRow \qquad (1)$$

$$rowDev(i) = \begin{cases} NaN & if \sum_j b_{i,j} = 0 \\ \frac{\Delta_{row}(i)}{\sum_{j \in StfCol} b_{i,j}} & otherwise \end{cases} \qquad (2)$$

$$Dev = \sum_{i \in StfRow} rowDev(i) \cdot \frac{\Delta_{row}(i)}{\sum_{j \in StfRow, k \in StfCol} b_{j,k}} \qquad (3)$$

The indices i and j correspond to the cell within the STF without taking the title column and row into account (e.g., i=1 and j=3 corresponds to “first row, A-Record third Party”, pointing to “$3_{PL,N,DT_5;PL,N,DT_3,L}$”). Equation (1) mirrors the total size of the mismatch between the STF’s and references’ row. The total size of the mismatch is put into perspective to the row size of the reference in Equation (2) as a deviation to the row of the references. The summands of the deviation are weighted to put the deviation of a row into perspective to the total size of the reference STF in Equation (3). This is done with the intention of reducing the distortion due to different sizes of the rows. As one can see, Equation (2) is only partially defined. We decided that in this research only rows from the referenced STF will be taken into account for the deviation to avoid the distortion of the resulting deviation value in Equation (3). That means, that the value of the deviation does not encompass the total deviation but is a measure for the minimum deviation of a STF from another. Furthermore, the STF-deviation is not satiated at 1.0 since, depending on the STFs chosen for comparison, a higher value than 1.0 may be achieved. Whether this issue might be fixable by inversion of the value or is a general problem of the metric, is not clear at this point. Future work should address the issues and aim for a total STF-deviation metric with a more percentage-wise approach. The implementation of the STF-deviation metric is shown in Listing 7 in Section V-B3.

While the calculation of STF-deviation and the collection of comparison results is carried out automated, the results are evaluated as interpretable trends. In the evaluation, the STF-deviation values are interpreted. The interpretations are based on the values themselves and the comparison of values between different analysis groups (intra-publisher, inter-publisher, etc.). There are two general rules for the interpretation:

- Smaller values are interpreted as small deviation, which indicates similar tracking behaviour and greater values vice versa,
- Values similar to a certain analysis group are interpreted as such.

For example, if the values in the intra-publisher comparison group are between values x and y and a STF-deviation of a comparison lies within the interval [x,y], the value is interpreted as being a trend to similar tracking behaviour.
This method of determining whether a STF is similar to the tracking behaviour has further drawbacks:

- interpreting lower values, even zeros, as similar might result in more false positives,
- interpreting higher values as not similar might result in more false negatives.

D. System landscape analysis for Science-Tracking Fingerprint examination

Extending and focusing our research from [1], we eliminate the off-premises examination by hosting our own Webbkoll server inside the examiner's System E1 and thus on-premises. Further, we limit ourselves to web-based access to scientific articles, enabling an in-depth analysis with substantially more tests. Figure 3 shows the altered setup.
It shows both the data flows from the user's perspective and the data flows from a digital forensics perspective. The data flow from the user's perspective consists of using a browser on a computer system that is part of the university's WLAN. In Figure 3 the user activity can be abstracted by the browsers provided by the VM of the examiner's VM DG1. Its network infrastructure can access the OPAC Gateway G1, which then uses the Internet connection of the university to access the publisher's web server delivering the papers (and potentially accessing first and third party trackers).
From the digital forensics perspective the data flow starts by capturing the data traffic at the bridged network interface as DS_N from the examiner's VM DG1. The captured network packets when using the tools Section III-B TShark and the results of using Webbkoll, Website Evidence Collector, Ungoogled Chromium and the script gather_data.py are stored onto mass storage as DS_T (see also Section IV-A). The data from the data gathering step is then transferred to the mass storage DS_T of the analysis workstation AW1 for further investigation and analysis (see also Sections IV-B and IV-C).

Compared to to the system landscape description from the companion article [1], the landscape is also altered by using the Online Public Access Catalogue (OPAC) gateway OPAC G1 hosted by the library system of our university, which routes any searches using the OPAC and provides access to articles under the subscription scheme of our universities' library and allows to answer ***Research Question RQ1*** from Section I. This shows a slightly different flow of data and information but does not prevent Science-Tracking (see Section V). Extending the system landscape with the OPAC gateway enables simulating a typical scientific literature research scenario, which addresses ***Research Question RQ3*** (see Section I) by means of the ***Extension E4*** see (Section I).

V. Implementation of the Automation

This section describes the implemented environment of our research, our analysis tools and components of our automation, the latter forming our ***Extension E5*** (see Section I).

A. System and tool chain

For our research, multiple platforms are used (see Section III-B). The acquisition of research data is performed on the "tester stick" already used in [1], which is in essence a Debian-64-bit-based VM running inside VirtualBox [29] and configured to use a bridged network adapter configured for low noise acquisition of the incoming web traffic, i.e., the system itself and the browser are configured to not actively connect to the network outside the research context; automatic system and browser updates, safebrowsing, certificate updates etc., is disabled.
To show the independence from a particular OS after the data gathering step, the acquired data is processed on Windows10-based PC with an Intel i5-8600k CPU, 16 GB RAM. For both the headless browser and the interactively used browser we employ Ungoogled Chromium [30].
To keep the automation mostly OS-agnostic, the tool chain was implemented in Python 3.12.2, though some adjustments have to be done for the acquisition. This is necessary since the terminating of an asynchronous process needs different signals to be sent, depending on the OS.

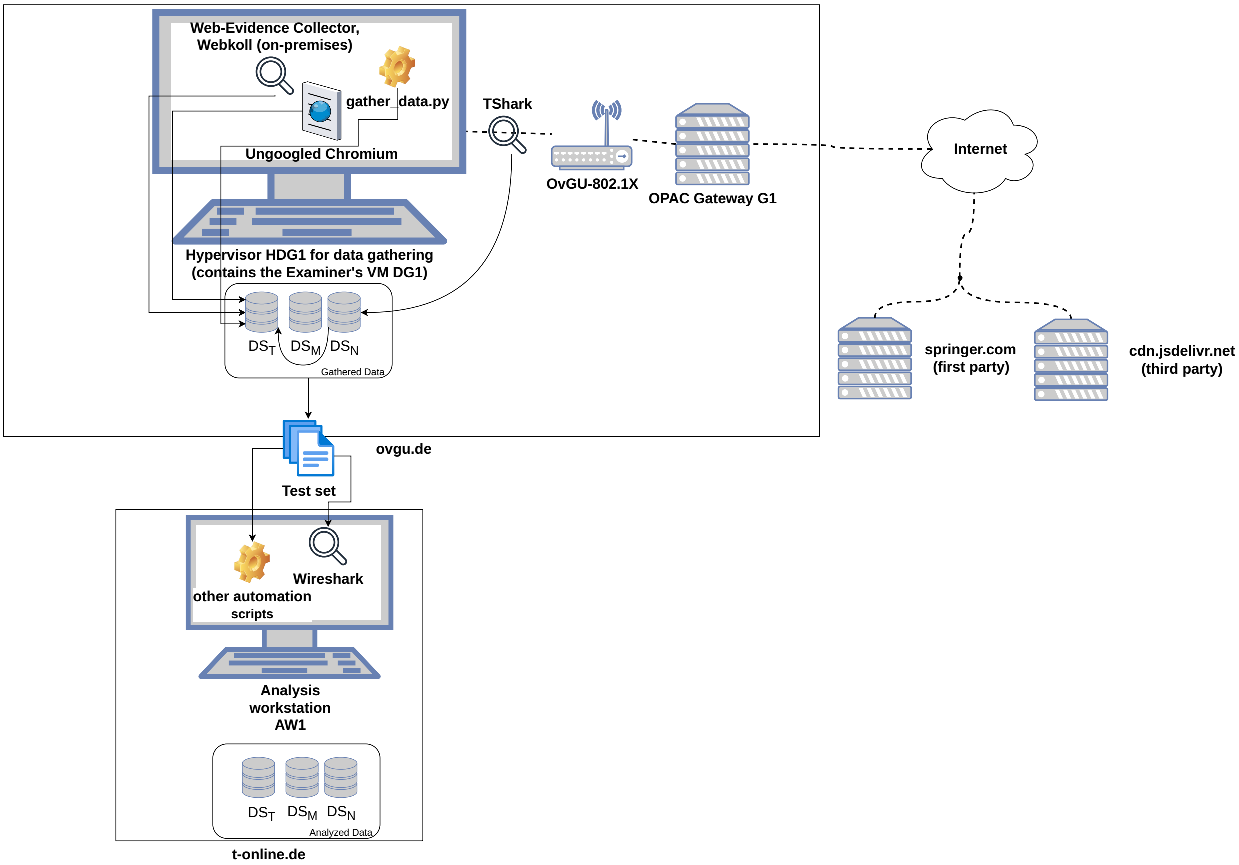


Figure 3. Simplified system landscape analysis for STF examinations visualizing components connections and data flows during the forensic examination, extending and focusing the research from [1]), the dashed lines represent the functional data flow from the user's perspective whilst solid lines represent the data flows of the examination from the digital forensics perspective.

