Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Digital preservation of academic heritage: a computational analysis of the Agora Discussion Club at Kryvyi Rih State Pedagogical University (2016-2025)

Natalia A. Pechenina
Kryvyi Rih State Pedagogical University
Serhii S. Korniienko
Kryvyi Rih State Pedagogical University
Serhiy O. Semerikov
semerikov@gmail.com

Kryvyi Rih State Pedagogical University

Article

Keywords:

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8663285/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Digital preservation of academic heritage: a computational analysis of the Agora Discussion Club at Kryvyi Rih State Pedagogical University (2016–2025)

1 Introduction

Academic discussion clubs have long served as informal spaces for intellectual exchange within university settings, fostering dialogue between students, faculty, and invited guests on topics spanning disciplinary boundaries. While such forums contribute significantly to the intellectual culture of their institutions, they frequently remain undocumented beyond institutional memory, leaving their contributions to academic discourse unpreserved and unanalysed (Barcelona & Dempsey, 2024; Becker, 2017). This lacuna is particularly pronounced for institutions in regions experiencing conflict or instability, where traditional archival practices may be disrupted and digital preservation becomes both more urgent and more challenging.

The present study addresses this gap through a computational analysis of the Agora Discussion Club, an academic forum established in October 2016 within the History Department at Kryvyi Rih State Pedagogical University (KSPU) in Ukraine. Named after the ancient Greek public gathering space, the club has hosted regular discussion sessions on historical, philosophical, cultural, and contemporary topics, maintaining its activities through the onset of Russia's full-scale invasion in February 2022 and continuing to the present.

This research pursues three primary objectives. First, we demonstrate methods for extracting, processing, and preserving web-based academic content that may otherwise be lost to link rot, server failures, or institutional changes. Second, we apply natural language processing techniques to analyse the thematic content, participant networks, and temporal evolution of academic discourse within a regional Ukrainian institution. Third, we release a bilingual corpus that may serve as a resource for future research on Ukrainian academic culture, digital humanities methodologies, and the impact of conflict on intellectual life.

The significance of this work extends beyond the specific case study. As universities increasingly document their activities through web platforms rather than traditional archives, computational methods become essential for preserving institutional memory. The methods developed here may be adapted for similar preservation efforts at

other institutions, particularly those operating in contexts where traditional archiving infrastructure is limited or threatened.

2 Background

2.1 Digital preservation of academic heritage

The digital turn in higher education has transformed how institutions document and disseminate their activities, yet this transformation has introduced new vulnerabilities. Studies of link decay demonstrate that web content has a median lifespan of approximately 9.3 years, with academic resources particularly susceptible to loss during institutional transitions (Burnhill, Mewissen, & Wincewicz, 2015). Web archiving initiatives such as the Internet Archive's Wayback Machine provide partial solutions, but coverage remains incomplete for non-English language content and regional institutions outside major metropolitan areas.

Recent scholarship in digital humanities has emphasised the importance of proactive digital preservation, particularly for materials documenting marginalised or endangered communities (Risam, 2019). This perspective aligns with broader movements toward "digital cultural heritage" preservation, which recognises that born-digital materials require intentional archival strategies distinct from those applied to physical collections.

2.2 Computational analysis of academic discourse

Natural language processing techniques have been increasingly applied to analyse academic communication patterns, from citation networks to collaborative authorship patterns (Fortunato et al., 2018). Topic modelling approaches such as Latent Dirichlet Allocation (LDA) and, more recently, transformer-based methods like BERTopic (Grootendorst, 2022), enable researchers to identify thematic structures within document collections without requiring manual coding.

For multilingual corpora, particularly those involving low-resource languages, recent advances in cross-lingual transfer learning have expanded analytical possibilities (Liashenko & Semerikov, 2024). Ukrainian NLP has benefited from developments such as the uk_core_news models for spaCy and multilingual sentence transformers that support Cyrillic scripts, though significant gaps remain compared to high-resource languages (Slobodianiuk & Semerikov, 2024).

Research has demonstrated that large language models (LLMs) can outperform traditional OCR systems for historical document transcription, with studies showing up to 55% reduction in character error rates when using models such as GPT-4 for page-level processing (Kim, Baudru, Ryckbosch, Bersini, & Ginis, 2025). These advances are particularly relevant for processing documents in low-resource languages where specialised OCR models may be unavailable.

