Hochschule **RheinMain**

GROUNDING ON SHAKY GROUND:

Wikipedia's Legal Articles, Editorial Integrity, and the Risk of Data Poisoning in Artificial Intelligence

Prof. Dr. Matthias Harter

July 2025

A (VERY) SHORT RÉSUMÉ

Contact information at the end of this presentation...

SHORT RÉSUMÉ

Some call it CV...

Hochschule RheinMain

Name	Prof. Dr. Matthias Harter
	Fields of interest / profession
	• Patents and IP
	• AI, AGI and humanity
	• ASICs, Circuits and Systems
	• Aviation, Simulators
since 07/11	Professor for Embedded Systems and Microcomputers Hochschule RheinMain University of Applied Sciences
10/12 – 09/18	Head of the Department of Electrical Engineering and Information Technology
10/17 – 10/23	Head (founder) of new study program „Electrical and Aviation Engineering“

STARTING POINT:
RAG TO FIGHT HALLUCINATION IN LLMS

Retrieval Augmented Generation (RAG) as Countermeasure to Hallucination

HALLUCINATON VS. CONFABULATION

Humans also hallucinate / confabulate

Hochschule **RheinMain**

Hallucination / Confabulation in LLMs:
Filling gaps in memory / knowledge / training

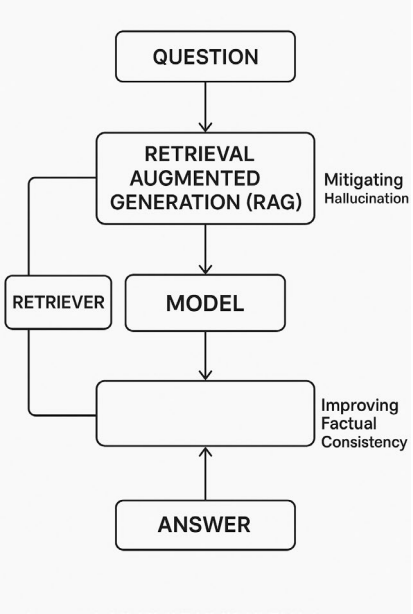
LEGAL DOMAIN IN GERMAN WIKIPEDIA

RAG as countermeasure against hallucination

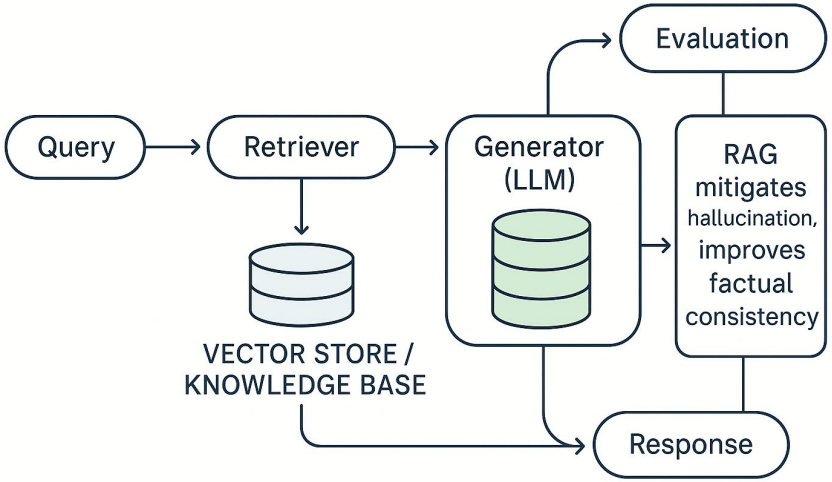
Hochschule **RheinMain**

- **Prompt for Text-to-Image Generative AI:**
  "Schematic drawing of an AI system which uses Retrieval Augmented Generation (RAG) as countermeasure against hallucination of the model to improve factual consistency."

ChatGPT o4



ChatGPT o3



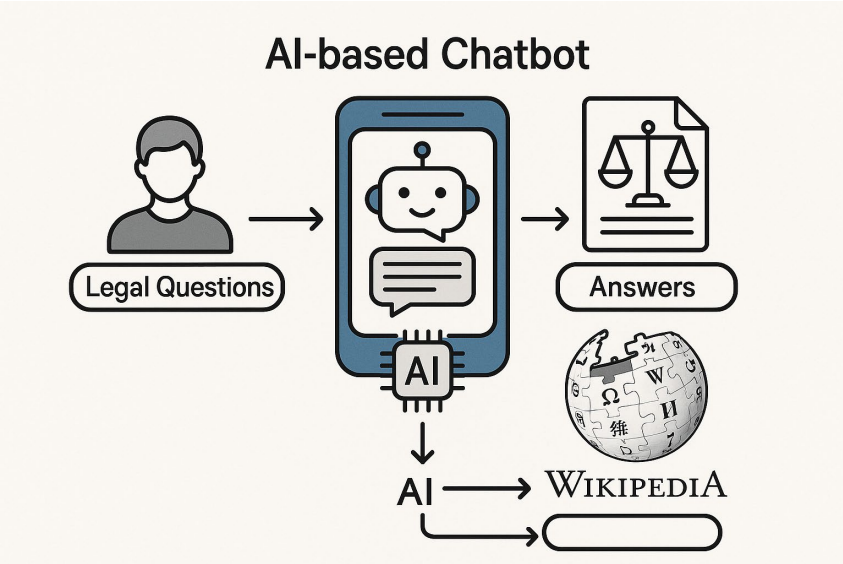
LEGAL DOMAIN IN GERMAN WIKIPEDIA

Idea of a RAG based AI chatbot for non-experts

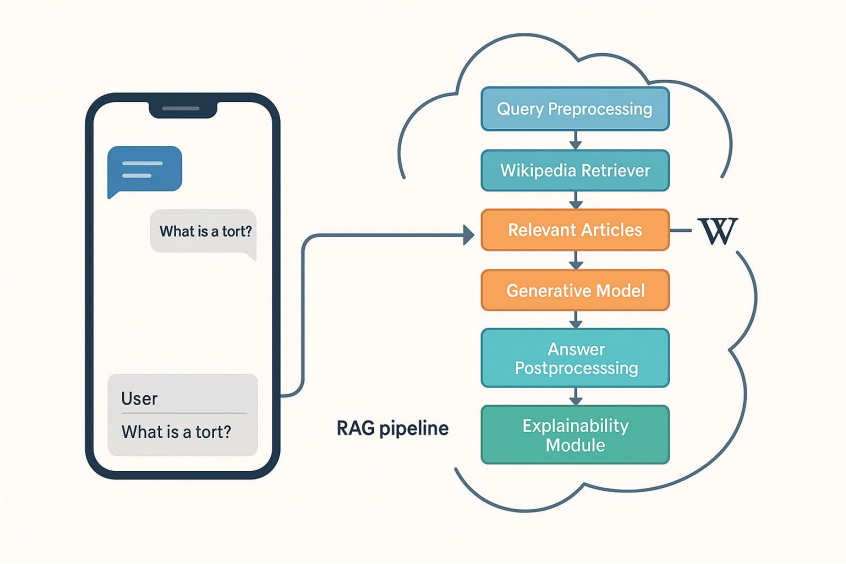
Hochschule **RheinMain**

- **Prompt for Text-to-Image Generative AI:**
  "Schematic representation of an AI-based chatbot in the form of an app that answers legal questions for laypersons and bases its answers solely on articles from Wikipedia according to the RAG approach."

ChatGPT 4o



ChatGPT o3



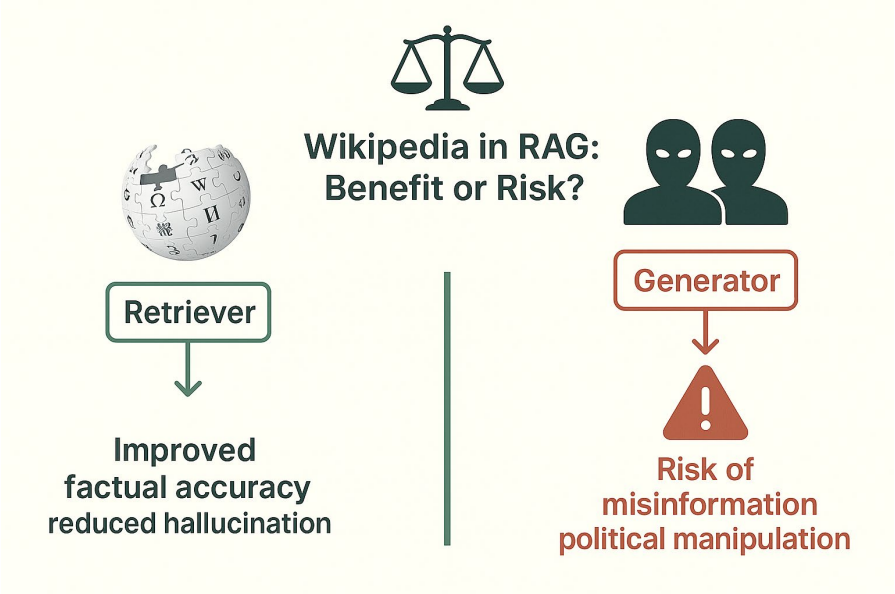
LEGAL DOMAIN IN GERMAN WIKIPEDIA

Idea of a RAG based AI chatbot for non-experts

Hochschule **RheinMain**

Research Question:

"Does relying on articles from Wikipedia in a RAG-based AI system improve the factual accuracy of answers and mitigate hallucation or do these articles have an inherent problem concerning infiltration by malicious authors and even politically motivated manipulations?"