B. Implementation of data acquisition and generation of results

This subsection describes the implementation of the aforementioned investigation concept in Section IV using 10 scripts (6 software tools and 4 automation scripts) in context to their respective steps.
The Webbkoll backend [28] (see Section III-B), is forked for necessary adjustments necessary for the automation, since consistent updating of the user-agent was necessary due to possible detection mechanisms of the publisher websites. The resulting .json file from the analysis is used for further estimation of third party trackers.
The Web-Evidence-Collector [21] (see Section III-B), is adjusted as well for the purposes of automation and a fork was created. Similar to Webbkoll backend, the user-agent is updated but also the tracker lists. A complete overview of the adjustments on both tools can be viewed in the commit list on the respective repositories.
In the following for the sake of brevity, we present pseudocode that represents the actions of our separate python scripts available from [31].

1) Data Gathering: For the implementation of the data gathering, a semi-parallel approach is used. A csv-table containing:

- the URL,
- the OPAC-URL,
- the publisher,
- an alternative URL from a different university,
- the state (open-access, non-open-access),

is provided as input, containing publications with their respective OPAC permalink and publisher website to be called. While the Webbkoll backend and TShark (see Section III-B) are started as an asynchronous process, the calls of the other analysis tools are initiated synchronously.
The TShark process is closed once the other tools completed analysis and restarted once the environment is ready for the next paper.
The Webbkoll backend only has to be terminated after the test series is completed, since the call to initiate the analysis is done by using curl [32].
The paper websites are called randomly with timeouts (randomized, 4-22 seconds) in between, to neither overload the publisher's server nor raise any suspicion, which could interfere with the acquisition process. Additionally, some timeouts are added, since the startup time of the asynchronous tools had to be considered.
The script *gather_data.py* implements the described approach.

```
TSharkInterface, PaperWebsiteList

startWebkollBackend()
randomize(PaperWebsiteList)
for paperWebsite in PaperWebsiteList:
    datetime = now()
    startTShark(TSharkInterface)
    webkollOutput <- webkollScan(paperWebsite)
    TSharkLogOutput = stopTShark()
    webEvidenceOutput <- webEvidenceCollect(
        paperWebsite)
    saveOutputsToFilesystem(webkollOutput,
        TSharkLogOutput, webEvidenceOutput)
stopWebkollBackend()
```

Listing 2. Pseudo algorithm for gathering data.

2) Data Investigation: The implementation of the generation of data investigation follows the pseudo-code algorithm in Listings 3 to 5. The general structure of the data investigation starts with importing and extracting DT_5 from the gathered sources (see Listing 3, lines 3 to 19). Once all information is gathered, the corresponding process to determine DT_5 and DT_3 matches, including the detection via tracking lists, is performed (see Listing 3, lines 20 to 27). The resulting STF is calculated iteratively for every host. The aggregation follows an iterative approach as well by first forming an intersected result table and generating from this the STF (see Listing 5).

Additional to the modules included in the distribution, the module Scapy [33], version 2.5.0, is used for processing the pcapng files from TShark and getting the required information from the DNS responses. As for its capabilities used in our research, Scapy offers extracting information from a byte stream, e.g., pcapng files, and presenting it in a human-readable format, like Wireshark does in its GUI. The generation utilizes an object-oriented approach to caching the output data from the tools and makes the script more readable.

Listing 3 describes the Generation of the result table and STF. It is implemented as *evaluator.py*.

```
1  Results, STF, Row, Trackerlists

3  WebkollData = parse(WebkollFile)
4  WebEvidenceData = parse(WebEvidenceFile)
5  TSharkData = parse(TSharkLog)

7  Hosts = getAllHosts(WebkollData, WebEvidenceData,
       TSharkData)

9  For every Host in Hosts:
10   If Host in WebkollData then
11     Row <- (WebkollHost: Hostname from WebkollData)
12     Row <- (WebkollIp: Ip from WebkollData)
13   If Host in WebEvidenceData then
14     Row <- (WebEvidenceHost: Hostname from
           WebEvidenceData)
15     Row <- (WebEvidenceParty: Party from
           WebEvidenceData)
16   If Host in TSharkData then
17     Row <- (TSharkHost: Hostname from TSharkData)
18     Row <- (TSharkIp: Ip from TSharkData)
19     Row <- (TSharkType: Type from TSharkData)
20   DT_5 = checkHosts(Row)
21   Row <- DT5
22   DT3 = checkTracker(Row, Trackerlists)
23   Row <- DT3
24   Results <- Row
25   If DT3 atleast UNC then
26     STF = updateSTF(STF, Row)
27     clear(Row)
28 createTable Results, STF
```

Listing 3. Generation of result table and STF.

Listing 4 outlines the update procedure for the generated STFs and is implemented in *evaluator.py* and *aggregation.py*.

```
Row, STF

If Row is CNAME then
    STF(Row->DT5,Row->DT3,Row->WebEvidenceParty,A-
        Record) += 1
    STF(Row->DT5,Row->DT3,Row->WebEvidenceParty,
        CNAME) += 1
Else
    STF(Row->DT5,Row->DT3,Row->WebEvidenceParty,A-
        Record) += 1
return STF
```

Listing 4. Updating the STF.

The aggregation of results for multiple papers is shown in Listing 5. It is implemented in the script *aggregation.py*. The scripts *auto_evaluation.py* and *automate_generate_eval_stuff.bat* automate the process of result table and STF generation.

```
Results, STF, Row, Trackerlists

Papers = getPaperResults(Filter...)
For every Paper in Papers:
  If Results is empty then
        Results = Paper
    Else
        Results = intersect(Results, Paper)
    If Results is empty then
        stop

STF = evaluateSTF(Results)
createTable(Results, STF)
```

Listing 5. Aggregation of results and STFs.

3) Data Analysis: The implementation of the data analysis focuses on the comparison of STFs and estimation of their deviation from each other.

The *fnc_module.py* provides the functions *scan_stf*, reading a STF that was saved to the disk, and *analyze_stfs*, comparing two STFs and generating a report as well as an estimation value for the deviation. The function *analyze_stfs* implements the metric from Section IV-C for grading the deviation of two STFs.

The script *diversity_analysis.py* performs the intra-publisher analysis and is automated over all test sets and publishers with the script *call_diversity_analysis.py*. The two aforementioned scripts focus on the implementation of a intra-publisher analysis.

The remainder of the functionality needed for our research is implemented in the scripts *inter_pub_diversity_analysis.py*. Those compare two STFs and generating reports. Further, *inter_pub_diversity_auto.py*, automate the process and generating a complete report. That functionality encompasses:

- an inter-publisher analysis within one test series,
- an inter-test-series analysis and an analysis of differences between open access and non open access literature.