2.3 Ukrainian higher education context

Ukrainian universities operate within a distinctive institutional context shaped by post-Soviet transitions, ongoing educational reforms, and, since 2014, the challenges

posed by Russian aggression. Kryvyi Rih State Pedagogical University, founded in 1930, exemplifies regional pedagogical institutions that serve as primary venues for teacher training and historical scholarship outside major urban centres (Kramarenko & Semerikov, 2025).

The Agora Discussion Club (Дискусійний клуб «Агора») emerged within KSPU's History Department as an initiative to foster student engagement with historical and contemporary issues beyond formal coursework. Founded in October 2016 through the initiative of Dr Natalia A. Pechenina and Dr Vadym O. Yashin, the club established a regular schedule of presentations and discussions that has continued through significant disruptions, including the COVID-19 pandemic and the 2022 invasion.

The club's mission is articulated in its founding documents: "The Agora Discussion Club is a platform for free discussion of any questions that have at least the slightest relation to history. The topics of sessions are freely determined by presenters. All sessions are open, admission is free. Anyone from the Academic community of historians can volunteer to be a presenter." (Pechenina & Yashin, 2020, p. 1) This open-access model reflects the club's commitment to democratising academic discourse within the university community.

A significant output of the club's activities is the publication *Studia Hierodotica* ("Herodotus Studies"), a 49-page compilation of selected session materials published in 2020 (Pechenina & Yashin, 2020). Named after Herodotus of Halicarnassus – the ancient Greek historian traditionally regarded as the "Father of History" – the publication preserves full-text versions of six presentations from sessions #48–50, including topics such as vampire imagery in mass culture, the Spanish Inquisition, gender in Soviet intelligence, cosmetics history, and the Space Race. The editorial preface quotes Herodotus's opening declaration from the *Histories*: "I have written this book so that nothing done by men may be forgotten over time, and that the great and wonderful deeds of both Greeks and barbarians may not go unrecorded." This classical allusion positions the club's activities within a tradition of preserving historical memory stretching back to antiquity.

3 Data and methods

3.1 Data collection

We collected data from the official web page of the Agora Discussion Club hosted on the KSPU website (https://kdpu.edu.ua/istorii/zahalna-informatsiia/pro-spetsialnist-istoriia/diskusijnij-klub-agora.html). The page utilised a Joomla content management system with paginated display of session announcements across nine pages. Using Python with Selenium WebDriver and BeautifulSoup, we implemented a two-phase scraping strategy: first collecting session URLs from the paginated index pages (`?start=0` through `?start=80`), then extracting detailed content from each individual session page.

For each session, we extracted: (1) session number and title (in Ukrainian); (2) date of publication; (3) description text; (4) view count as a measure of engagement; (5) embedded image URLs; and (6) links to external video content on Facebook and YouTube platforms. Sessions spanning from session #1 (October 2016) to session #80

(January 2025) were collected, representing the complete documented history of the club at the time of data collection.

Images were downloaded and processed using Tesseract OCR with the Ukrainian language model (tesseract-ocr-ukr) to extract text from presentation slides, posters, and informational graphics embedded within session announcements. Of 748 session-specific images downloaded (after removing 4 template images that appeared across all sessions), 428 contained unique content not duplicated across sessions. OCR extraction yielded meaningful text from 38 sessions, with 446 images containing extractable text content with reasonable confidence scores, primarily from presentation materials used during sessions. This OCR-extracted text was integrated with session descriptions to enrich the corpus for named entity recognition.

Additionally, we obtained the *Studia Hierodotica* publication, a 49-page PDF compilation of selected session materials available from the club's archive on Google Drive. This publication, containing 64,513 characters of Ukrainian text, provided full presentation transcripts from six sessions, substantially enriching the corpus with extended discourse beyond the web summaries. The PDF was processed using PyMuPDF (fitz) for text extraction, yielding complete article texts for sessions #48 (Lipstick: History Through the Ages), #49 (Vampires: Evolution of the Image in Mass Culture), and #50 (Inquisition: Debunking the "Black Legend"), along with three additional presentations on art and stereotypes, women in Soviet intelligence, and the Space Race.

Video recordings from four sessions were obtained from the club's Google Drive archive and transcribed using OpenAI Whisper (medium model) with Ukrainian language specification. The seven video files, totalling approximately 170 minutes, were processed on CPU using FP32 precision. Whisper produced timestamped transcripts in multiple formats (plain text, SRT subtitles, and JSON with word-level timing). Session #75 featured a literary discussion of Lesya Ukrainka's "Forest Song", while session #80 addressed artificial intelligence opportunities and threats.

