Capturing Trends Using OpenAlex and Wikipedia Page Views as Science Indicators: The Case of Data Protection and Encryption Technologies

Sarah Ismail[*], Alain Mermoud[**], Loïc Maréchal[***], Samuel Orso[****] and Dimitri Percia David[*****]

[*]*sarah.ismail.ar@gmail.com*
0009-0002-1397-435X
Cyber-Defence Campus, armasuisse Science and Technology, EPFL Innovation Park, Lausanne, CH-1015, Switzerland

[**]*alain.mermoud@ar.admin.ch*
0000-0001-6471-772X
Cyber-Defence Campus, armasuisse Science and Technology, EPFL Innovation Park, Lausanne, CH-1015, Switzerland

[***] *loic.marechal@unil.ch*
0000-0001-8039-5097
Faculty of Business and Economics (HEC Lausanne), University of Lausanne, Bâtiment Internef, Lausanne, CH-1015, Switzerland

[****] *samuel.orso@unil.ch*
0000-0002-2064-3586
Geneva School of Economics and Management, University of Geneva, Geneva, CH-1211, Switzerland

[*****] *dimitri.perciadavid@hevs.ch*
0000-0002-9393-1490
Institute of Entrepreneurship & Management (HES-SO), University of Applied Sciences of Western Switzerland, Sierre, CH-3960, Switzerland

1. Introduction

Technological change is often measured using indicators such as patent analysis (Chen et al., 2017), research, and bibliometric analyses of scientific publications. For instance, Zhang et al. (2020) design a bibliometric framework combining bibliometric methods and a novel approach to charting the evolutionary pathways of scientific innovation. Meanwhile, Percia David et al. (2023) establish a scientometric analysis on arXiv e-prints to study the different patterns of security development in computer technologies through various factors such as the attention paid to security among technologies as well as the effect of opinion on security development. They study 20 categories of computer technologies on arXiv and find that the category Cryptography and Security has the highest share of security attention, exceeding 75%.

An emerging approach to measuring change is using public attention as a metric. Indeed, public attention arouses great curiosity among researchers in all fields. Several studies have been conducted to capture the public's attention. In the environmental domain, Guedes-Santos et al. (2021) use Wikipedia page views to measure public interest to study the popularity of protected areas. In 2021, Wang et al. find that public attention could improve wastewater treatment, and therefore, the public could be a very effective supervisor of environmental issues. On a societal level, Alis et al. (2015) find that studying the views of Wikipedia articles opens up the possibility

of improving estimates of the current number of tourists leaving the UK. On the biodiversity side, Roll et al. (2016) measure public interest in reptiles by analyzing Wikipedia page views, providing insight into the cultural importance of reptiles. While this approach is applied in different fields, it has not yet been explored in the technological field.

Thus, we propose a metric to quantify public attention in encryption and data protection technologies based on Wikipedia page views. We use page views as a proxy of public attention. We analyze which technologies are likely to gain popularity and examine differences in interest across the technology life cycles. We use a webometrics approach to treat the number of views on Wikipedia for 36 technologies, previously identified by field experts using the Delphi method. Additionally, we use OpenAlex data (Priem et al., 2022) to compare expert and public attention to the 36 technologies studied. By comparing expert and public attention, we gain a better understanding of the technology landscape and identify areas that may need more attention. This analysis provides information on potential future developments in data protection and encryption technologies. We identify emerging technologies and predict future consumer behavior. Ultimately, this study sheds light on the benefits of using data reflecting the public interest to assess the temporal position of a technology.

2. Data and methods

Wikipedia pageview statistics permit downloading the number of page visits over a given period at the chosen frequency. It provides daily, monthly, and yearly data. The statistics do not consider the time spent by Internet users on a page; irrespective of its duration, it is counted as a view. To obtain an accurate reading of the statistics of the pages, we enter the name of each of our 36 technologies in the search box of the site {https://pageviews.wmcloud.org/} to get to the main page of the technology, while paying attention to redirects and page shortcuts that can lead us to a secondary page. Next, we filter the data by specifying that we want all data available in monthly frequency on all types of platforms and agents: desktops, mobile applications, and mobile web. We download the data for the 36 technologies over 82 months, from July 2015 to April 2022. Therefore, this dataset serves as a measure of global popularity and is a proxy for public attention to encryption and data protection technologies.