All reports are generated as a CSV file and our results can be found in our provided repository at [31].

```
Path, Filters

STF = initializeSTF()
files = listFilesIn(Path)
for file in files:
    if not isDir(file) and filenameStartsWith(file, ”stf”) and satisfiesFilters(file, Filters):
        for row in file:
            category <- determineCategory(row)
            STF <- readRow(category, row)
return STF
```

Listing 6. Pseudo-code algortihm of scan_stf.

```
STF, ReferenceSTF, Title

deviation, totalDeviation, referenceSTFtotalSize = 0
deviationList = []

stfDifferences = initializeSTF()
report = intializeReport(Title)

for row in rows(stfDifferences):
    rowSizeReferenceSTF = sum(values(ReferenceSTF[row]))
    referenceSTFtotalSize += rowSizeReferenceSTF
    stfDifferences <- getDifference(STF[row], ReferenceSTF[row])
    differenceRowSize = sum(stfDifferences[row])
    if rowSizeReferenceSTF is 0:
        deviation <- 0.0
        report <- stfDifferences, ’not gradable’
    else:
        deviation = abs(differenceRowSize / rowSizeReferenceSTF)
        report <- stfDifferences, deviation
    deviationList <- (deviation, differenceRowSize)

for (deviation, differenceRowSize) in deviationList:
    totalDeviation += deviation * (differenceRowSize / referenceSTFtotalSize)

finishUp(report)
return report, totalDeviation
```

Listing 7. Pseudo-code algorithm of analyse_stfs.

```
Input: TestSeries, Publisher, Version, Reference

intraPublisherReport = initializeReport(Publisher, Version)
referenceSTF = scanSTF(Reference)

for paper in TestSeries:
    if isPublisher(paper, Publisher) and isVersion(paper, Version):
        stf <- scan_stf(paper, Version)
        paperReport, totalDeviation <- analyse_stfs(stf, referenceSTF)
        intraPublisherReport <- addToReport(paper, totalDeviation, Version)
        saveToFS(paperReport)

finishUp(intraPublisherReport)
saveToFS(intraPublisherReport)
```

Listing 8. Intra-publisher analysis for a specific test series.

```
FstPaper, SndPaper, FstPublisher, SndPublisher, TestSeries, Versions, CompleteReport

for version in Versions:
    fstStf <- scan_stf(getFile(FstPaper, TestSeries[0]))
    if SndPaper not undefined:
        sndStf <- scan_stf(getFile(SndPaper, TestSeries[0]))
    else:
        sndStf <- scan_stf(getFile(FstPaper, TestSeries[1]))
    comparisonReport, deviation <- analyse_stfs(fstStf, sndStf)
    saveToFS(comparisonReport)
    if SndPaper not undefined:
        CompleteReport <- addToReport(FstPublisher, SndPublisher, version, deviation)
    else:
        CompleteReport <- addToReport(FstPublisher, TestSeries, version, deviation)
```

Listing 9. General algortihm for comparing two STFs.

In the following section we evaluate the approach from Section IV in its implementation as described in Section V.

VI. Evaluation

The automation is tested successfully and enables to process a larger amount of literature in a smaller time frame than in [1]. The tool chain enables an almost OS-agnostic automation approach for the generation of the STF (excluding the data gathering step). About 2.1 GB of research data is collected due to the use of the tool chain.

Due to the greater set of publications, some additional insights are gained into the capabilities of the STF. With the automation, a more fine granular examination on the publisher can be performed.

The investigation centres around gaining a greater insight into the possible tracking behaviour on the website of publisher with respect to different points of acquiring a publication and open access status (see Sections IV and V).

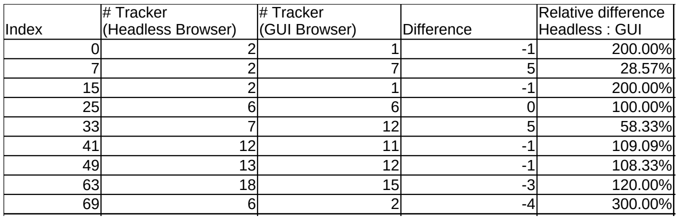
The results in Section VI-A to Section VI-G show, that using the OPAC gateway from our university does not prevent tracking, answering ***Research Question RQ1*** (see Section I). Furthermore, the influence of time (t_i, t_{i+1} see [1]) regarding the specific date the tracker lists are acquired, on the recognition/classification of third party hosts is investigated. Also, as an additional comparison to the publishers investigated in [1], the publisher Wiley is also partially added; only in test series 2 and 3 due to mid-experiment inclusion after examination of test series 1.

A. Influence of browser type on tracking

This subsection describes our findings of our research about the influence of the browser type used on tracking behaviour mentioned in Section IV-A and enhances our findings described in the companion article [1] as ***Extension E8*** (see Section I). The recordings of our comparative research are saved in a repository and can be provided on request.

Our first observation shows a difference in the recorded network traffic in all test cases. As for our second observation, there is no clear trend in behaviour depending on the type of browser. In both cases, one type gathered more data than the other.

In the following, we show an example using our list of literature represented in the file paper.csv, which can be viewed in our repository [31]. On paper No.7 the recordings of the graphical browser show, that 5 URLs have been additionally called in comparison to the headless browser. But on paper No.15 only 1 host has been called on the graphical browser. Further information is contained in Figure 4. It can be surmised based on the results of this comparative research that there is an influence of the browser type. We argue for the usage of the headless browser on the grounds of getting results on a larger scale although we are potentially missing some trackers by using the headless browser.



Index	# Tracker (Headless Browser)	# Tracker (GUI Browser)	Difference	Relative difference Headless : GUI
0	2	1	-1	200.00%
7	2	7	5	28.57%
15	2	1	-1	200.00%
25	6	6	0	100.00%
33	7	12	5	58.33%
41	12	11	-1	109.09%
49	13	12	-1	108.33%
63	18	15	-3	120.00%
69	6	2	-4	300.00%

Figure 4. Results of the probe for tracking based on browser type.

Finding the source of the different tracking behaviour, however, is a very valid research goal for future work.

B. Time dependency of tracker lists

This research enhances the companion article [1] as ***Extension E6*** (see)Section I). During our investigation, tracker lists are downloaded from every provider at the last time of change before the acquisition.

To check if the classification behaviour changes over a short period of time, the tracker list versions are grouped by a date that signifies the last change on one of the lists before the recording, and applied on every test series.

The generated reports, e.g., Figure 14, show the same results and thus being independent of the version of the tracker lists for the tests conducted. Future research should examine the time dependency over a broader time span.

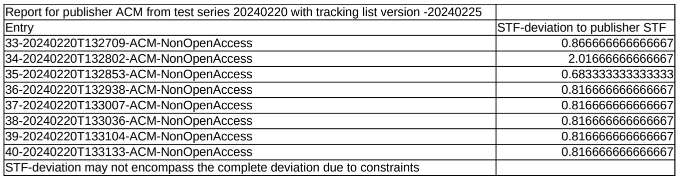
Due to these results, some of the result tables will be abridged due to there not being any benefit for showing the results with respect to every version of the tracker lists. The full set of results can be found at [31].

C. Intra-publisher diversity

By comparing the result tables and STF of single papers with the intersected results of a test series during our research enhancing our companion article [1] with the ***Extension E3*** (see Section I), a diversity within the third party hosts classified as probable tracker (classified via external data in the form of tracker lists, see Section III-B) is observed in all but one publisher, namely Springer. Figures 5 to 11 show exemplary, how strongly the STF of a paper can differ from the intersected STF of its publisher in comparison to its peers.

The papers from the publisher ACM show the second strongest intra-publisher diversity. Throughout every test series, there is no paper that does not match the intersection completely. Also, in comparison to Figure 12, every deviation value is higher. An interesting detail is that some deviation values are the same and on closer inspection with the STFs, the STFs are the same. While this is no proof that the same deviation signals an equal STF, it may indicate heterogeneity in the set of STFs.

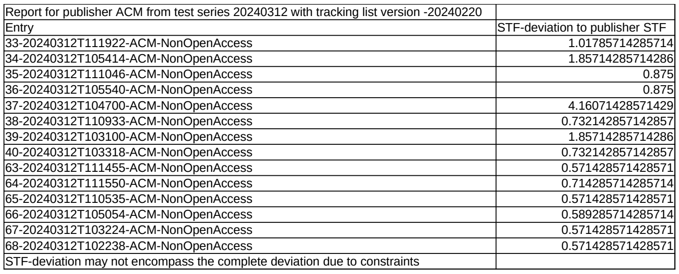
The following Figure 5 shows the intra publisher diversity based on the STF-deviation for the publisher ACM according to the test series conducted at 20/02/2024.



Report for publisher ACM from test series 20240220 with tracking list version -20240225	
Entry	STF-deviation to publisher STF
33-20240220T132709-ACM-NonOpenAccess	0.866666666666667
34-20240220T132802-ACM-NonOpenAccess	2.01666666666667
35-20240220T132853-ACM-NonOpenAccess	0.683333333333333
36-20240220T132938-ACM-NonOpenAccess	0.816666666666667
37-20240220T133007-ACM-NonOpenAccess	0.816666666666667
38-20240220T133036-ACM-NonOpenAccess	0.816666666666667
39-20240220T133104-ACM-NonOpenAccess	0.816666666666667
40-20240220T133133-ACM-NonOpenAccess	0.816666666666667
STF-deviation may not encompass the complete deviation due to constraints	

Figure 5. Intra-publisher comparison results for ACM (test series 2024-02-20).

The following Figure 6 shows the intra publisher diversity based on the STF-deviation for the publisher ACM according to the test series conducted at 12/03/2024.