3.2 Translation and bilingual corpus construction

Given that all source materials were in Ukrainian, we employed the Claude API (model: claude-sonnet-4-20250514) to translate session titles and descriptions to English. Translation prompts included contextual information about the academic setting and historical terminology to improve domain-specific translation quality. The prompt instructed the model to preserve academic terminology accurately and maintain the scholarly tone of the original text.

The resulting corpus maintains parallel Ukrainian-English text for all 80 session titles and 77 session descriptions (three sessions lacked substantial description text), enabling both monolingual analysis and cross-linguistic research applications. All translations were conducted programmatically with consistent prompting to ensure methodological reproducibility.

3.3 Natural language processing pipeline

Our NLP pipeline comprised several analytical components:

1. *Named entity recognition:* We employed a hybrid NER approach combining spaCy's Ukrainian language model (uk_core_news_lg) with post-processing for morphological normalization and entity enhancement. The spaCy model provided initial entity extraction across categories including persons (PER), organizations (ORG), and locations (LOC). Given the inflectional nature of Ukrainian, we applied morphological normalization using pymorphy3 to merge variant forms of the same entity (e.g., "Сергій Корнієнко" and "Сергія Корнієнка" represent the same person in nominative and genitive cases). Additional EVENT and WORK entities were extracted through pattern matching against known historical events and cultural works mentioned in session titles. This approach, enhanced with OCR-extracted text from session images, yielded 714 unique entities: 445 persons, 134 locations, 121 organizations, 7 historical events, and 7 works (Table 3).
2. *Network analysis:* Participant co-occurrence networks were constructed using the NetworkX library, where nodes represent individuals mentioned across multiple sessions and edges indicate co-mention within the same session. Edge weights correspond to the number of sessions in which two individuals were co-mentioned. Centrality metrics including degree centrality and betweenness centrality were calculated to identify key participants and potential bridges between different discussion communities.
3. *Temporal analysis:* Sessions were grouped by calendar year (extracted from publication dates) and analysed for thematic shifts using keyword-based classification. Eight thematic categories were defined based on preliminary qualitative analysis of session titles: ancient history, Ukrainian history, military topics, philosophy, religion, art and culture, social issues, and technology. Each session was assigned to one or more categories based on keyword presence in titles and descriptions.
4. *Sentiment analysis:* A lexicon-based approach was employed using curated lists of positive and negative terms relevant to academic discourse. While this approach has known limitations for nuanced sentiment detection, it provides a baseline measure of evaluative language distribution across sessions.
5. *Engagement analysis:* Spearman correlation coefficients were calculated between thematic category presence and session view counts to identify topics associated with differential audience engagement. View counts were retrieved directly from the website's display metadata.

3.4 Ethical considerations

All data were collected from publicly accessible web pages with no authentication requirements. The research analyses institutional content documenting public academic events rather than private communications. Extracted person names refer to individuals participating in public academic presentations, whose involvement was already documented on the institutional website. No ethics approval was required as the study does not involve human participants but rather computational analysis of publicly available institutional records documenting academic activities.

4 Results

4.1 Corpus statistics

The final corpus comprises 80 sessions documented between October 2016 and January 2025, spanning eight academic years. Of these, 77 sessions contained substantive description text suitable for text analysis. The corpus includes 748 session-specific images (after deduplication to remove 4 template images appearing across all sessions), with OCR extraction yielding meaningful text from 38 sessions with unique content. Additionally, the *Studia Hierodotica* publication contributed 64,513 characters of extended presentation text from six sessions, substantially enriching the available discourse data for sessions #48–50 and three additional presentations.

Total view counts across all sessions exceeded 56,900, with individual session views ranging from 276 (session #25: "Changes Happening Yesterday?") to 2,709 (session #70: "Doric Dialect"), yielding a mean of 712.1 views per session (SD = 483.7). The high-engagement sessions typically featured specialised linguistic or archaeological content, suggesting audience interest in scholarly depth.

4.2 Thematic coverage

Session topics demonstrated remarkable diversity, ranging from ancient history (e.g., "The Scythian Era: Weapons, Society, Woman", session #1) to contemporary technology (e.g., "AI 2025: Opportunities and Threats", session #80). Manual classification identified eight primary thematic categories with the distribution shown in Table 1.