The OpenAlex dataset describes learned entities and how these entities connect to each other using a graph structure. Each scholarly article or scientific publication has associated concepts represented in the document. OpenAlex organizes publications in a tree structure, where the general concepts are the parents of finer concepts. In total, OpenAlex has 65,026 concepts ranging from political science to physics. Posts are tagged automatically using a classification model trained on Microsoft Academic Graph (MAG). Thus, OpenAlex provides a taxonomy of topics covered in the scientific literature (Scheidsteger et al., 2020), which we use in the following to retrieve articles tagged to encryption and data protection technologies. We scrape the papers of each technology, and we count the number of papers published monthly. This gives a time series for each technology. To balance the frequency and the quantity of data, we use the monthly frequency for all data sources.

3. Results

The evolution of public attention over time for the 36 technologies

The level of public attention given to encryption and data protection technologies considerably varies over time and is heavily influenced by technological advancements. Among the 36 technologies we examine, there are significant differences in popularity and maturity. For instance, "Blockchain" emerges as one of the most widely used encryption technologies, possibly due to a range of factors such as external events, public opinion, or the popularity of the technology itself. Comparing these technologies can be challenging given their differences. Yet, to gain a better understanding of public attention, we plot the monthly page views of each technology.

Figure 1: Multi-plot of public attention from Wikipedia page views on 36 encryption and data protection technologies. The observations are monthly, range from July 2015 to April 2022 and the frequency is monthly.

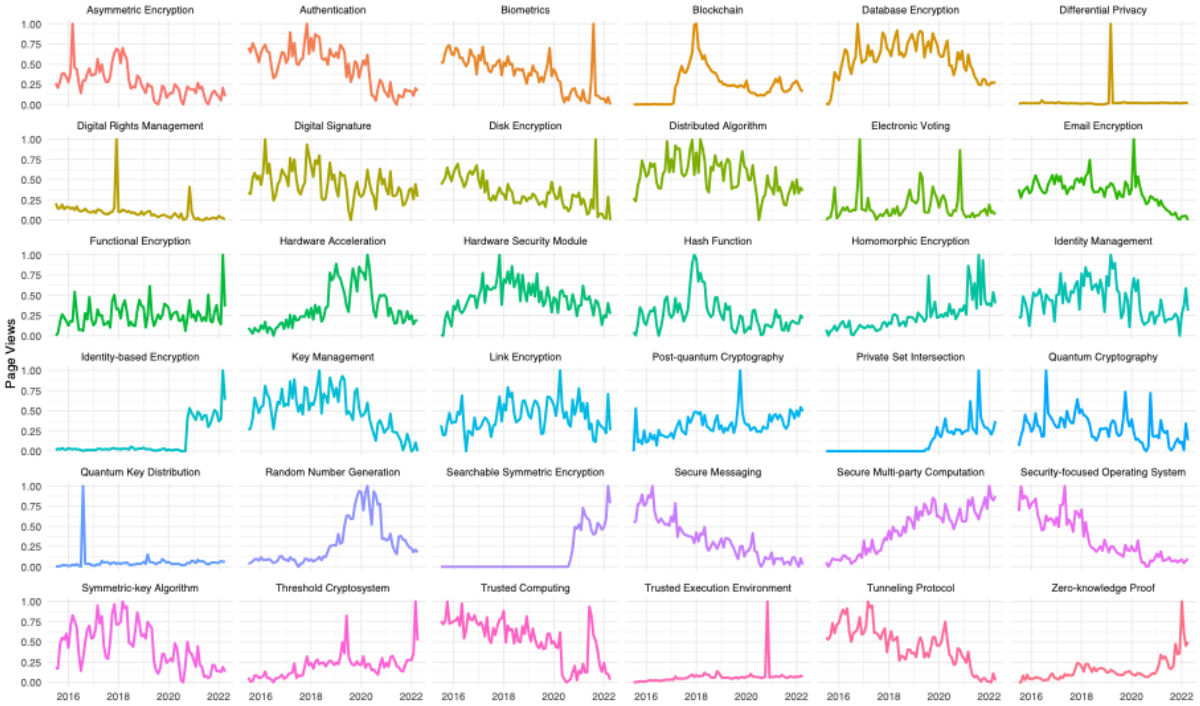
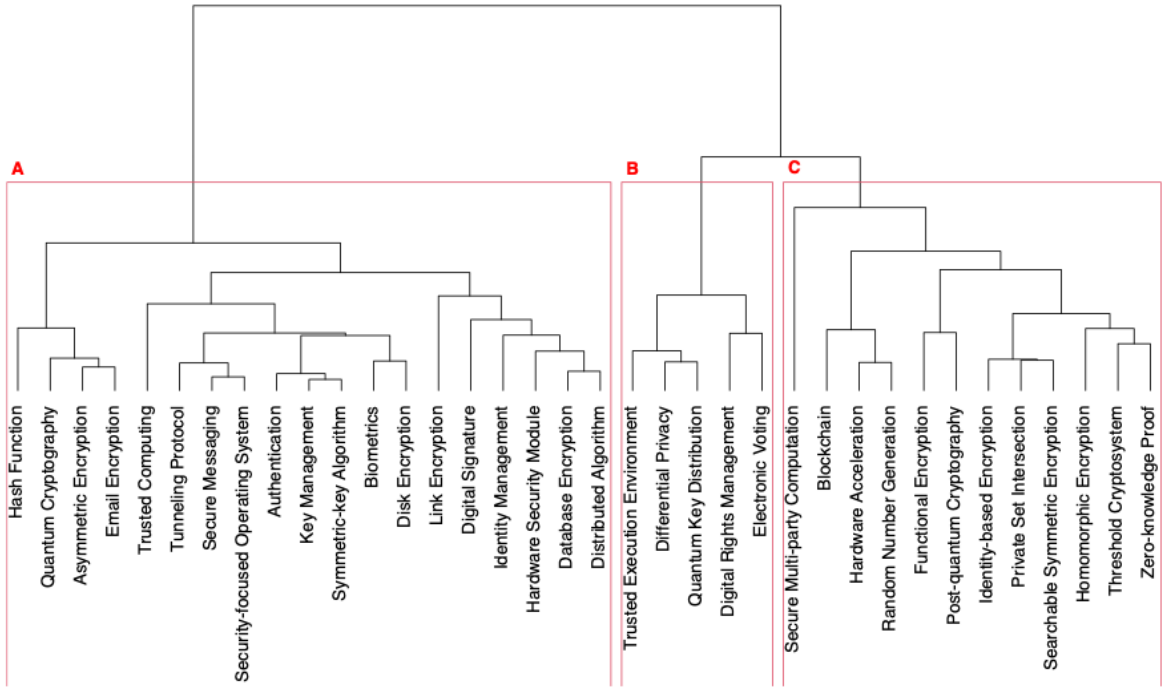