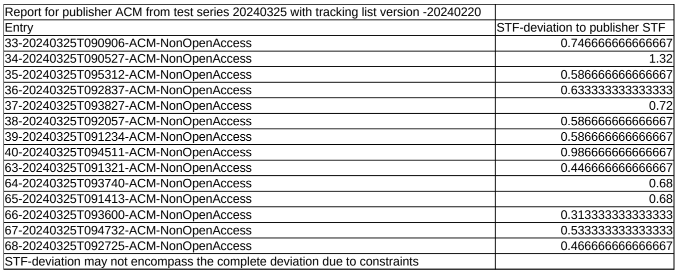


Report for publisher ACM from test series 20240312 with tracking list version -20240220	
Entry	STF-deviation to publisher STF
33-20240312T111922-ACM-NonOpenAccess	1.01785714285714
34-20240312T105414-ACM-NonOpenAccess	1.85714285714286
35-20240312T111046-ACM-NonOpenAccess	0.875
36-20240312T105540-ACM-NonOpenAccess	0.875
37-20240312T104700-ACM-NonOpenAccess	4.16071428571429
38-20240312T110933-ACM-NonOpenAccess	0.732142857142857
39-20240312T103100-ACM-NonOpenAccess	1.85714285714286
40-20240312T103318-ACM-NonOpenAccess	0.732142857142857
63-20240312T111455-ACM-NonOpenAccess	0.571428571428571
64-20240312T111550-ACM-NonOpenAccess	0.714285714285714
65-20240312T110535-ACM-NonOpenAccess	0.571428571428571
66-20240312T105054-ACM-NonOpenAccess	0.589285714285714
67-20240312T103224-ACM-NonOpenAccess	0.571428571428571
68-20240312T102238-ACM-NonOpenAccess	0.571428571428571
STF-deviation may not encompass the complete deviation due to constraints	

Figure 6. Intra-publisher comparison results for ACM (test series 2024-03-12).

The following Figure 7 shows the intra publisher diversity based on the STF-deviation for the publisher ACM according to the test series conducted at 25/03/2024.

From Figures 8 and 9 it can be assumed, that in that specific test series the STFs of the papers from Elsevier show a strong difference in observed tracking behaviour and a mentionable intersection between the STFs could not be formed.

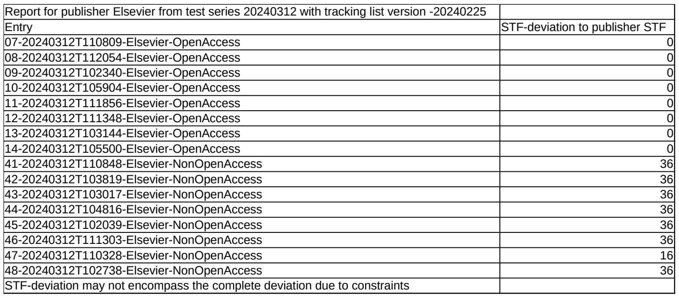


Report for publisher ACM from test series 20240325 with tracking list version -20240220	
Entry	STF-deviation to publisher STF
33-20240325T090906-ACM-NonOpenAccess	0.746666666666667
34-20240325T090527-ACM-NonOpenAccess	1.32
35-20240325T095312-ACM-NonOpenAccess	0.586666666666667
36-20240325T092837-ACM-NonOpenAccess	0.633333333333333
37-20240325T093827-ACM-NonOpenAccess	0.72
38-20240325T092057-ACM-NonOpenAccess	0.586666666666667
39-20240325T091234-ACM-NonOpenAccess	0.586666666666667
40-20240325T094511-ACM-NonOpenAccess	0.986666666666667
63-20240325T091321-ACM-NonOpenAccess	0.446666666666667
64-20240325T093740-ACM-NonOpenAccess	0.68
65-20240325T091413-ACM-NonOpenAccess	0.68
66-20240325T093600-ACM-NonOpenAccess	0.313333333333333
67-20240325T094732-ACM-NonOpenAccess	0.533333333333333
68-20240325T092725-ACM-NonOpenAccess	0.466666666666667
STF-deviation may not encompass the complete deviation due to constraints	

Figure 7. Intra-publisher comparison results for ACM (test series 2024-03-25).

This might also be connected to the findings in Section VI-E, as they show a difference in tracking behaviour between open access and non-open access groups.

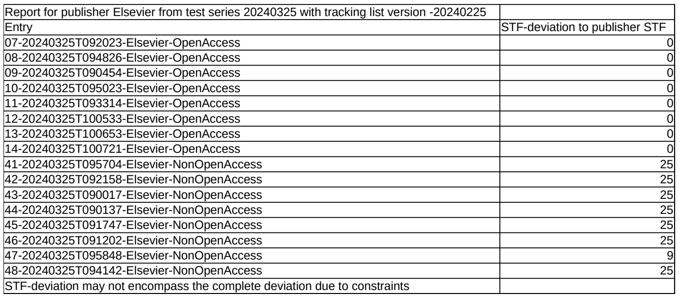
It should also be mentioned that there are STFs of Elsevier publications in test series 2024-02-20 but since no tracking data is gathered for some publications, the intersected STF was empty. Therefore, an analysis on the intra-publisher diversity is impossible for this test series. The following Figure 8 shows the intra publisher diversity based on the STF-deviation for the publisher Elsevier according to the test series conducted at 12/03/2024.



Report for publisher Elsevier from test series 20240312 with tracking list version -20240225	
Entry	STF-deviation to publisher STF
07-20240312T110809-Elsevier-OpenAccess	0
08-20240312T112054-Elsevier-OpenAccess	0
09-20240312T102340-Elsevier-OpenAccess	0
10-20240312T105904-Elsevier-OpenAccess	0
11-20240312T111856-Elsevier-OpenAccess	0
12-20240312T111348-Elsevier-OpenAccess	0
13-20240312T103144-Elsevier-OpenAccess	0
14-20240312T105500-Elsevier-OpenAccess	0
41-20240312T110848-Elsevier-NonOpenAccess	36
42-20240312T103819-Elsevier-NonOpenAccess	36
43-20240312T103017-Elsevier-NonOpenAccess	36
44-20240312T104816-Elsevier-NonOpenAccess	36
45-20240312T102039-Elsevier-NonOpenAccess	36
46-20240312T111303-Elsevier-NonOpenAccess	36
47-20240312T110328-Elsevier-NonOpenAccess	16
48-20240312T102738-Elsevier-NonOpenAccess	36
STF-deviation may not encompass the complete deviation due to constraints	

Figure 8. Intra-publisher comparison results for Elsevier (test series 2024-03-12).

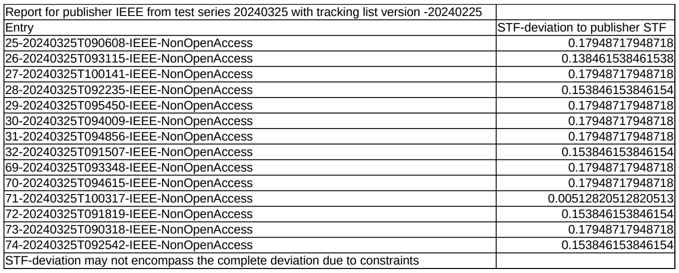
The following Figure 9 shows the intra publisher diversity based on the STF-deviation for the publisher Elsevier according to the test series conducted at 25/03/2024.



Report for publisher Elsevier from test series 20240325 with tracking list version -20240225	
Entry	STF-deviation to publisher STF
07-20240325T092023-Elsevier-OpenAccess	0
08-20240325T094826-Elsevier-OpenAccess	0
09-20240325T090454-Elsevier-OpenAccess	0
10-20240325T095023-Elsevier-OpenAccess	0
11-20240325T093314-Elsevier-OpenAccess	0
12-20240325T100533-Elsevier-OpenAccess	0
13-20240325T100653-Elsevier-OpenAccess	0
14-20240325T100721-Elsevier-OpenAccess	0
41-20240325T095704-Elsevier-NonOpenAccess	25
42-20240325T092158-Elsevier-NonOpenAccess	25
43-20240325T090017-Elsevier-NonOpenAccess	25
44-20240325T090137-Elsevier-NonOpenAccess	25
45-20240325T091747-Elsevier-NonOpenAccess	25
46-20240325T091202-Elsevier-NonOpenAccess	25
47-20240325T095848-Elsevier-NonOpenAccess	9
48-20240325T094142-Elsevier-NonOpenAccess	25
STF-deviation may not encompass the complete deviation due to constraints	

Figure 9. Intra-publisher comparison results for Elsevier (test series 2024-03-25).