METHODOLOGY

The process to download all legal domain articles from the German Wikipedia

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API

Hochschule **RheinMain**

Maintenance-Category Retrieval

Maintenance categories used by Wikipedia to tag outdated or problematic pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

Download of Legal Subcategories

Starting from the top-level category „Recht“ („law“) in the German Wikipedia, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

Template-Based Retrieval

Identification of all articles that utilize one of all existing law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics.

Keyword-based Title Search

Using a list of legal terms from a specialized law dictionary [18], a title-search is performed.

Expansion via Internal Links

From all articles in the database, the most frequently occurring interal Wikipedia links are extracted and the same category-based filtering is applied.

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API

Hochschule **RheinMain**

Maintenance-Category Retrieval

Maintenance categories used by Wikipedia to tag outdated or problematic pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

Download of Legal Subcategories

Starting from the top-level category „Recht“ („law“) in the German Wikipedia, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

Template-Based Retrieval

Identification of all articles that utilize one of all existing law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics.

Keyword-based Title Search

Using a list of legal terms from a specialized law dictionary [18], a title-search is performed.

Expansion via Internal Links

From all articles in the database, the most frequently occurring interal Wikipedia links are extracted and the same category-based filtering is applied.

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API

Hochschule **RheinMain**

Maintenance-Category Retrieval

Maintenance categories used by Wikipedia to tag outdated or problematic pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

Download of Legal Subcategories

Starting from the top-level category „Recht“ („law“) in the German Wikipedia, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

Template-Based Retrieval

Identification of all articles that utilize one of all existing law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics.

Keyword-based Title Search

Using a list of legal terms from a specialized law dictionary [18], a title-search is performed.

Expansion via Internal Links

From all articles in the database, the most frequently occurring interal Wikipedia links are extracted and the same category-based filtering is applied.

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API

Hochschule **RheinMain**

Maintenance-Category Retrieval

Maintenance categories used by Wikipedia to tag outdated or problematic pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

Download of Legal Subcategories

Starting from the top-level category „Recht“ („law“) in the German Wikipedia, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

Template-Based Retrieval

Identification of all articles that utilize one of all existing law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics.

Keyword-based Title Search

Using a list of legal terms from a specialized law dictionary [18], a title-search is performed.

Expansion via Internal Links

From all articles in the database, the most frequently occurring interal Wikipedia links are extracted and the same category-based filtering is applied.

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API

Hochschule **RheinMain**

Step 1

Maintenance-Category Retrieval

Maintenance categories used by Wikipedia to tag outdated or problematic pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

Step 2

Download of Legal Subcategories

Starting from the top-level category „Recht“ („law“) in the German Wikipedia, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

Step 3

Template-Based Retrieval

Identification of all articles that utilize one of all existing law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics.

Step 4

Keyword-based Title Search

Using a list of legal terms from a specialized law dictionary [18], a title-search is performed.

Step 5

Expansion via Internal Links

From all articles in the database, the most frequently occurring interal Wikipedia links are extracted and the same category-based filtering is applied.

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API: Article Count

Hochschule RheinMain

Step 1 | Step 2 | Step 3 | Step 4 | Step 5

Maintenance-Category Retrieval

Maintenance categories used by W… t… p… pages are downloaded. These serve as one filtering criterion to exclude articles from subsequent steps if they are deemed insufficiently maintained or not in compliance with editorial standards.

175 Maintenance Categories

Download of Legal Subcategories

Starting from the top-level cat… t“ (… W…, all subcategories are recursively traversed, collecting any articles placed under these nested categories.

~15,000 Legal Subcategories

Template-Based Retrieval

Identification of all artic… e o… sp… templates (e.g., infoboxes or structured references) design… to pro… ut for…

24 Law-Specific Templates

~30,000 Articles returned

Keyword-based Title Search

Using a li… law dictionary [18], a title-search is performed…

~14,000 Legal Terms from Dictionary

~9,000 Articles returned

Expansion via Internal Links

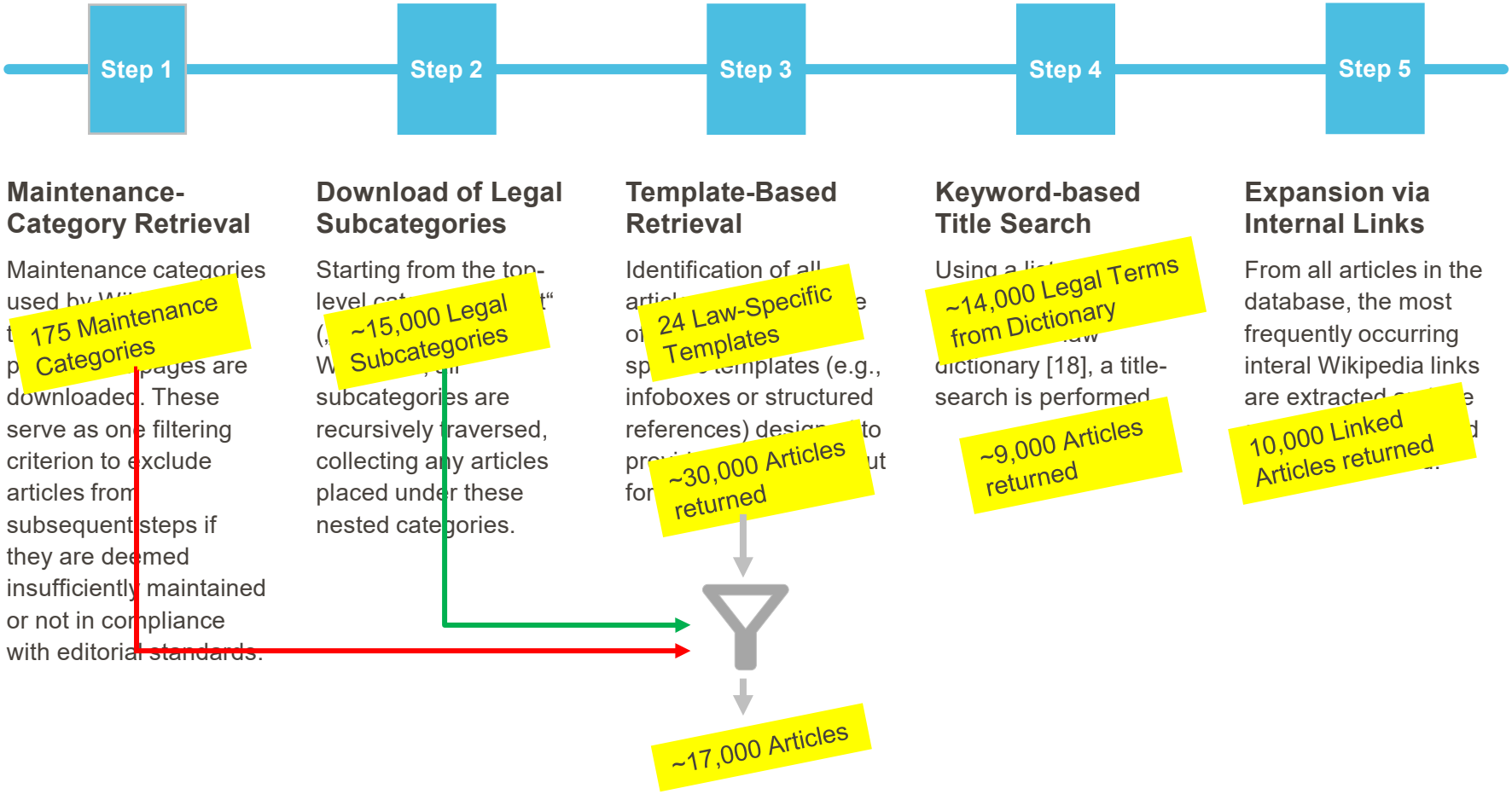
From all articles in the database, the most frequently occurring interal Wikipedia links are extracted … e … d …