Figure 1 highlights the variability of public interest in encryption and data protection technologies, dependent on both time and specific technology. Some technologies, such as "Secure multi-party computing" and "Homomorphic Encryption", exhibit a consistent and steady rise in public attention. Others, such as "Secure messaging", experience a decline in public attention. Additionally, certain technologies, including "Hardware security module" and "Database encryption", exhibit a long-term concave curve, with a peak of interest followed by a gradual decrease. The observed long-term concave curve is characterized by an initial increase in public attention, which reaches a peak after a certain number of views, followed by a gradual decrease in interest. This phenomenon can be explained by the fact that the public tends to be highly curious about new technology during its inception, resulting in a surge of interest and heightened web traffic. However, as technology becomes more widely understood and integrated into daily life, the novelty factor fades, resulting in a decline in public attention. This trend is commonly observed in the maturity stage of the technological life cycle, wherein the technology reaches widespread adoption and has become more commonplace, leading to a decline in the public interest.

Type of trends

We use a clustering algorithm to group technologies based on their similarities to provide further evidence of these trends. This algorithm aims to identify patterns in the data that are not immediately obvious from the graphical analysis. Specifically, we use the Dynamic Time Warping method to measure the similarity between two sequences, regardless of their temporal alignment. This allows us to identify similarities and differences between technologies that are not apparent from a simple comparison of their graphs. The clustering analysis produces three main clusters highly consistent with the types of curves we observe in our graphical analysis: those with increasing curves, decreasing curves, and no clear trends. We present the results of this clustering analysis in a dendrogram, in Figure 2. The dendrogram illustrates the relationships between the different technologies based on their similarities regarding public attention. Technologies that are closely related are grouped together, while those that are less similar are located farther apart. By examining the dendrogram, we gain insights into how different technologies are associated.

Figure 2: Dendrogram showing the selected 36 technologies grouping technologies by similarities: (A) negative trends, (B) constant trends, and (C) positive trends. We use the Dynamic Time Warping (DTW) method to estimate the similarity between two temporal sequences that do not align exactly in time, speed, or length. The observations range from July 2015 to April 2022 and are normalized and decomposed to filter out the noise and potential seasonality and keep only the trend.