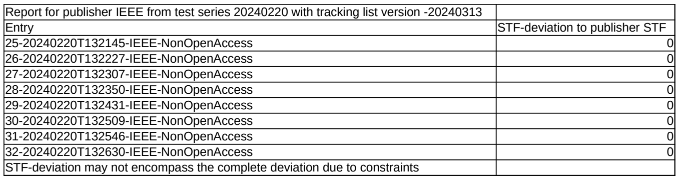
The findings for the paper from the publisher IEEE show almost no intra-publisher diversity throughout the whole test series. Still, even in the case shown in Figure 12 the intra-publisher diversity is low in comparison to other publishers like ACM or Elsevier. This might indicate that the intersected STF of IEEE encompasses almost every detected host of the test series or in the case Figures 10 and 11 every host. The following Figure 10 shows the intra publisher diversity based on the STF-deviation for the publisher IEEE according to the test series conducted at 20/02/2024.



Report for publisher IEEE from test series 20240325 with tracking list version -20240225	
Entry	STF-deviation to publisher STF
25-20240325T090608-IEEE-NonOpenAccess	0.17948717948718
26-20240325T093115-IEEE-NonOpenAccess	0.138461538461538
27-20240325T100141-IEEE-NonOpenAccess	0.17948717948718
28-20240325T092235-IEEE-NonOpenAccess	0.153846153846154
29-20240325T095450-IEEE-NonOpenAccess	0.17948717948718
30-20240325T094009-IEEE-NonOpenAccess	0.17948717948718
31-20240325T094856-IEEE-NonOpenAccess	0.17948717948718
32-20240325T091507-IEEE-NonOpenAccess	0.153846153846154
69-20240325T093348-IEEE-NonOpenAccess	0.17948717948718
70-20240325T094615-IEEE-NonOpenAccess	0.17948717948718
71-20240325T100317-IEEE-NonOpenAccess	0.00512820512820513
72-20240325T091819-IEEE-NonOpenAccess	0.153846153846154
73-20240325T090318-IEEE-NonOpenAccess	0.17948717948718
74-20240325T092542-IEEE-NonOpenAccess	0.153846153846154
STF-deviation may not encompass the complete deviation due to constraints	

Figure 10. Intra-publisher comparison results for IEEE (test series 2024-02-20).

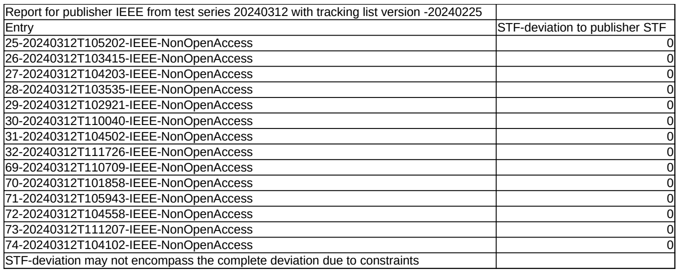
The following Figure 11 shows the intra publisher diversity based on the STF-deviation for the publisher IEEE according to the test series conducted at 12/03/2024.



Report for publisher IEEE from test series 20240220 with tracking list version -20240313	
Entry	STF-deviation to publisher STF
25-20240220T132145-IEEE-NonOpenAccess	0
26-20240220T132227-IEEE-NonOpenAccess	0
27-20240220T132307-IEEE-NonOpenAccess	0
28-20240220T132350-IEEE-NonOpenAccess	0
29-20240220T132431-IEEE-NonOpenAccess	0
30-20240220T132509-IEEE-NonOpenAccess	0
31-20240220T132546-IEEE-NonOpenAccess	0
32-20240220T132630-IEEE-NonOpenAccess	0
STF-deviation may not encompass the complete deviation due to constraints	

Figure 11. Intra-publisher comparison results for IEEE (test series 2024-03-12).

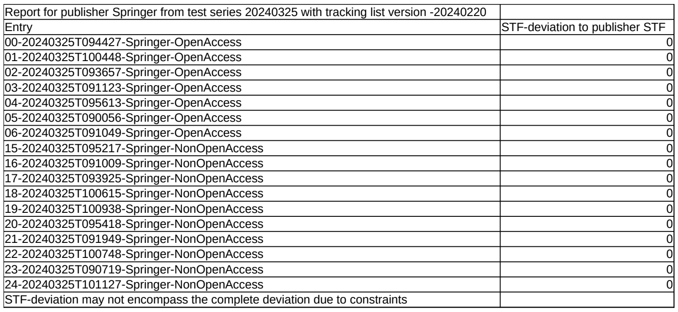
The following Figure 12 shows the intra publisher diversity based on the STF-deviation for the publisher IEEE according to the test series conducted at 25/03/2024.



Report for publisher IEEE from test series 20240312 with tracking list version -20240225	
Entry	STF-deviation to publisher STF
25-20240312T105202-IEEE-NonOpenAccess	0
26-20240312T103415-IEEE-NonOpenAccess	0
27-20240312T104203-IEEE-NonOpenAccess	0
28-20240312T103535-IEEE-NonOpenAccess	0
29-20240312T102921-IEEE-NonOpenAccess	0
30-20240312T110040-IEEE-NonOpenAccess	0
31-20240312T104502-IEEE-NonOpenAccess	0
32-20240312T111726-IEEE-NonOpenAccess	0
69-20240312T110709-IEEE-NonOpenAccess	0
70-20240312T101858-IEEE-NonOpenAccess	0
71-20240312T105943-IEEE-NonOpenAccess	0
72-20240312T104558-IEEE-NonOpenAccess	0
73-20240312T111207-IEEE-NonOpenAccess	0
74-20240312T104102-IEEE-NonOpenAccess	0
STF-deviation may not encompass the complete deviation due to constraints	

Figure 12. Intra-publisher comparison results for IEEE (test series 2024-03-25).

In the set of publications of the publisher Springer during the test period no diversity in the set of classified trackers is observable, see Figure 13. This behaviour within STFs of publications from Springer is observable throughout the complete test series and during our tests is unique to the publisher Springer.



Report for publisher Springer from test series 20240325 with tracking list version -20240220	
Entry	STF-deviation to publisher STF
00-20240325T094427-Springer-OpenAccess	0
01-20240325T100448-Springer-OpenAccess	0
02-20240325T093657-Springer-OpenAccess	0
03-20240325T091123-Springer-OpenAccess	0
04-20240325T095613-Springer-OpenAccess	0
05-20240325T090056-Springer-OpenAccess	0
06-20240325T091049-Springer-OpenAccess	0
15-20240325T095217-Springer-NonOpenAccess	0
16-20240325T091009-Springer-NonOpenAccess	0
17-20240325T093925-Springer-NonOpenAccess	0
18-20240325T100615-Springer-NonOpenAccess	0
19-20240325T100938-Springer-NonOpenAccess	0
20-20240325T095418-Springer-NonOpenAccess	0
21-20240325T091949-Springer-NonOpenAccess	0
22-20240325T100748-Springer-NonOpenAccess	0
23-20240325T090719-Springer-NonOpenAccess	0
24-20240325T101127-Springer-NonOpenAccess	0
STF-deviation may not encompass the complete deviation due to constraints	

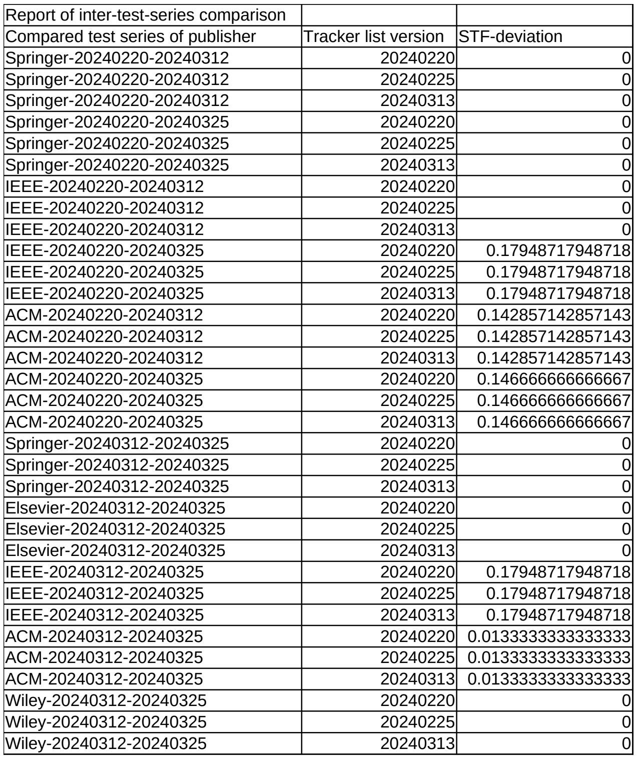
Figure 13. Intra-publisher comparison results for Springer (test series 2024-03-25).