10,000 Linked Articles returned

LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API: Filtering Process

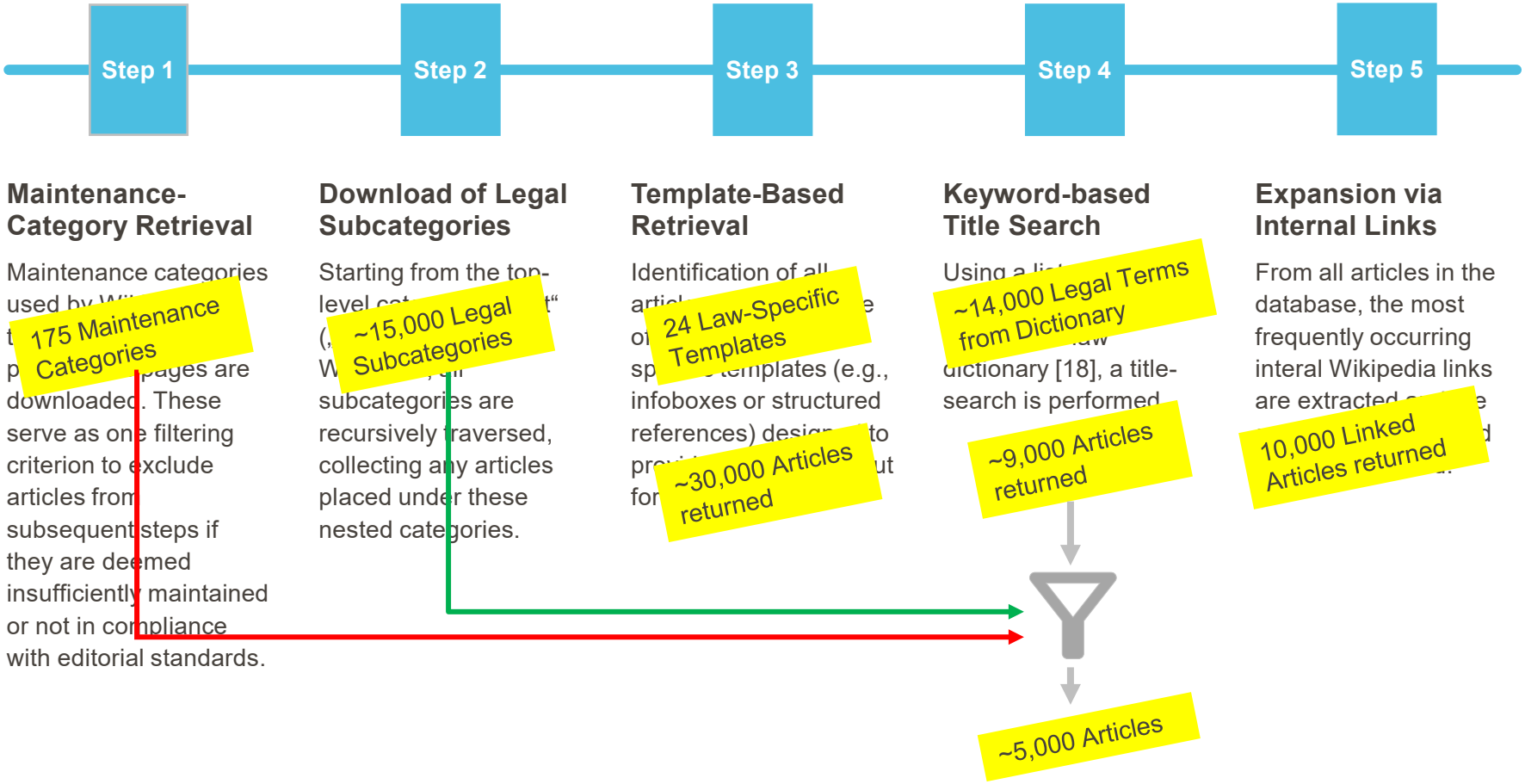
Hochschule **RheinMain**



LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API: Filtering Process

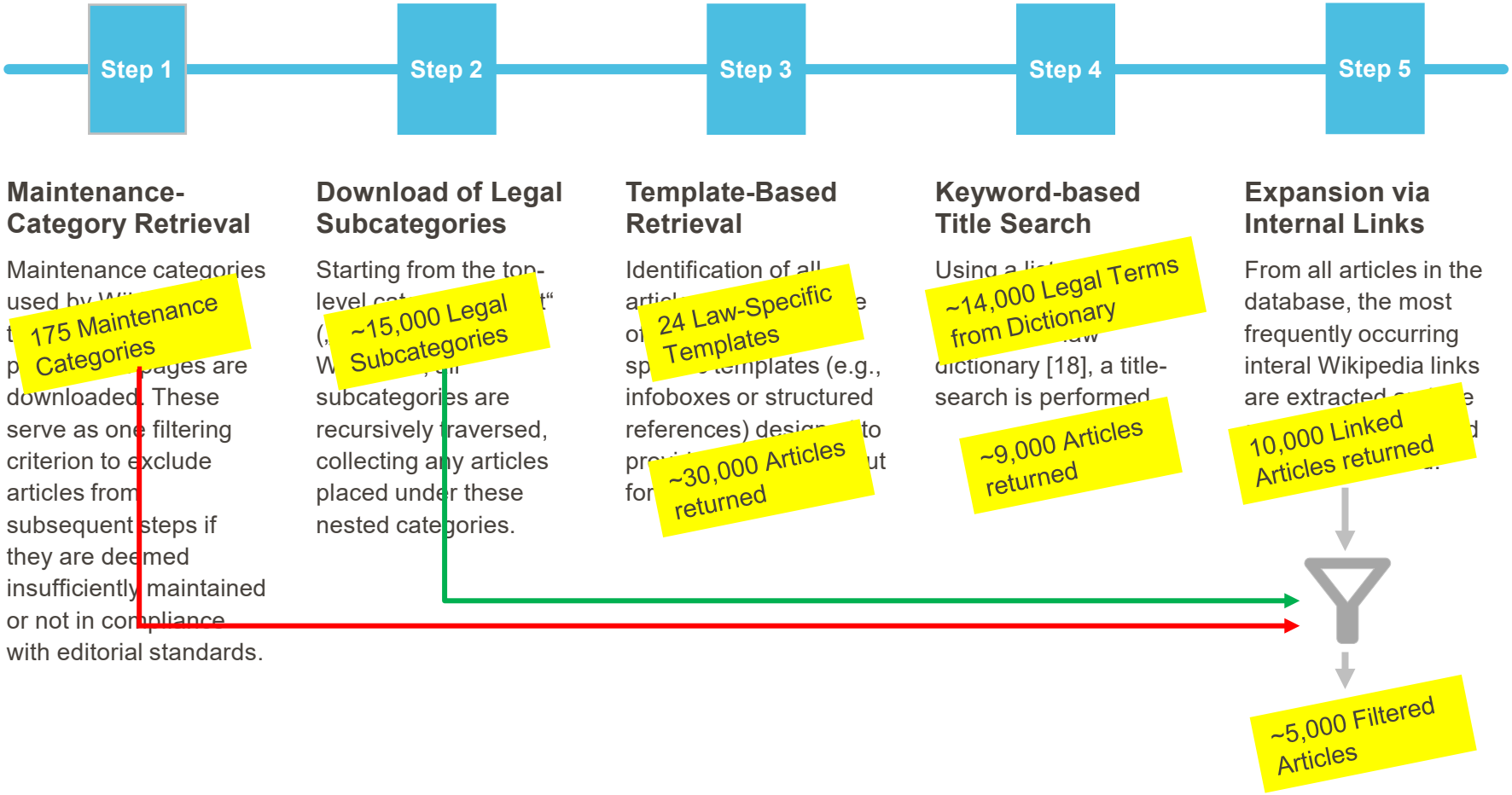
Hochschule **RheinMain**



LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API: Filtering Process

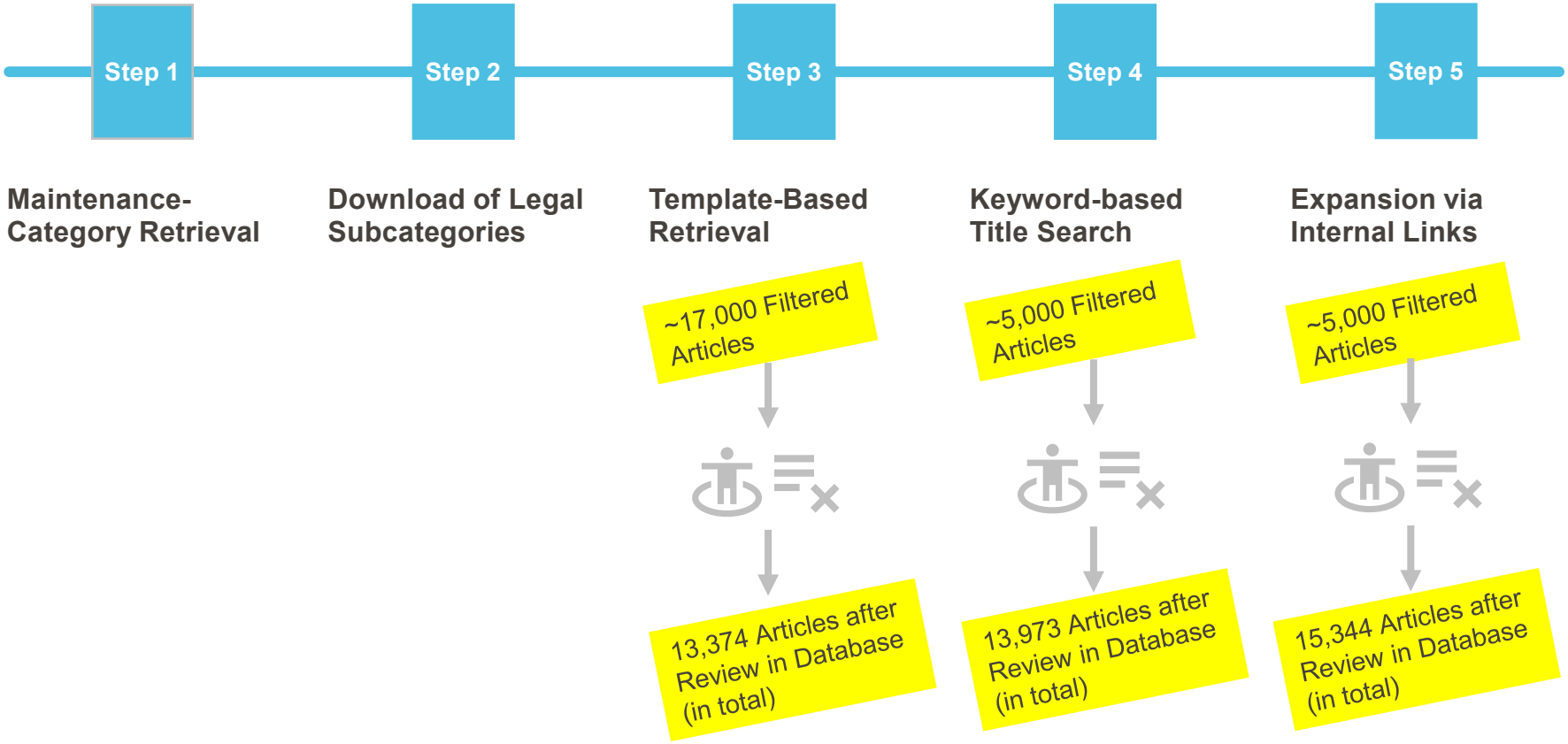
Hochschule RheinMain



LEGAL DOMAIN IN GERMAN WIKIPEDIA