The first cluster (A) includes 19 technologies with a negative trend, indicating a decline in public attention over time. This cluster includes well-established technologies such as Hash function, Email encryption, and Database encryption, which have been in use for a while and are no longer considered novel. The second cluster (B) consists of five technologies with a stable trend in public attention, indicating that they maintain a consistent level of interest over time. These technologies

include Electronic Voting and may be less controversial or exciting to the public. Finally, the third cluster (C) contains 12 technologies with a positive trend in public attention. These technologies are currently booming and well-known. Examples of technologies in this cluster include Blockchain, Homomorphic encryption, and Zero-knowledge proof. These technologies are driving innovation in the field and attracting significant interest from the public.

Different patterns of development

Different patterns of development are observed in technological time series, and based on these patterns, we categorize them into three classes: no growth, moderate growth, and high growth. To calculate the clusters, we divide the average page views of the last three months by the average page views of the last three months from six years ago. We refer to the resulting ratio as growth ratio. If the growth ratio < 1.05, the technology is not growing. The technology exhibits moderate growth if 1.05 < growth\ ratio < 2 and finally, the technology thrives if 2 < growth ratio. Furthermore, we also classify technologies into low, moderate, and high-interest categories, with a technology considered high-interest if it receives an average of c ≥ 50,000 or more page views per month, moderate-interest if it receives between 25, 000 ≤ c < 50,000 page views per month, and low-interest if it receives less than c < 25,000 page views per month. The clustering of technologies based on their growth is presented in Table 1 as a two-dimensional matrix.

Table 1. Technology pageviews are associated with a two-dimensional matrix created by grouping technologies based on their past growth and public interest.

		Interest	
Growth Pattern	Low Interest	Moderate Interest	High Interest
No Growth	Authentication, Trusted Computing, Disk Encryption, Electronic Voting, Email Encryption, Hardware Security Module, Identity Management, Key Management, Quantum Cryptography, Secure Messaging, Security-focused Operating System, Symmetric-key Algorithm, Tunneling Protocol	Biometrics, Digital Rights Management, Digital Signature	Asymmetric Encryption, Hash Function
Moderate Growth	Differential Privacy, Functional Encryption, Hardware Acceleration, Homomorphic Encryption, Quantum Key Distribution	Random Number Generation	
Strong Growth	Identity-based Encryption, Secure Multi-party Computation, Post-quantum Cryptography, Private Set Intersection, Searchable Symmetric Encryption, Secure Multi-Party Computation, Trusted Execution Environment, Zero-knowledge Proof		Blockchain

The technologies attracting significant public interest are Blockchain, Hash Function, and Asymmetric Encryption. Blockchain shows strong growth, unlike Hash Function, which shows no growth. Again, technologies with more specialized techniques and methodologies, such as Digital Signature and Biometrics, are seeing moderate interest. Low-interest and no-growth technologies are niche technologies, such as Disk Encryption, or long-standing technologies, such as Email Encryption. However, some low-interest technologies, such as Post-quantum Cryptography, still show strong growth.

Relationship between technologies

To better understand the relationships between the 36 technologies, we calculate the correlation between their page view time series. To ensure that we capture the correlation between the actual specificity of the technologies, we filter the time series data from the trend and seasonality, leaving us with the noise component of the time series. This approach allowed us to remove any potential confounding factors and focus solely on the correlation between the technologies.

Figure 3: This figure illustrates the full correlation study between Wikipedia pageviews of the 36 technology pages. The observations range from July 2015 to April 2022 and are normalized, decomposed, and filtered to keep only the noise.