In the following we compare different publishers using the proposed STF-deviation metric.

D. Inter-test series diversity

Besides checking for diversity within the set of probable tracker within one test series, the results in between test series are compared for any observable difference. This enables the detection of a potential diversity in the time dimension, enhancing our research from the companion article [1] as the ***Extension E3*** (see Section I).

Figure 14 shows the results of the comparison within our tests.



Report of inter-test-series comparison		
Compared test series of publisher	Tracker list version	STF-deviation
Springer-20240220-20240312	20240220	0
Springer-20240220-20240312	20240225	0
Springer-20240220-20240312	20240313	0
Springer-20240220-20240325	20240220	0
Springer-20240220-20240325	20240225	0
Springer-20240220-20240325	20240313	0
IEEE-20240220-20240312	20240220	0
IEEE-20240220-20240312	20240225	0
IEEE-20240220-20240312	20240313	0
IEEE-20240220-20240325	20240220	0.17948717948718
IEEE-20240220-20240325	20240225	0.17948717948718
IEEE-20240220-20240325	20240313	0.17948717948718
ACM-20240220-20240312	20240220	0.142857142857143
ACM-20240220-20240312	20240225	0.142857142857143
ACM-20240220-20240312	20240313	0.142857142857143
ACM-20240220-20240325	20240220	0.146666666666667
ACM-20240220-20240325	20240225	0.146666666666667
ACM-20240220-20240325	20240313	0.146666666666667
Springer-20240312-20240325	20240220	0
Springer-20240312-20240325	20240225	0
Springer-20240312-20240325	20240313	0
Elsevier-20240312-20240325	20240220	0
Elsevier-20240312-20240325	20240225	0
Elsevier-20240312-20240325	20240313	0
IEEE-20240312-20240325	20240220	0.17948717948718
IEEE-20240312-20240325	20240225	0.17948717948718
IEEE-20240312-20240325	20240313	0.17948717948718
ACM-20240312-20240325	20240220	0.0133333333333333
ACM-20240312-20240325	20240225	0.0133333333333333
ACM-20240312-20240325	20240313	0.0133333333333333
Wiley-20240312-20240325	20240220	0
Wiley-20240312-20240325	20240225	0
Wiley-20240312-20240325	20240313	0

Figure 14. Results of the comparison of intersected STFs between test series.

In regard to the inter-test series diversity, it is observed that, except for the publisher Springer, every publisher has between at least two test series differences in the intersected STFs, see Figure 14.

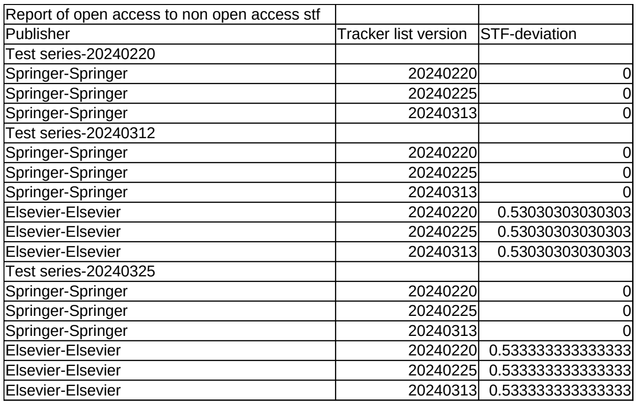
This might indicate that changes in tracking behaviour reflect on the STF and can therefore be noticed by the application of the STF.

E. Intra-publisher differences for open access and non-open access papers (OA/NOA)

The investigation of possible differences within the observed trackers answers the ***Research Question RQ4*** (see Section I). It is, however, limited by the constraints of our approach, environment and tools. For instance, OPAC only listed open access publications from the publishers Springer and Elsevier. IEEE and ACM do feature open access publications, but, at least in the case of IEEE, during our tests open access publications are not offered on the publisher's usual website (e.g., IEEE Xplore) but rather a platform specifically for open access publications. ACM itself offers open access literature through searching specifically for it within OPAC results in matches (e.g., using filters for publisher and keyword *open access* or *non-open access*). This necessitates specialized queries.

In addition, a problem is encountered with Webbkoll and Elsevier open access publications, which results in failure to acquire analysis data, and therefore no tracker could be classified plausible for DT_3 or DT_5. From the available analysis data for Elsevier publications a difference in classified trackers between open access and non-open access publications could be observed, see Figure 15.

As for the paper from the publisher Springer, no deviations were observable in the data sets.



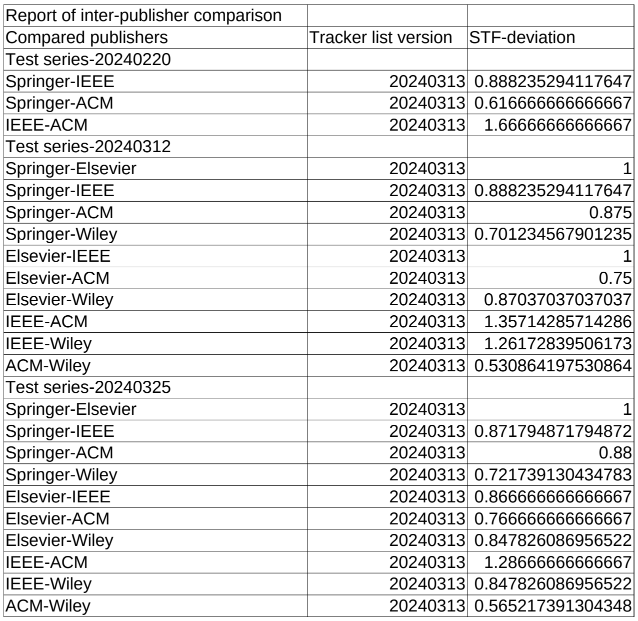
Report of open access to non open access stf		
Publisher	Tracker list version	STF-deviation
Test series-20240220		
Springer-Springer	20240220	0
Springer-Springer	20240225	0
Springer-Springer	20240313	0
Test series-20240312		
Springer-Springer	20240220	0
Springer-Springer	20240225	0
Springer-Springer	20240313	0
Elsevier-Elsevier	20240220	0.53030303030303
Elsevier-Elsevier	20240225	0.53030303030303
Elsevier-Elsevier	20240313	0.53030303030303
Test series-20240325		
Springer-Springer	20240220	0
Springer-Springer	20240225	0
Springer-Springer	20240313	0
Elsevier-Elsevier	20240220	0.533333333333333
Elsevier-Elsevier	20240225	0.533333333333333
Elsevier-Elsevier	20240313	0.533333333333333

Figure 15. Results of the comparison of open access to non-open access literature.

Future work should point to an enhanced environment and tools to address the existing challenges.

F. Inter-publisher difference

With the automated approach, a comparison of the intersected STFs of different publishers is performed successfully, enhancing the findings from our companion article [1] as the ***Extension E1*** (see Section I). Figure 16 shows an abridged version of the complete reports, since there is no need to consider the different versions of tracking lists due to the mentioned points in Section VI-B.



Report of inter-publisher comparison		
Compared publishers	Tracker list version	STF-deviation
Test series-20240220		
Springer-IEEE	20240313	0.888235294117647
Springer-ACM	20240313	0.616666666666667
IEEE-ACM	20240313	1.66666666666667
Test series-20240312		
Springer-Elsevier	20240313	1
Springer-IEEE	20240313	0.888235294117647
Springer-ACM	20240313	0.875
Springer-Wiley	20240313	0.701234567901235
Elsevier-IEEE	20240313	1
Elsevier-ACM	20240313	0.75
Elsevier-Wiley	20240313	0.87037037037037
IEEE-ACM	20240313	1.35714285714286
IEEE-Wiley	20240313	1.26172839506173
ACM-Wiley	20240313	0.530864197530864
Test series-20240325		
Springer-Elsevier	20240313	1
Springer-IEEE	20240313	0.871794871794872
Springer-ACM	20240313	0.88
Springer-Wiley	20240313	0.721739130434783
Elsevier-IEEE	20240313	0.866666666666667
Elsevier-ACM	20240313	0.766666666666667
Elsevier-Wiley	20240313	0.847826086956522
IEEE-ACM	20240313	1.28666666666667
IEEE-Wiley	20240313	0.847826086956522
ACM-Wiley	20240313	0.565217391304348

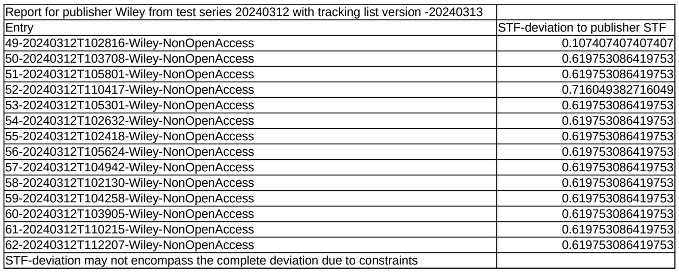
Figure 16. Results of the inter-publisher comparison (abridged).