Table 1: Thematic distribution across Agora sessions (N=80)

Theme	Sessions	Percentage
Ancient History	18	22.5%
Art & Culture	14	17.5%
Ukrainian History	12	15.0%
Philosophy	10	12.5%
Social Issues	9	11.3%
Military	8	10.0%
Technology	5	6.3%
Religion	4	5.0%

The predominance of ancient history topics reflects the historical specialisation of the sponsoring department, while the presence of technology and social issues topics indicates engagement with contemporary debates beyond traditional disciplinary boundaries.

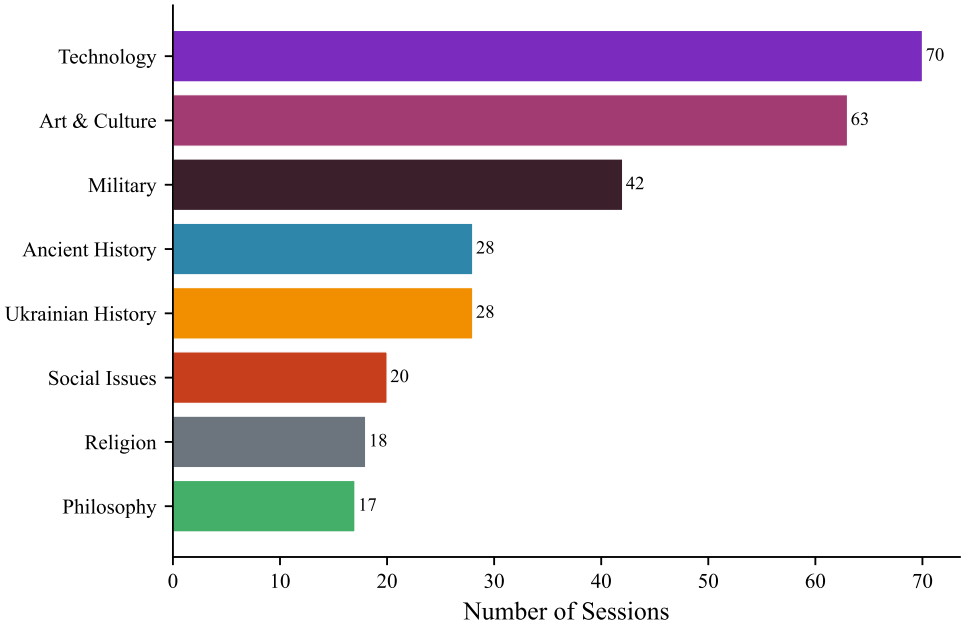


Figure 1: Thematic distribution of Agora Discussion Club sessions (N=80). Sessions were classified using keyword-based analysis of titles and descriptions. Multiple themes may apply to a single session.

4.3 Temporal evolution

Activity levels varied across the club's history, with peak activity during academic years 2017–2019 and notable continuity despite external disruptions. Figure 2 illustrates the temporal distribution of sessions by academic year.

Session #77, held on 22 February 2022 – one day before Russia's full-scale invasion – addressed the topic of "Assassins vs. Templars", reflecting the club's engagement with both historical and popular culture themes. Subsequent sessions demonstrated adaptation to wartime conditions, with session #78 (September 2022) marking the club's sixth anniversary and demonstrating institutional resilience.

The most recent sessions (#79–80) addressed artificial intelligence and its implications for historical research, indicating continued engagement with contemporary methodological developments despite ongoing conflict. Session #79, "ChatGPT: What Can Historians Ask It?" (March 2023), received 1,468 views, suggesting strong interest in digital methodology among the audience.

4.4 Participant network

Named entity recognition identified 445 unique person entities across the corpus, with 32 individuals appearing in two or more sessions (Table 4). The most active presenter, Serhii Korniienko, appeared in 14 sessions, followed by Dmytro Chyzhyk with 6