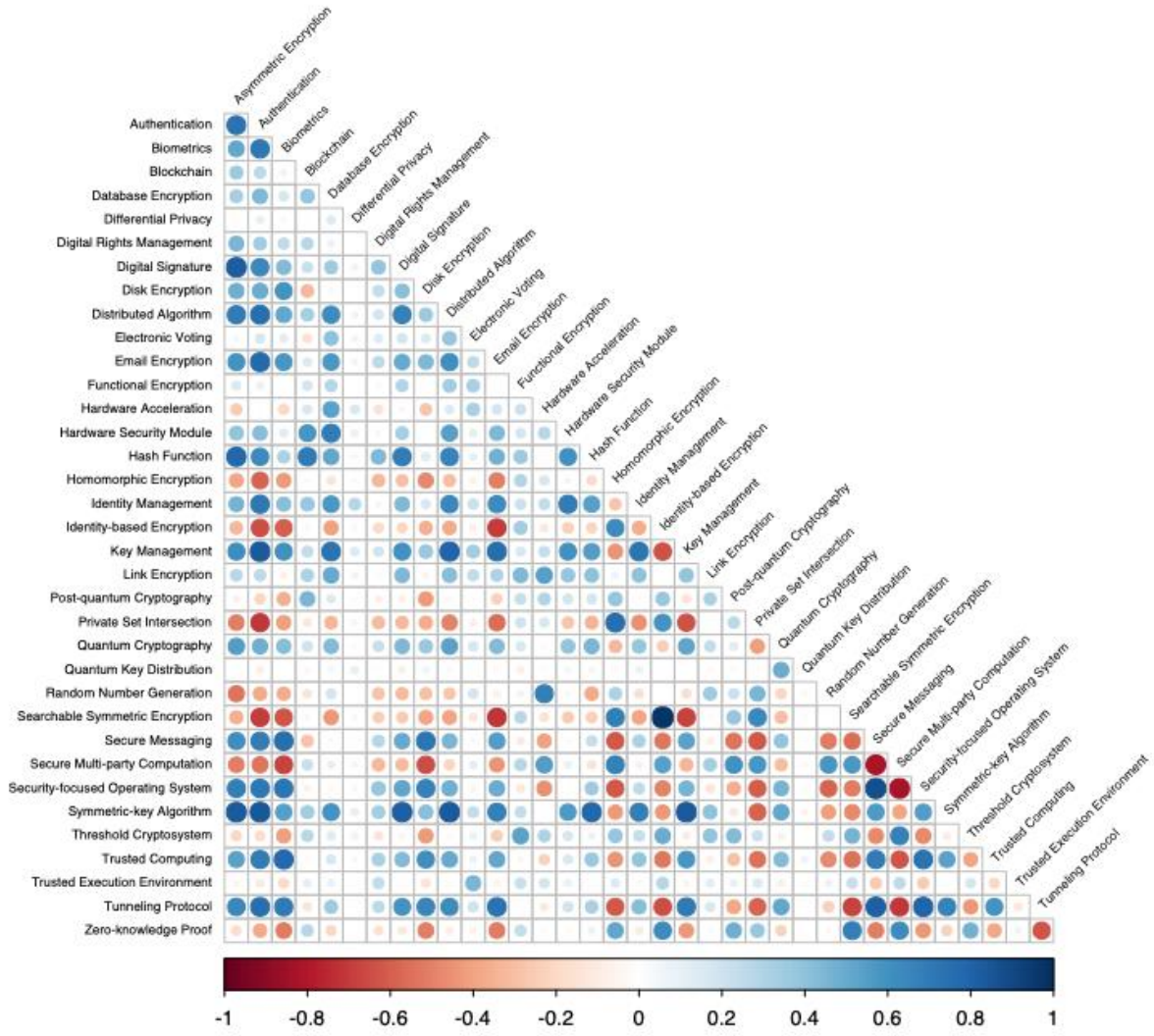


Figure 3 displays the correlation matrix of Wikipedia's monthly page views. One of the highest correlation coefficients is 0.71, and it exists between two related technologies: Quantum cryptography and Quantum key distribution. Given their fields, it is not surprising that these two technologies are closely related. We assume that users have a strong tendency to click on the hyperlink between these two Wikipedia pages, so when they visit one page, they would likely often click on the other. The correlation between Biometrics and Zero-knowledge proof is -0.53, indicating that the public attention on these two technologies is countercyclical. While both technologies allow identification, Biometrics determines a person's identity, while Zero-knowledge proof enables users to demonstrate the veracity of a situation without revealing any information about it. One interpretation of this negative correlation is that Biometrics is a long-standing technology that is already well-established (part of cluster (A) in Figure 2, while Zero-knowledge proof is relatively new and topical (part of cluster (C) in Figure 2). Hence, individuals interested in Zero-knowledge proof may be more likely to focus on more current and growing technologies. These correlations can potentially contain confounding factors and spurious relationships.

Comparison of public and expert attention

We use OpenAlex time series as a proxy for expert attention, as it reflects the number of publications related to each technology in the scientific community. By comparing the OpenAlex time series (in red color) with the Wikipedia page view time series (in blue color), which represents public attention, we can gain insights into the relationship between public attention and the expert attention in the field of encryption and data protection technologies. Overall, analyzing both public attention and expert attention can provide a more comprehensive understanding of trends and developments in the field.