Downloading process using Wikipedia‘s API: Manual Review of Categories

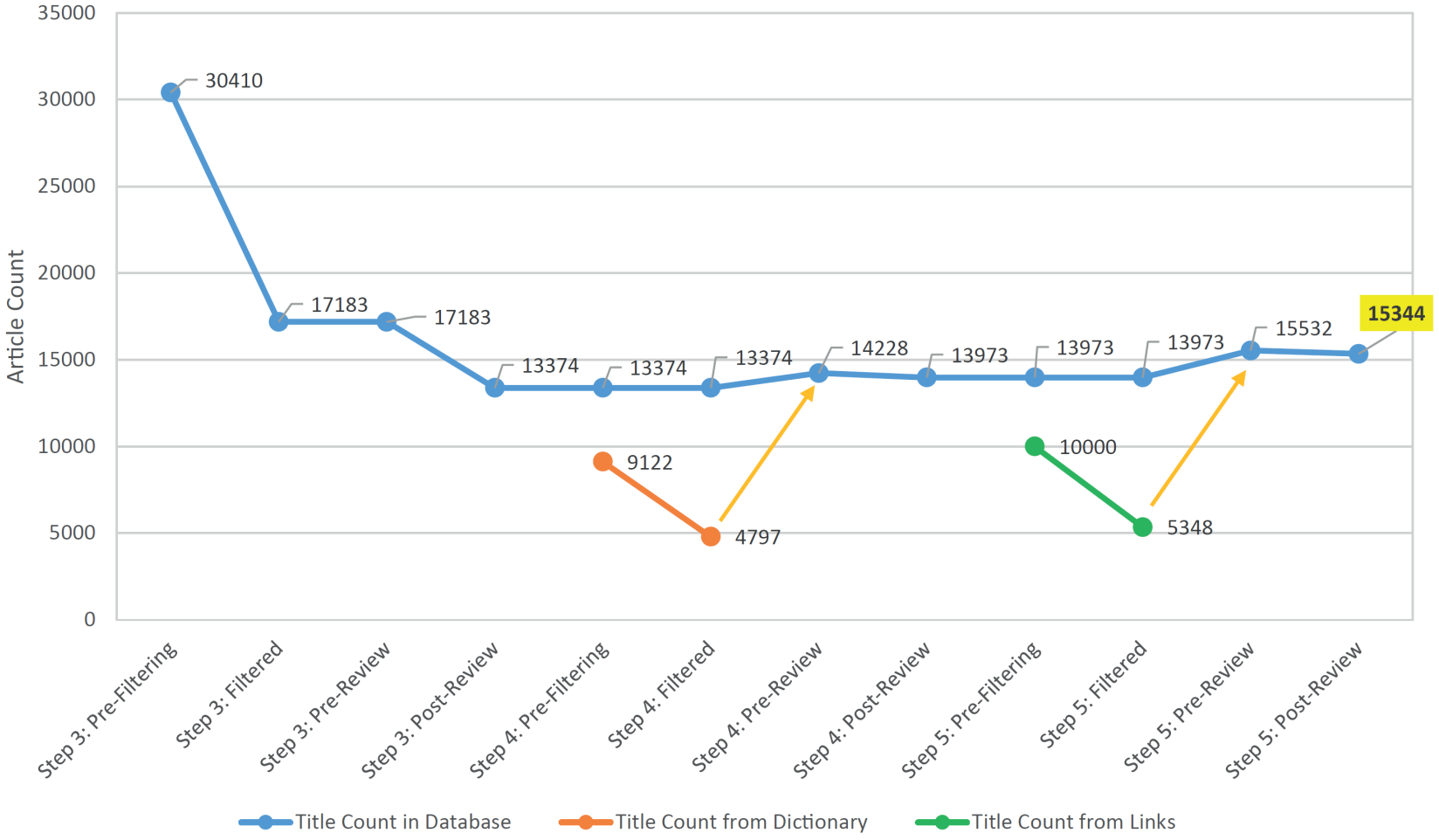
Hochschule **RheinMain**



LEGAL DOMAIN IN GERMAN WIKIPEDIA

15344 articles in database after 5 step procedure

Hochschule **RheinMain**



FINDINGS

Analysis of articles edited by permanently blocked users

BANNED USERS

Reasons for permanently blocking (banning) users

Hochschule RheinMain

Reason	No. of Users	Percentage
Non-Compliance	183,756	99%
At Own Request	1,455	0.8%
Deceased	344	0.2%
All Permanently Blocked Users	185,555	100%