A complete version with all tracker list version can be found in [31]. The results of Figure 16 show that there is a noticeable difference in tracking behaviour between publishers, which could give hints/leads towards identifying specific publishers based on their tracking behaviour (see also our companion conference article [1]. It can be assumed, based on our results, that the tracking behaviour may strongly differ from publisher to publisher. Future research on an even larger scale (both in number of papers and the time span observed) is needed to have a qualified opinion as to how discriminating the STF-deviation with respect to publishers is.

The full set of tables is available under [31].

G. Addendum Wiley

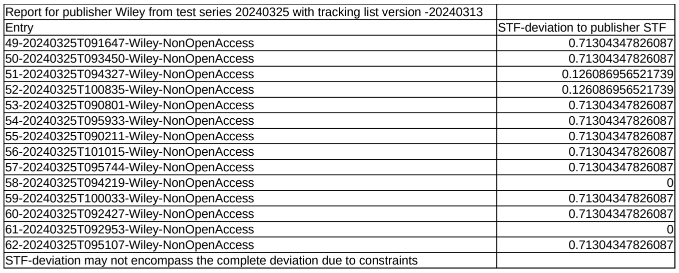
The publisher Wiley is additionally investigated to expand our group of subjects using the same setup and procedures, enhancing the findings from our companion article [1] as the ***Extension E7*** (see Section I). As Wiley is added mid-investigation, publications of it are only considered in the second and third test series. The following Figure 17 shows the intra-publisher comparison results for the publisher Wiley from the test series conducted at 12/03/2024.



Report for publisher Wiley from test series 20240312 with tracking list version -20240313	
Entry	STF-deviation to publisher STF
49-20240312T102816-Wiley-NonOpenAccess	0.107407407407407
50-20240312T103708-Wiley-NonOpenAccess	0.619753086419753
51-20240312T105801-Wiley-NonOpenAccess	0.619753086419753
52-20240312T110417-Wiley-NonOpenAccess	0.716049382716049
53-20240312T105301-Wiley-NonOpenAccess	0.619753086419753
54-20240312T102632-Wiley-NonOpenAccess	0.619753086419753
55-20240312T102418-Wiley-NonOpenAccess	0.619753086419753
56-20240312T105624-Wiley-NonOpenAccess	0.619753086419753
57-20240312T104942-Wiley-NonOpenAccess	0.619753086419753
58-20240312T102130-Wiley-NonOpenAccess	0.619753086419753
59-20240312T104258-Wiley-NonOpenAccess	0.619753086419753
60-20240312T103905-Wiley-NonOpenAccess	0.619753086419753
61-20240312T110215-Wiley-NonOpenAccess	0.619753086419753
62-20240312T112207-Wiley-NonOpenAccess	0.619753086419753
STF-deviation may not encompass the complete deviation due to constraints	

Figure 17. Intra-publisher comparison results for Wiley (test series 2024-03-12).

The following Figure 18 shows the intra-publisher comparison results for the publisher Wiley from the test series conducted at 25/03/2024.



Report for publisher Wiley from test series 20240325 with tracking list version -20240313	
Entry	STF-deviation to publisher STF
49-20240325T091647-Wiley-NonOpenAccess	0.71304347826087
50-20240325T093450-Wiley-NonOpenAccess	0.71304347826087
51-20240325T094327-Wiley-NonOpenAccess	0.126086956521739
52-20240325T100835-Wiley-NonOpenAccess	0.126086956521739
53-20240325T090801-Wiley-NonOpenAccess	0.71304347826087
54-20240325T095933-Wiley-NonOpenAccess	0.71304347826087
55-20240325T090211-Wiley-NonOpenAccess	0.71304347826087
56-20240325T101015-Wiley-NonOpenAccess	0.71304347826087
57-20240325T095744-Wiley-NonOpenAccess	0.71304347826087
58-20240325T094219-Wiley-NonOpenAccess	0
59-20240325T100033-Wiley-NonOpenAccess	0.71304347826087
60-20240325T092427-Wiley-NonOpenAccess	0.71304347826087
61-20240325T092953-Wiley-NonOpenAccess	0
62-20240325T095107-Wiley-NonOpenAccess	0.71304347826087
STF-deviation may not encompass the complete deviation due to constraints	

Figure 18. Intra-publisher comparison results for Wiley (test series 2024-03-25).

In Figures 17 and 18 it is shown that there is an intra-publisher diversity within the tracking. The deviations seem to form a middle ground between ACM and IEEE, compared to the findings in Section VI-C. Besides being analysed for intra-publisher diversity, an inter-publisher comparison as well as an inter-test series comparison is conducted. Their results are shown in Figures 14 and 16 and indicate, that there was no significant difference in the tracking behaviour over time, but the tracking behaviour deviates from other publishers. Since no open access publications from Wiley in OPAC are to be found by filtering and keyword search during our tests, no examinations with respect to open-access status are conducted at the time of the research.

While those results are not a full addition to the test series, they still show a tendency and underline the unique result position for the publisher Springer so far.

VII. Conclusion and future work

In this article we extended the work from the companion conference article [1] centred around the topic of Science-Tracking and the usage of the Science-Tracking Fingerprint (STF) as a means to gain hints for the originator of the tracking. The extension covers 8 separate aspects.

First we altered the system landscape by measuring the amount of Science-Tracking behind our universities' Online Public Access Catalog (OPAC) in order to see whether the Science-Tracking is altered by tunnelling our paper requests and downloads through that system. This is not the case according to our current results. Even after placing queries through this OPAC system, tracking by the publishers still takes place. We swapped broadness for detail and thus restricted ourselves to the examination of Web-based Science-Tracking.

Secondly, as placed in the future work section of [1], we automated the processes for the detection of Science-Tracking and the calculation of the STF. In total 10 scripts (6 Software tools and 4 automation scripts) that cover mostly the steps of data gathering and data investigation were released as Open Source. We also changed the number of lists of known trackers from originally 1 to 3 to increase the hit ratio for known tracker domains. We were able to examine 60 papers from 4 selected publishers.

Enabled by the automatization and larger numbers of STF-based examinations as a result (60 in total for all examinations), we could observe multiple documents from an individual publisher at 3 different points in time to obtain a measure for the intra-publisher diversity using the Science-Tracking Fingerprint. Our results show for 3 of the 4 publishers there is a notable diversity between the third party hosts suspected to be trackers.

The STF-deviation metric introduced in this paper allows for the comparison of the differences between STFs of different publishers (inter-publisher comparison). The first results show a noticeable difference between the tracking behaviour of the different publishers, giving hope to idea that the publishers could be distinguished from one another and the STF and STF-deviation could give first hints/leads towards identifying a publisher by its tracking behaviour.

We have shown that the tracking behaviour of publishers can differ whether their papers are accessed using an interactive browser as compared to a headless browser. Although these are first results, this points towards interesting research topics to find the cause and mechanisms for detecting the browser type.

We could show that for the duration of our tests the tracking lists used to classify third party hosts as trackers did not change noticeably for the trackers employed by the publishers under examination. We still argue for maintaining the procedure keeping the possibility to check against updated lists of trackers.

Our results highlight the need for future work with regards to the examination environment. First results show there are differences between open access and non-open access papers for some publishers during our tests.

The inclusion of the publisher Wiley, albeit late in the research and lacking open access papers with the universities' OPAC gateway, bolstered our research and showed an intra publisher diversity and inter publisher differences within our tests .

The introduction of the STF-deviation metric allowed for the evaluation of the intra and inter publisher differences.

Future work should address the shortcomings of the STF-deviation metric:

- Not encompassing the total deviation of a STF
- Distortion of the STF-deviation, see Figures 8 and 9
- Limiting the value of the STF-deviation to a range of [0,1], to make it more interpretable
- Reducing the possibility of false positives and negatives

The source of the altered tracking behaviour of interactive vs. headless browsers should be identified and this behaviour mitigated. This would allow for a better quality of the results. The time span for observing changes in tracker lists for relevant tracker entries should be expanded to yield more insights into the relevance of a retrospective evaluation of tracking.