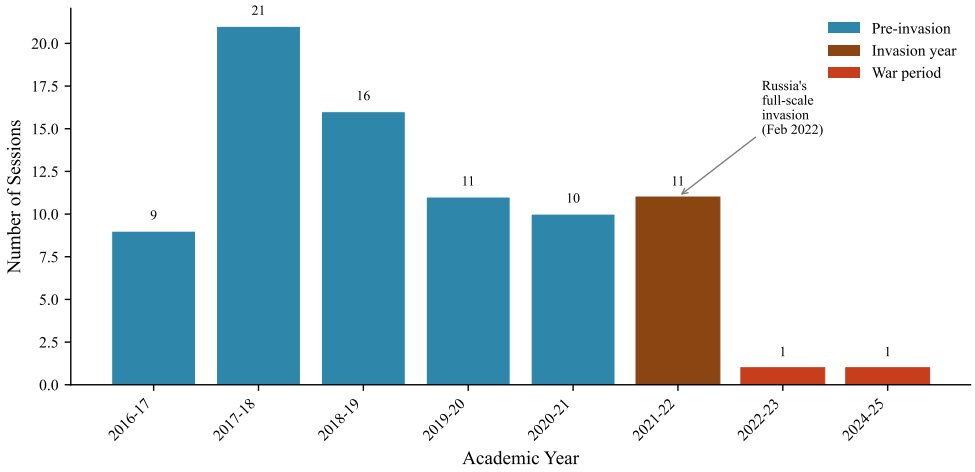


Figure 2: Temporal distribution of Agora Discussion Club sessions by academic year (2016–2025). Blue bars indicate pre-invasion sessions; red bars indicate sessions during the war period following Russia's full-scale invasion (February 2022). Academic years run September–June.

appearances. The co-occurrence network comprised 15 nodes (individuals appearing across three or more sessions) connected by 22 edges. Network analysis revealed a core group of regular participants with the following characteristics:

- Mean degree centrality: 2.93
- Maximum betweenness centrality: 0.42
- Network density: 0.21

The network structure suggests a central core of faculty and active student participants, with peripheral involvement from guest speakers and occasional presenters. Several student presenters appeared across multiple academic years, suggesting the club's role in developing student presentation skills and fostering ongoing intellectual engagement beyond single-event participation (Figure 3).

4.5 Engagement patterns

Correlation analysis between thematic categories and view counts revealed differential audience engagement across topics (Table 2, Figure 4).

Military and technology themes showed the strongest positive associations with engagement, while philosophy and religion themes showed slight negative associations. These patterns may reflect broader public interest in topics with contemporary relevance or novelty, though the modest correlation magnitudes suggest that engagement is influenced by multiple factors beyond thematic category alone.

Figure 5 presents a timeline view of all sessions, with bubble size indicating view count and colour intensity showing relative engagement levels.

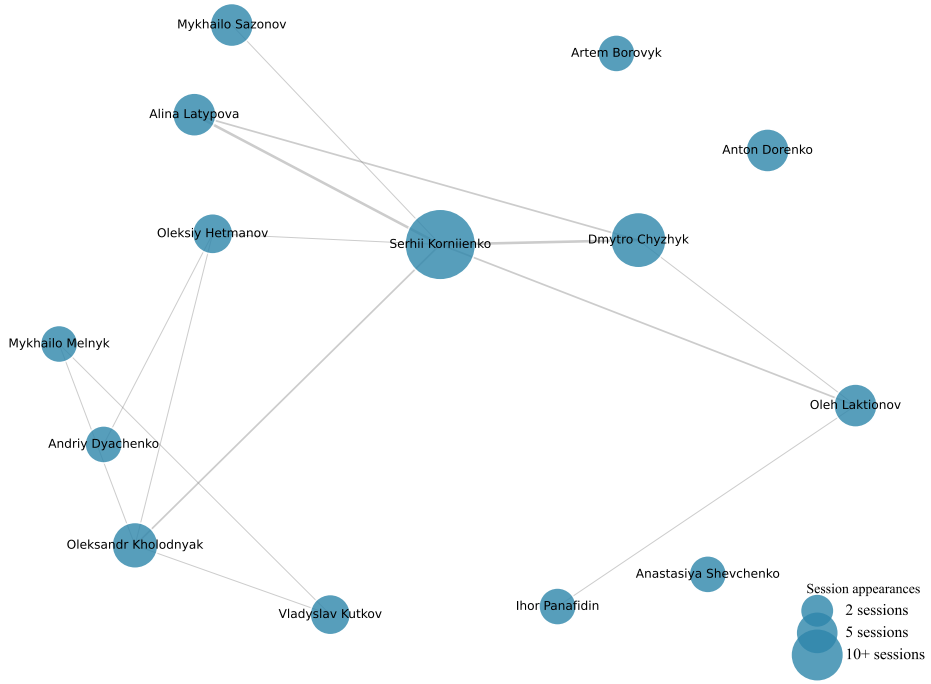


Figure 3: Participant co-occurrence network of the Agora Discussion Club. Nodes represent individuals mentioned across multiple sessions; edges indicate co-occurrence within the same session. Node size reflects the number of sessions in which the participant appeared.