BANNED USERS

Reasons for „Non-Compliance“

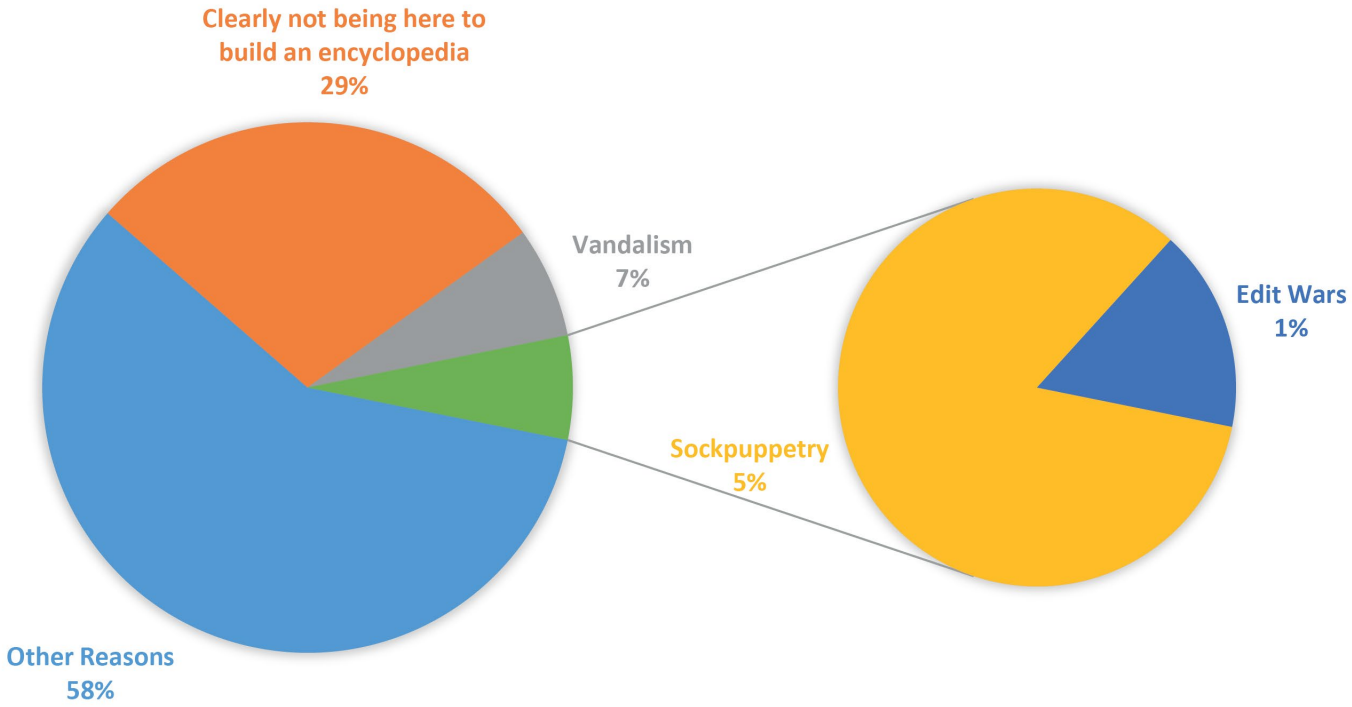
- “**Sockpuppetry**” denotes a clear **strategy for infiltration**: manipulated accounts operated under pseudonyms, sometimes identified through investigations documented by both investigative journalism and Wikipedia itself [19] [20] [21]

Source: Wikimedia Commons

BANNED USERS

Reasons for „Non-Compliance“

- “**Sockpuppetry**” denotes a clear **strategy for infiltration**: manipulated accounts operated under pseudonyms, sometimes identified through investigations documented by both investigative journalism and Wikipedia itself [19] [20] [21]

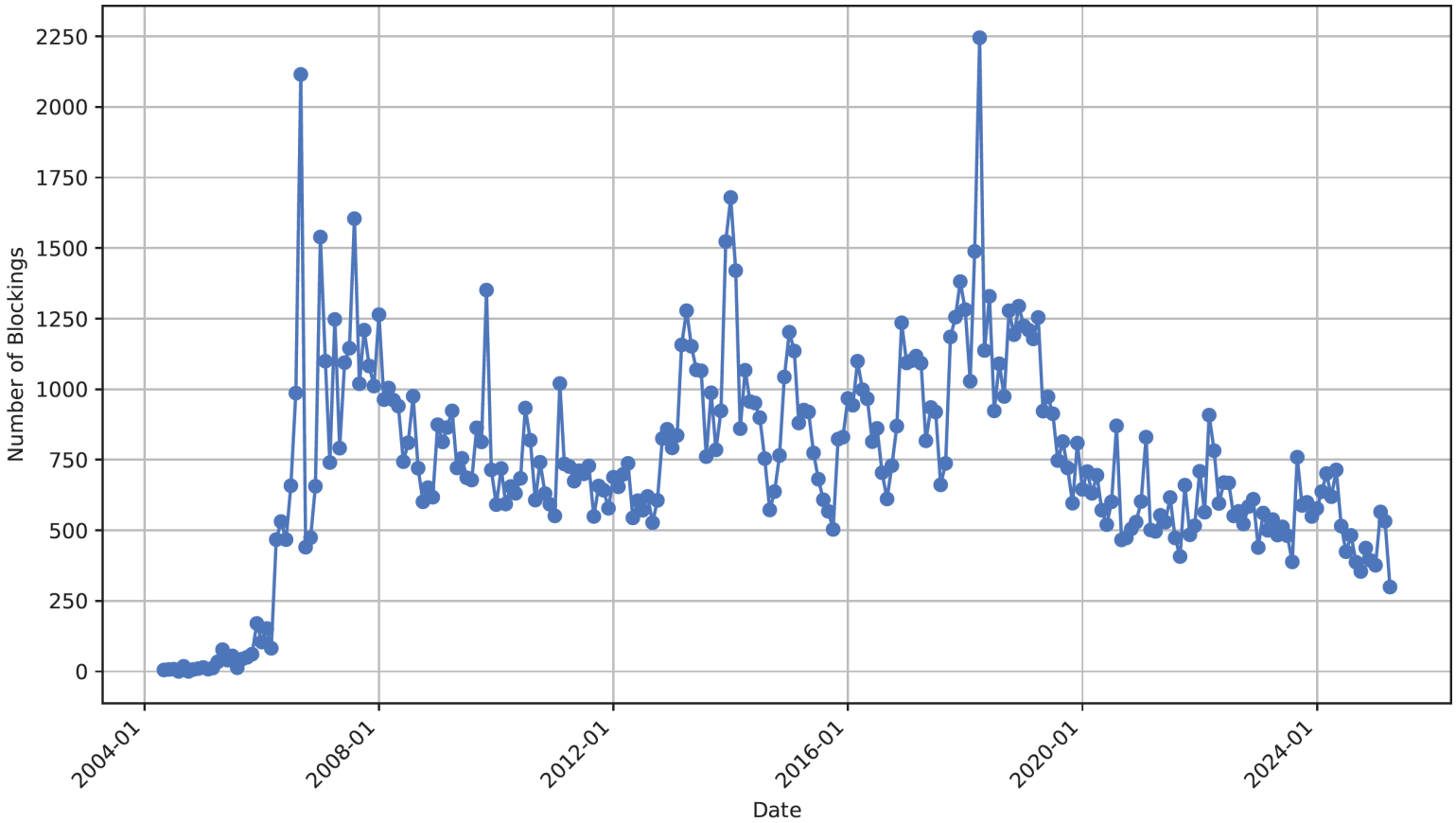


BANNED USERS PER MONTH

Permanently blocked, trend over the last 21 years

Hochschule **RheinMain**

Absolute numbers:

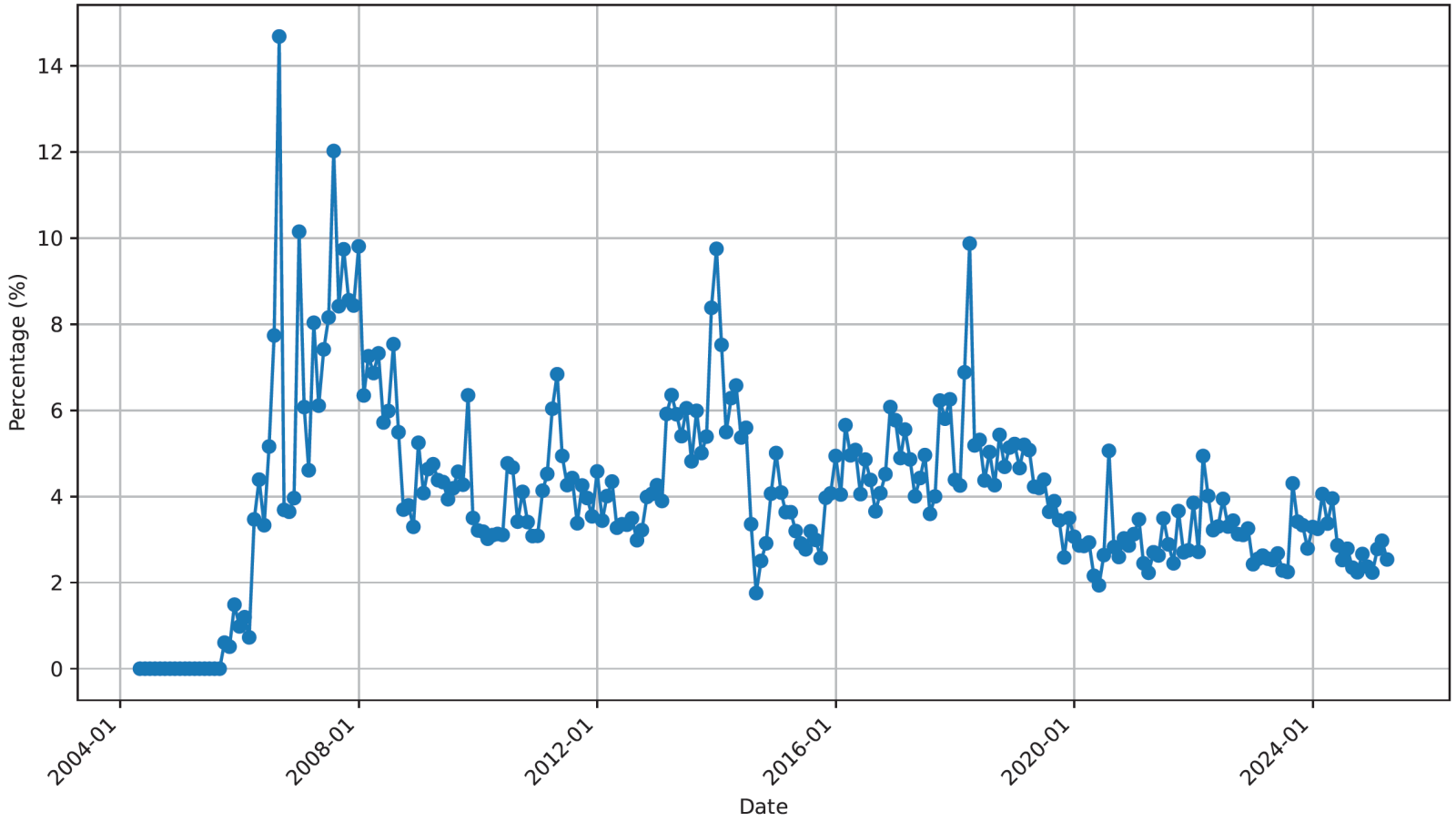


BANNED USERS PER MONTH

Permanently blocked, trend over the last 21 years

Hochschule **RheinMain**

Relative numbers (banned users in relation to new user registrations):

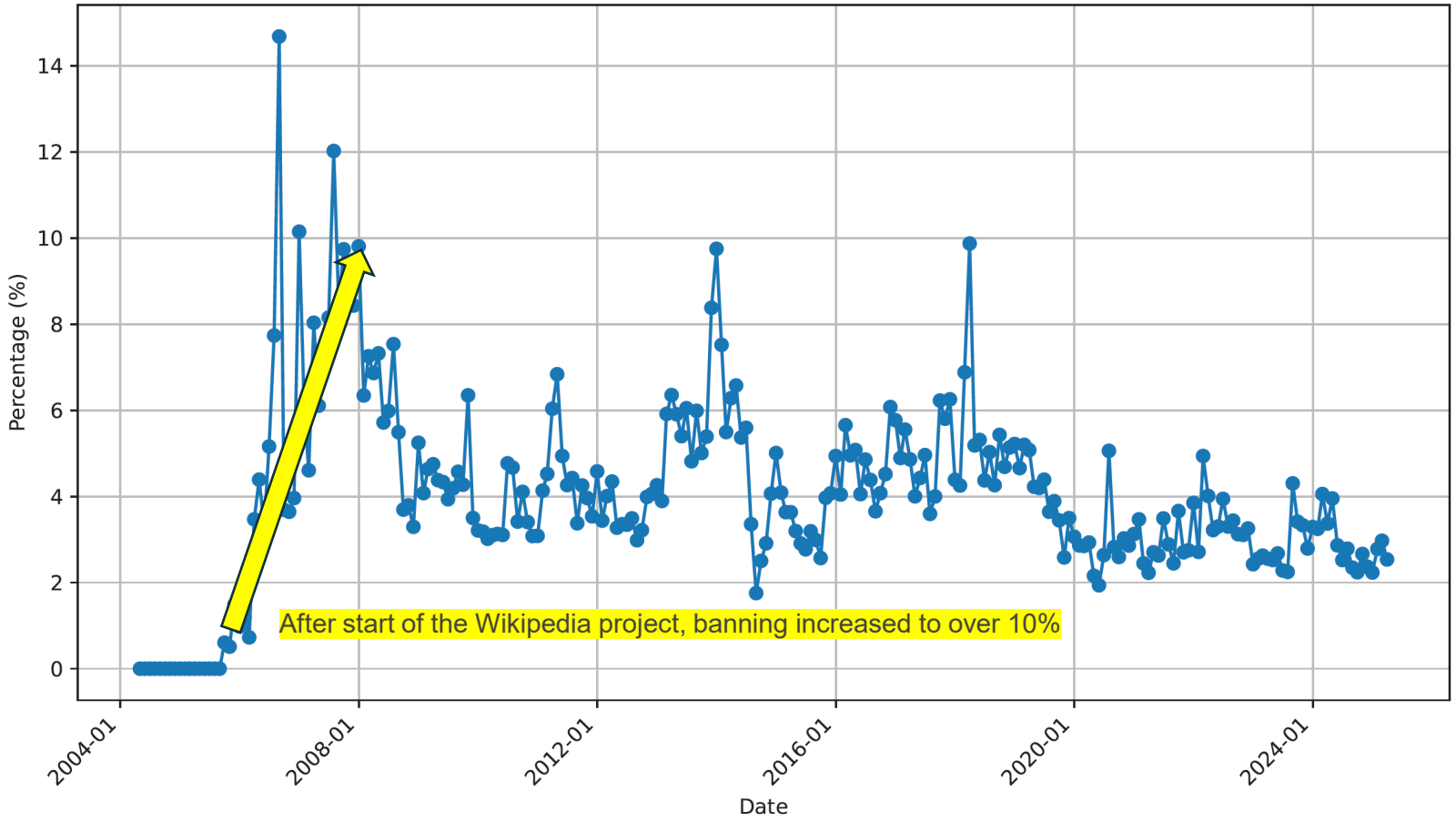


BANNED USERS PER MONTH

Permanently blocked, trend over the last 21 years

Hochschule **RheinMain**

Relative numbers (banned users in relation to new user registrations):

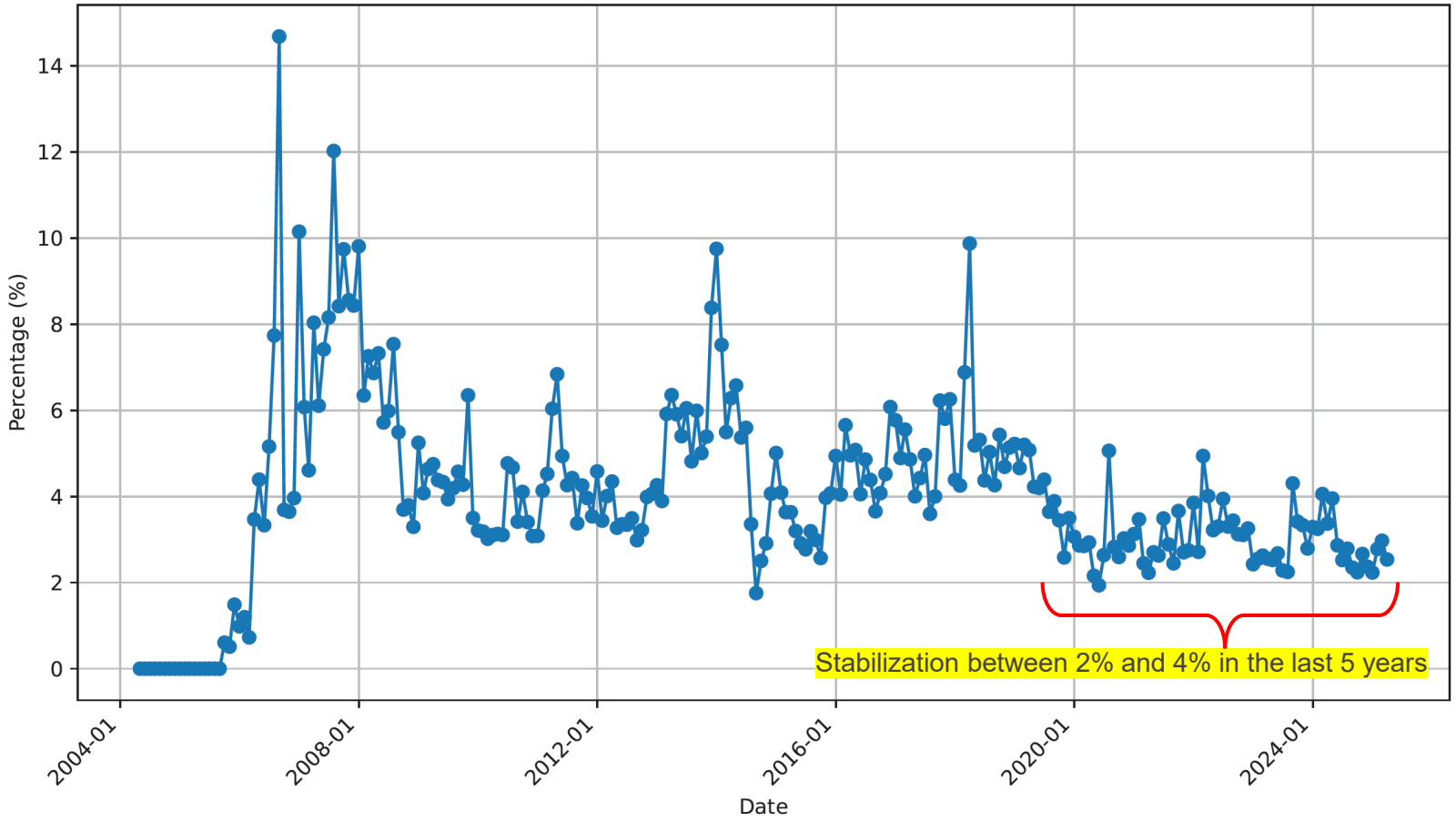


BANNED USERS PER MONTH

Permanently blocked, trend over the last 21 years

Hochschule **RheinMain**

Relative numbers (banned users in relation to new user registrations):