The setup and the software needs adaption to incorporate more sources for the comparison of open access vs. non-open access papers, e.g., the flexibility to add other publishers websites (some publishers have different sites for open and non-open access papers).

Also, some tools (e.g., Webbkoll) are barred from accessing some publisher websites, here mitigation to circumvent the restrictions or alternative tools could be a focus of future research.

Acknowledgements

The research from Stefan Kiltz, Robert Altschaffel and Jana Dittmann is partly funded by the EFRE project ”CyberSecurity-Verbund LSA II – Prävention, Detektion und Reaktion mit Open Source-Perspektiven” https://forschung-sachsen-anhalt.de/project/cybersecurity-verbund-lsa-praevention-27322.

References

[1] S. Kiltz, R. Altschaffel, and J. Dittmann, “Science-tracker fingerprinting with uncertainty: Selected common characteristics of publishers from network to application trackers on the example of web, app and email,” in *Proceedings of the Seventeenth International Conference on Emerging Security Information, Systems and Technologies (Securware)*, Porto, Portugal, 2023, pp. 88–97.

[2] Deutsche Forschungsgemeinschaft, “Data tracking in research: aggregation and use or sale of usage data by academic publishers,” (last access 2024.11.29). [Online]. Available: https://www.dfg.de/resource/blob/174924/d99b797724796bc1a137fe3d6858f326/datentracking-papier-en-data.pdf

[3] E. Bettinger, M. Bursic, and A. Chandler, “Disrupting the digital status quo: Why and how to staff for privacy in academic libraries,” (last access 2024.11.29). [Online]. Available: https://publish.illinois.edu/licensingprivacy/files/2023/06/Whitepaper-on-Privacy-Staffing-Licensing-Privacy.pdf

[4] M. Bambot, “How we hacked the sourcecon 2018 attendee list in 2 hours - by murtaza bambot - medium,” (last access 2024.11.29). [Online]. Available: https://medium.com/@MurtazaBambot/how-we-hacked-the-sourcecon-2018-attendee-list-in-2-hours-645bf26d2825

[5] R. Altschaffel, S. Kiltz, T. Lucke, and J. Dittmann, “Introduction to being a privacy detective: Investigating and comparing potential privacy violations in mobile apps using forensic methods,” in *Proceedings of the Fourteenth International Conference on Emerging Security Information, Systems and Technologies (Securware)*, Valencia, Spain, 2020, pp. 60–68.

[6] University Library of the Otto von Guericke University Magdeburg, “OPC4 - start/welcome,” (last access 2024.11.29). [Online]. Available: https://opac.lbs-magdeburg.gbv.de/DB=1/LNG=EN/

[7] S. Kiltz, "Data-centric examination approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics," Ph.D. dissertation, Otto-von-Guericke-University, Magdeburg, Germany, 2020, (last access 2024.11.29). [Online]. Available: https://opendata.uni-halle.de/bitstream/1981185920/34842/1/Kiltz_Stefan_Dissertation_2020.pdf

[8] W. Christl, "Corporate surveillance in everyday life," (last access 2024.11.29). [Online]. Available: https://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf

[9] H. Mildebrath, "Unpacking 'commercial surveillance': The state of tracking," (last access 2024.11.29). [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/739266/EPRS_BRI(2022)739266_EN.pdf

[10] N. Samarasinghe and M. Mannan, "Towards a global perspective on web tracking," *Computers & Security*, vol. 87, p. 101569, 2019, (last access 2024.11.29). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404818314007

[11] K. Sim, H. Heo, and H. Cho, "Combating web tracking: Analyzing web tracking technologies for user privacy," *Future Internet*, vol. 16, no. 10, 2024. [Online]. Available: https://www.mdpi.com/1999-5903/16/10/363

[12] R. Pan and A. Ruiz-Martínez, "Evolution of web tracking protection in chrome," *Journal of Information Security and Applications*, vol. 79, p. 103643, 2023, (last access 2024.11.29). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212623002272

[13] B. Krupp, J. Hadden, and M. Matthews, "An analysis of web tracking domains in mobile applications," in *Proceedings of the 13th ACM Web Science Conference 2021*, ser. WebSci '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 291–298, (last access 2024.11.29). [Online]. Available: https://doi.org/10.1145/3447535.3462507

[14] R. Altschaffel, M. Beurskens, J. Dittmann, W. Horstmann, S. Kiltz, G. Lauer, J. Ludwig, B. Mittermaier, and K. Stump, "Data tracking and DEAL: On the 2022/2023 negotiations and the consequences for academic libraries," *Recht und Zugang (RuZ)*, vol. 5, no. 1, Nov. 2024, (last access 2024.11.29). [Online]. Available: https://doi.org/10.5281/zenodo.14006196

[15] C. Hanson, "User tracking on academic publisher platforms," (last access 2024.11.29). [Online]. Available: https://www.codyh.com/writing/tracking.html

[16] European Commission and Directorate-General for Communications Networks, Content and Technology, C. Armitage, N. Botton, L. Dejeu-Castang, and L. Lemoine, *Study on the impact of recent developments in digital advertising on privacy, publishers and advertisers – Final report*. Publications Office of the European Union, 2023.

[17] Volatility Foundation, "Github - volatilityfoundation/volatility3: Volatility 3.0 development," (last access 2024.11.29). [Online]. Available: https://github.com/volatilityfoundation/volatility3?tab=readme-ov-file

[18] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of ACM CCS 2016*, 2016, pp. 1388–1401.

[19] Dataskydd.net Sverige, "Analyze — webbkoll - dataskydd.net," (last access 2024.11.29). [Online]. Available: https://webbkoll.dataskydd.net/en

[20] Wireshark Foundation, "Wireshark - go deep," (last access 2024.11.29). [Online]. Available: https://www.wireshark.org/

[21] European Data Protection Supervisor, "European data protection supervisor / website-evidence-collector · gitlab," (last access 2024.11.29). [Online]. Available: https://code.europa.eu/EDPS/website-evidence-collector/

[22] ungoogled-software, "Release 121.0.6167.184-1 · ungoogled-software/ungoogled-chromium · github," (last access 2024.11.29). [Online]. Available: https://github.com/ungoogled-software/ungoogled-chromium/releases/tag/121.0.6167.184-1

[23] Mathias Bynens, "Github - puppeteer/puppeteer: Javascript api for chrome and firefox," (last access 2024.11.29). [Online]. Available: https://github.com/puppeteer/puppeteer

[24] Disconnect Inc., "Github - disconnectme/disconnect-tracking-protection: Canonical repository for the disconnect services file," (last access 2024.11.29). [Online]. Available: https://github.com/disconnectme/disconnect-tracking-protection

[25] Fanboy, MonztA, Khrin, Yuki2718, and piquark6046, "Github - easylist/easylist: Easylist filter subscription (easylist, easyprivacy, easylist cookie, fanboy's social/annoyances/notifications blocking list):," (last access 2024.11.29). [Online]. Available: https://github.com/easylist/easylist

[26] E. Casey, "Error, uncertainty and loss in digital evidence," *International Journal of Digital Evidence*, vol. 1, no. 2, pp. 1–45, 2002.

[27] K. Inman and N. Rudin, *Principles and Practises of Criminalistics: The Profession of Forensic Science*. Boca Raton Florida, USA: CRC Press LLC, 2001.

[28] Dataskydd.net Sverige, "dataskydd.net/webbkoll-backend: Express.js app that runs puppeteer as a service; visits specified url with chromium and sends back various data (requests, cookies, etc.) as json. - codeberg.org," (last access 22/11/2024). [Online]. Available: https://codeberg.org/dataskydd.net/webbkoll-backend

[29] Oracle Inc., "Oracle VM virtualbox," (last access 2024.11.29). [Online]. Available: https://www.virtualbox.org

[30] ungoogled-chromium Authors, "GitHub - ungoogled-software/ungoogled-chromium: Google Chromium, sans integration with Google," (last access 2024.11.29). [Online]. Available: https://github.com/ungoogled-software/ungoogled-chromium

[31] S. Kiltz, N. Weiler, T.-F. Riechard, R. Altschaffel, and J. Dittmann, "Securware-journal-download-folder," (last access 2024.11.29). [Online]. Available: https://cloud.ovgu.de/s/RWEHi9wSqH3xbjQ

[32] Daniel Stenberg, "curl - command line tool and library for transferring data with urls (since 1998)," (last access 2024.11.29). [Online]. Available: https://curl.se/

[33] Philippe Biondi, "Scapy," (last access 2024.11.29). [Online]. Available: https://scapy.net/