4.6 Corpus complexity and word count analysis

Analysis of session description lengths revealed variation in documentation depth across the corpus. Figure 6 shows the distribution of word counts across session descriptions.

4.7 Pre-war and wartime comparison

Comparing sessions before Russia's full-scale invasion (February 2022) with those during the wartime period reveals patterns of institutional adaptation. Figure 7 illustrates key differences between the two periods.

4.8 Topic evolution

Thematic emphases shifted across academic years, reflecting both disciplinary interests and contemporary events. Figure 8 traces the evolution of major topic categories over the club's history.

Table 2: Correlation between thematic categories and view counts

Theme	Correlation (r)
Military	0.117
Social issues	0.111
Art & Culture	0.109
Technology	0.093
Ancient History	0.083
Ukrainian History	0.062
Religion	−0.146
Philosophy	−0.176

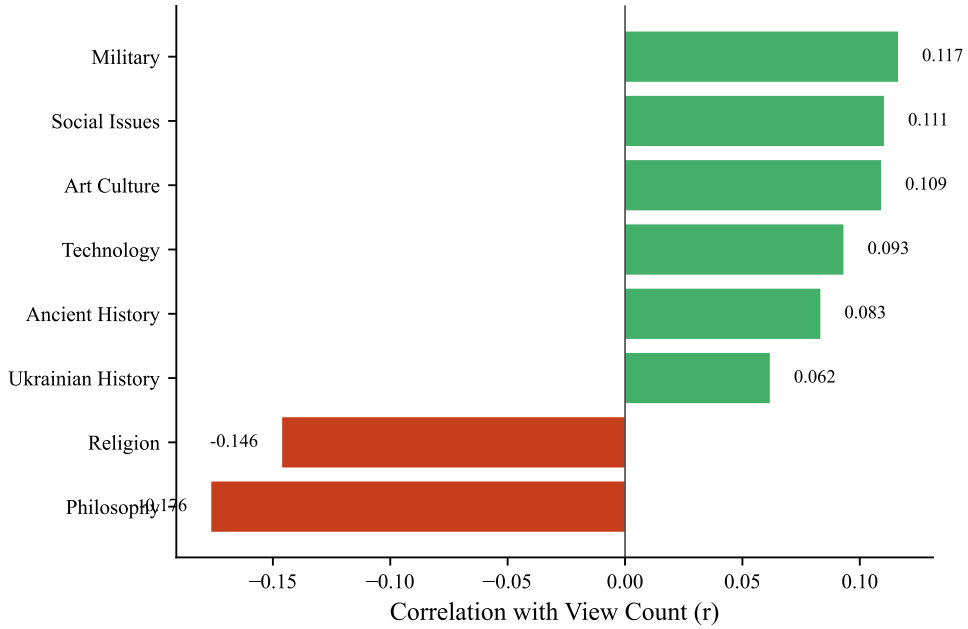


Figure 4: Correlation between thematic categories and session view counts. Green bars indicate positive correlations (higher engagement); red bars indicate negative correlations. Correlation coefficients (Pearson's r) shown for each theme.

4.9 Named entity analysis

Our enhanced NER pipeline, enriched with OCR-extracted text from 38 sessions containing unique image content, extracted 714 unique entities across five categories (Table 3). Person entities (445) dominated the corpus, reflecting the discussion format where individual presenters and historical figures are prominently mentioned.

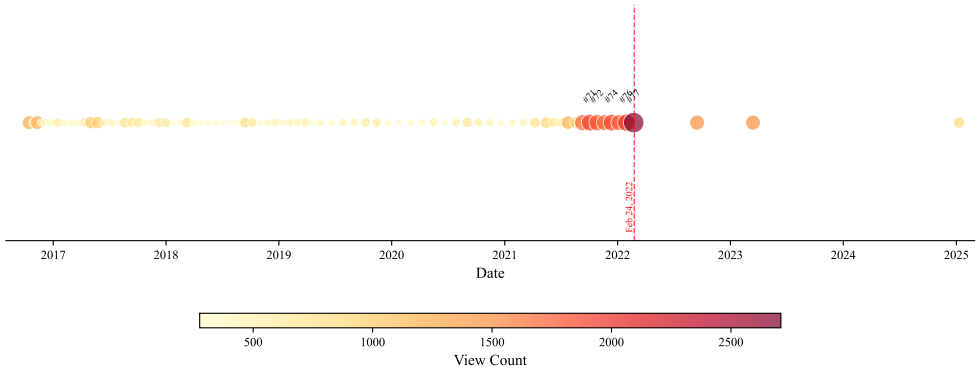


Figure 5: Timeline of Agora Discussion Club sessions (2016–2025). Bubble size and colour intensity indicate view count. The vertical dashed line marks Russia's full-scale invasion (24 February 2022). Session numbers of highest-engagement sessions are annotated.