Figure 4: These plots present public attention from Wikipedia page views (blue line) and expert attention from the number of publications on OpenAlex (red line). We discard outliers, and the study period is from July 2015 to April 2022. Data is provided on a monthly frequency. The method used on the data is the robust z-score with a scale from -5 to 5.

The relationship between the two types of attention proxies shows an interesting pattern over time. Graphically, we observe that expert attention tends to follow public attention by a few months. This time lag between the two types of attention proxies may be due to various factors, such as the time it takes for researchers to conduct, write, and publish in scientific journals. Additionally, public attention may serve as a precursor or indicator of emerging research trends that experts pick up. Understanding the dynamics of the relationship between public and expert attention provides insights into how trends develop and how to effectively communicate emerging technologies to the public.

4. Conclusion

In conclusion, this paper introduces a novel webometric methodology based on open science practices. We present a framework that focuses on public attention and expert attention to technologies to predict trends in the field and monitor the technological life cycle. Our study offers a novel perspective on the position of technologies over time and their classification. The analysis reveals distinct trends for 36 technologies. Notably, Blockchain, Hash Function, and Asymmetric Encryption are the technologies with the largest public interest. In contrast, Disk Encryption and Email Encryption have the lowest interest with no growth. We also find that Post-quantum Cryptography, although a low-interest technology, still shows strong growth. These findings give support to the validity of our framework for identifying, analyzing, and predicting technology-related trends. Our results indicate that monitoring public attention on Wikipedia can be a key indicator and offer essential insights into technology trends.

Several limitations should be considered when interpreting our results. First, our sample is limited to seven years, which may make it difficult to accurately determine the position of a technology within its technological life cycle. However, we uncover that the inception year of a technology provides valuable information for predicting the public attention curve. Specifically, we assume that a brand-new technology will initially have an increasing curve. Second, our analysis is limited by the number of external datasets used to explain page views variations. Indeed, we work with only two external databases, which may limit the quality of our forecasts. Additionally, page views are used as a proxy of public attention, which does not reflect a technology's true level of public attention. Furthermore, it is also essential to note that our analysis only concerns the English version of Wikipedia. Finally, marketing campaigns or newswires may affect web traffic data, which could be mitigated by combining the data to provide a complete picture of technology trends.

In future research, there is potential to further advance our approach by incorporating additional external datasets. Collecting more data, such as each technology's Twitter hashtags, can provide more explanatory variables to our models, increasing their explanatory power. Finally, measuring and analyzing public consensus, or controversy on a topic, through the number of edits made to Wikipedia pages could be an innovative way to understand public perception of different technologies over time. These avenues of research could potentially improve our understanding of technological change and thus provide valuable insights.

Open science practices

Our research is using publicly available data sources, such as Wikipedia pageviews and the Openalex dataset. This allows transparency, reproducibility, and accessibility of our results. We have not shared a research plan in advance but have documented our analysis in a clear and reproducible manner and are willing to share our code with anyone interested in reproducing or expanding our work. Our commitment to open science practices ensures that our research is accessible to a broader audience, and we hope this brief reflection highlights the importance of transparency in research.

Competing interests

There are no competing interests.

References

Alis, C. M., Letchford, A., Moat, H. S., & Preis, T. (2015). Estimating tourism statistics with Wikipedia page views. In *Proceedings of the ACM Web Science Conference*, 1-2.

Chen, H., Zhang, G., Zhu, D., & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, *119*, 39-52.

David, D. P., Maréchal, L., Lacube, W., Gillard, S., Tsesmelis, M., Maillart, T., & Mermoud, A. (2023). Measuring security development in information technologies: A scientometric framework using arXiv e-prints. *Technological Forecasting and Social Change*, *188*, 122316.

Priem, J., Piwowar, H.,, & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada, Spain.*

Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., & Grenyer, R. (2016). Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation*, *204*, 42-50.

Scheidsteger, T., & Haunschild, R., (2022). Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020. *26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada, Spain.*

Wang, C., Chu, Z., & Gu, W. (2021). Assessing the role of public attention in China's wastewater treatment: A spatial perspective. *Technological Forecasting and Social Change*, *171*, 120984.

Zhang, Y., Porter, A. L., Cunningham, S., Chiavetta, D., & Newman, N. (2020). Parallel or intersecting lines? Intelligent bibliometrics for investigating the involvement of data science in policy analysis. *IEEE Transactions on Engineering Management*, *68*(5), 1259-1271.