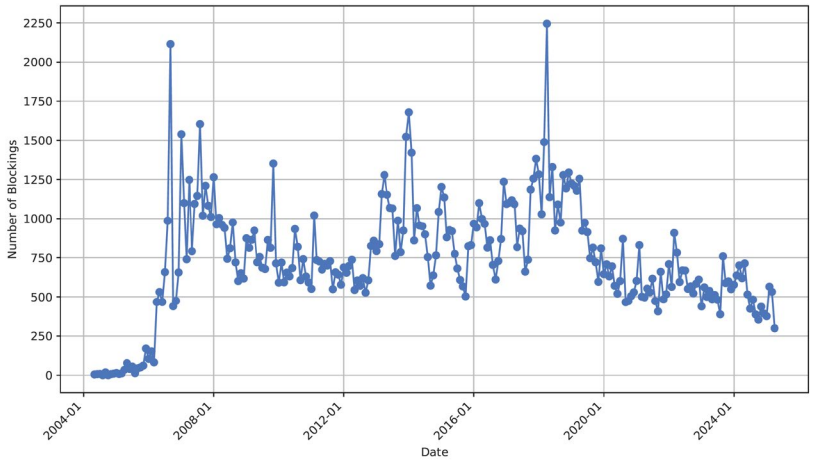


BANNED USERS PER MONTH

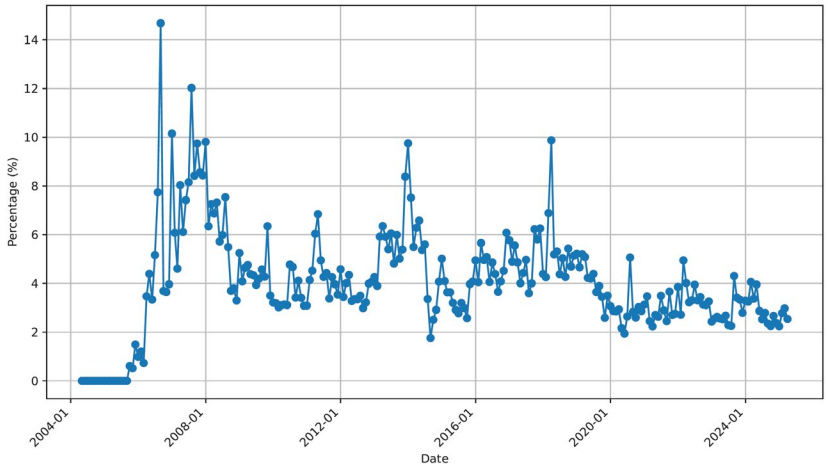
Permanently blocked, trend over the last 21 years

Hochschule **RheinMain**

Absolute numbers:



Relative numbers (banned users in relation to new user registrations):



BANNING RATIO IN REVISIONS

Number of revisions with bans in relation to all

Hochschule **RheinMain**

- **Banning Ratio** for **Revisions** (`BR_rev`):
  Number of revisions of a certain article (in the legal domain) originating from a banned user in relation to the total number of revisions of that article

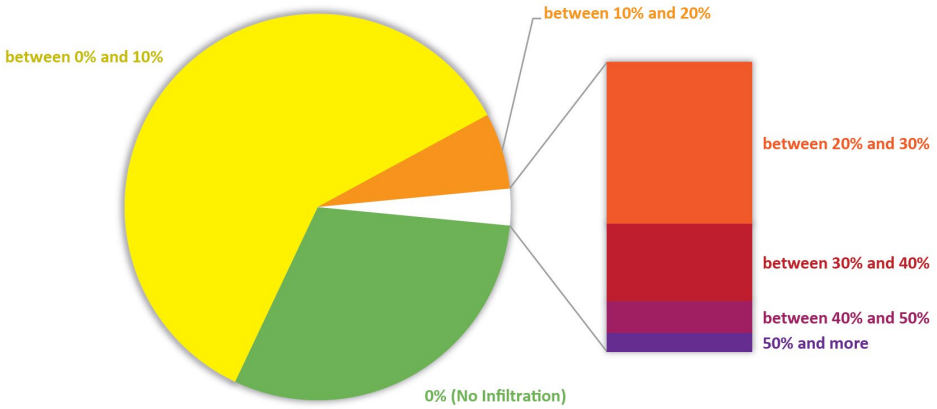
$$\texttt{BR\_rev} = \frac{\texttt{\#rev\_banned\_users}}{\texttt{\#rev\_total}} \quad \forall articles$$

BANNING RATIO IN REVISIONS

Number of revisions with bans in relation to all

Hochschule RheinMain

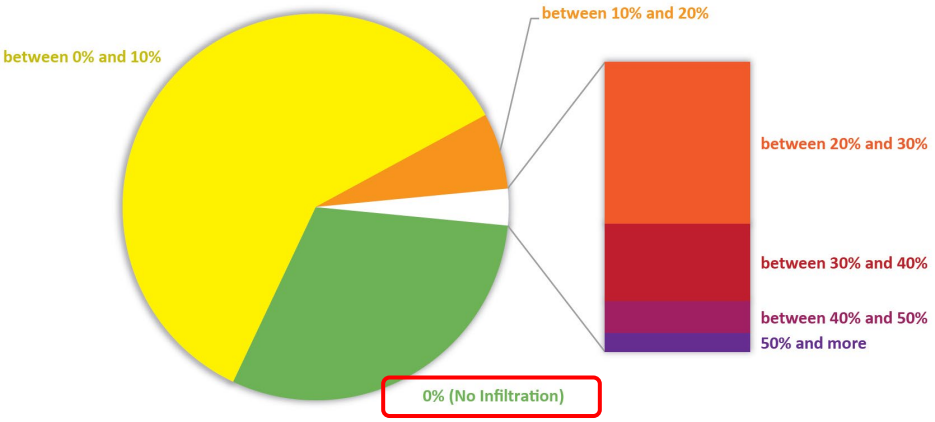
Banning Ratio	Articles	Percentage
0% (no infiltration)	4,682	30.51%
between 0% and 10%	9,216	60.00%
between 10% and 20%	978	6.37%
between 20% and 30%	261	1.70%
between 30% and 40%	125	0.81%
between 40% and 50%	52	0.34%
more than 50%	30	0.20%
	15,344	100%



BANNING RATIO IN REVISIONS

Number of revisions with bans in relation to all

Banning Ratio	Articles	Percentage
0% (no infiltration)	4,682	30.51%
between 0% and 10%	9,216	60.00%
between 10% and 20%	978	6.37%
between 20% and 30%	261	1.70%
between 30% and 40%	125	0.81%
between 40% and 50%	52	0.34%
more than 50%	30	0.20%
	15,344	100%



BANNING RATIO IN REVISIONS

Number of revisions with bans in relation to all

Hochschule **RheinMain**

- **Banning Ratio** for **Revisions**:
  Number of revisions of a certain article (in the legal domain) originating from a banned user in relation to the total number of revisions of that article

- Only **30.51%** of articles have revision **from good-faith editors solely**.

- **~1%** of articles have their **last revision** authored by a user **later banned**.

BANNING RATIO IN DISCUSSION

Posts in discussion with bans in relation to all

Hochschule **RheinMain**

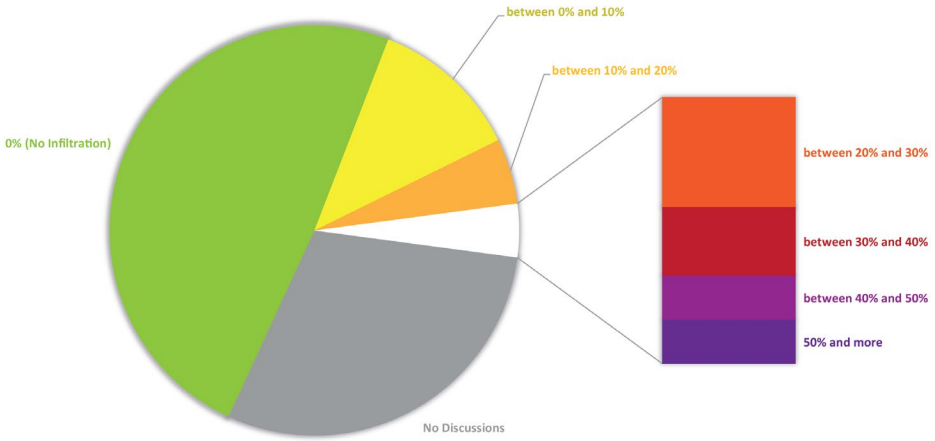
- **Banning Ratio** for **Discussion** (BR_disc):
  Number of contributions (posts) on the discussion page of a certain article (in the legal domain) originating from a banned user in relation to the total number of contributions for that article

$$\text{BR\_disc} = \frac{\#\text{posts\_banned\_users}}{\#\text{posts\_total}} \quad \forall articles$$

BANNING RATIO IN DISCUSSION

Number of discussions with bans in relation to all

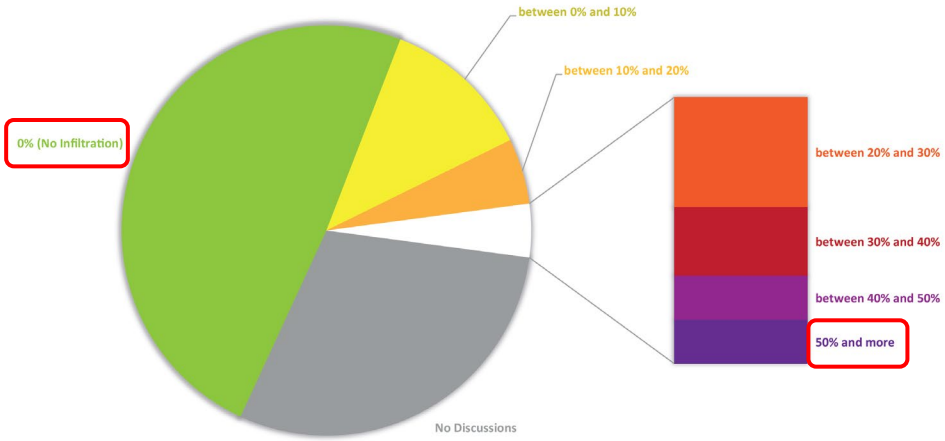
Banning Ratio	Articles	Percentage
No discussions for these articles	4,572	29.80%
0% (no banned contributors)	7,517	48.99%
between 0% and 10%	1,818	11.85%
between 10% and 20%	792	5.16%
between 20% and 30%	266	1.73%
between 30% and 40%	166	1.08%
between 40% and 50%	107	0.70%
more than 50%	106	0.69%
	15,344	100%



BANNING RATIO IN DISCUSSION

Number of discussions with bans in relation to all

Banning Ratio	Articles	Percentage
No discussions for these articles	4,572	29.80%
0% (no banned contributors)	7,517	48.99%
between 0% and 10%	1,818	11.85%
between 10% and 20%	792	5.16%
between 20% and 30%	266	1.73%
between 30% and 40%	166	1.08%
between 40% and 50%	107	0.70%
more than 50%	106	0.69%
	15,344	100%



BANNING RATIO IN DISCUSSION

Posts in discussion with bans in relation to all

Hochschule **RheinMain**

- **Banning Ratio** for **Discussion**:
  Number of contributions (posts) on the discussion page of a certain article (in the legal domain) originating from a banned user in relation to the total number of contributions for that article

- ~30% of articles have no associated dicussion page

- ~50% of articles have no banned contributors on discussions page

- ~21% of articles have banned contributors on discussions page, but only in 0.7% of articles these banned contributors dominated the discussion (majority of posts from banned users)

BANNING RATIO IN DISCUSSION

Posts in discussion with bans in relation to all

Hochschule **RheinMain**

- **Banning Ratio** for **Discussion**:
  Number of contributions (posts) on the discussion page of a certain article (in the legal domain) originating from a banned user in relation to the total number of contributions for that article

- ~30% of articles have no associated dicussion page

- ~50% of articles have no banned contributors on discussions page

- ~21% of articles have banned contributors on discussions page, but only in 0.7% of articles these banned contributors dominated the discussion (majority of posts from banned users)

➡ Not surprising: Malicious users tend to **avoid discussions** in favor of infiltrating the article itself in the first place

SUMMARY

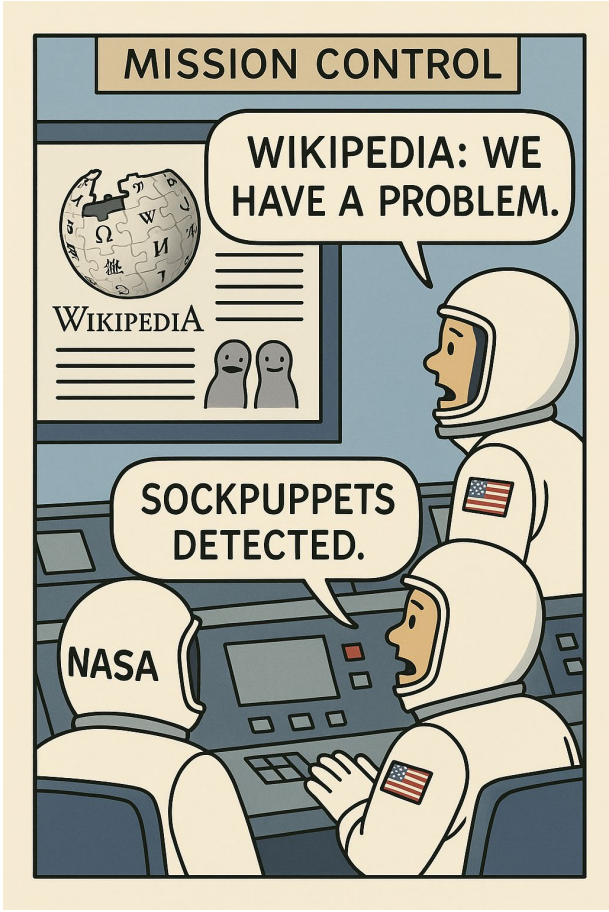
Extend of infiltration in German Wikipedia in a selected domain

SUMMARY

„Wikipedia: We Have a Problem!“

Hochschule **RheinMain**

- In summary, **even a specialized,** seemingly **neutral topic**, such as “Recht” (law) on the German **Wikipedia**, exhibits **clear patterns of infiltration** by permanently banned contributors.
- The belief that **crowdsourced content** will **always self-correct** through sheer volume of contributors is challenged by the persistent manipulation attempts observed here.
- **Countermeasures**, including **refined scraping**, filtering processes, and real-time oversight, are **crucial steps toward ensuring the continued integrity** of open knowledge ecosystems.



OUTLOOK

Things to do ...

Hochschule RheinMain

- **Broader Multi-Domain Analysis**: Replicating this methodology for additional subject areas would clarify whether the observed infiltration patterns are specific to the legal sphere or mirrored across other domains.

- **Refined Filtering for AI Data**: For RAG-based systems or training pipelines, removing or downweighting articles that show high infiltration scores and introducing a reliability metric into prompts can reduce the risk of providing manipulated content to end users.

- **Ongoing Community Oversight**: As infiltration continues to evolve, a coordinated effort by Wikipedia administrators and community volunteers is essential.

REFERENCES AND CONTACT INFORMATION

Comments and discussion always welcome!

SELECTED REFERENCES

Full list: see paper

1. A. Wan, E. Wallace, S. Shen, and D. Klein, “Poisoning language models during instruction tuning,” in Proceedings of the 40th International Conference on Machine Learning, ICML’23, pp. 35413–35425, JMLR.org, 2023.

…

19. Wikipedia contributors, “Sock puppet account.” https://en.wikipedia.org/wiki/Sock puppet account, 2025. Accessed: 2025-05-15.
20. Wikipedia contributors, “Wikipedia:sockpuppet investigations.” https://en.wikipedia.org/wiki/Wikipedia:Sockpuppet investigations, 2025. Accessed: 2025-05-15.
21. Wikipedia contributors, “Wikipedia:sockpuppetry.” https://en.wikipedia.org/wiki/Wikipedia:Sockpuppetry, 2025. Accessed: 2025-05-15.
22. C. Schattleitner and D. Laufer, “Sock puppet zoo – attack on wikipedia.” https://www.ardaudiothek.de/sendung/sockenpuppenzooangriff-auf-wikipedia/13996869/, 2025. Podcast.

THANK YOU FOR LISTENING

Hochschule **RheinMain**

Contact
Prof. Dr. Matthias Harter
Faculty of Engineering
Department of Electrical Engineering and Information Technology

Am Brückweg 26
D-65428 Rüsselsheim

+49 6142 898-4223
matthias.harter@hs-rm.de