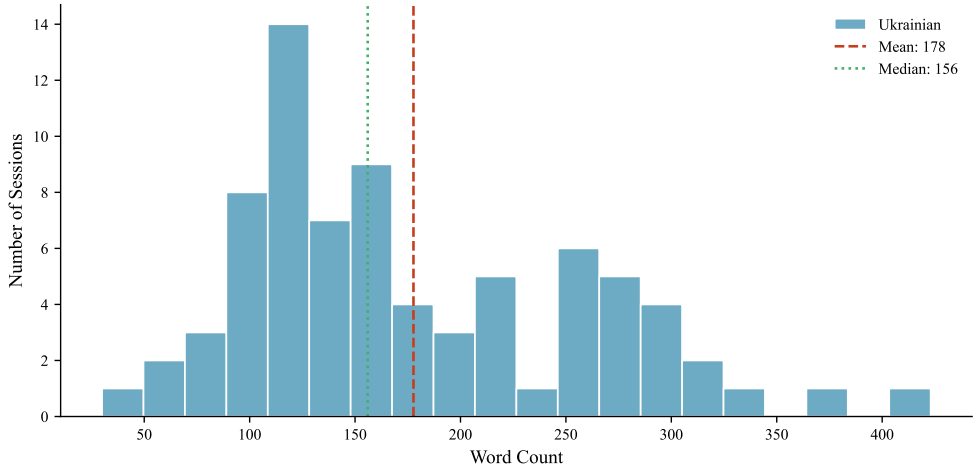


Figure 6: Distribution of word counts across Agora session descriptions. The mean description length was 125 words (Ukrainian), with substantial variation reflecting differences in documentation detail. Sessions with empty or minimal descriptions (N=3) were excluded from text analysis.

Analysis of presenter frequency revealed individuals appearing in two or more sessions (Table 4). The most frequent presenter, Serhii Korniienko, moderated 14 sessions, establishing him as a central figure in the club's activities. Historical events discussed included the Third Wave Experiment, Ukrainian Revolution, Battle of Kruty, and the Yugoslav Wars (Table 5).

Figure 9 illustrates the distribution of entity types.

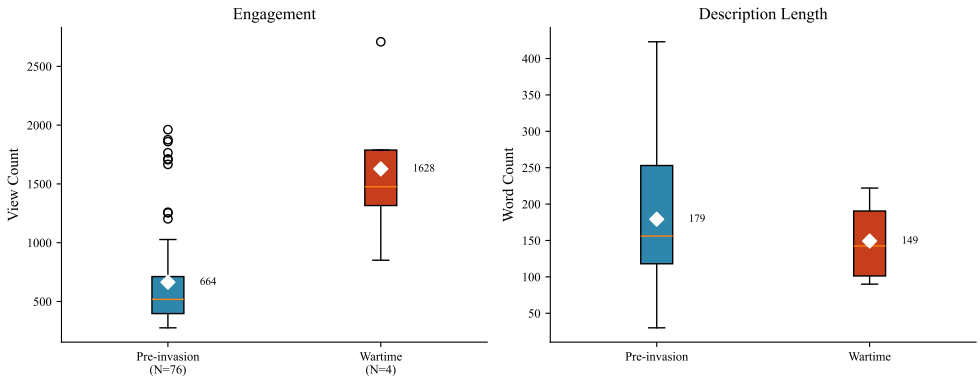


Figure 7: Comparison of session characteristics between pre-invasion (October 2016–February 2022) and wartime periods (February 2022–January 2025). Pre-invasion sessions (N=62) showed higher mean view counts (831.3 vs 677.5), while wartime sessions demonstrated longer average description lengths, suggesting more detailed documentation under constrained circumstances.

Table 3: Named entity extraction results

Entity type	**Count**
PERSON	445
LOCATION	134
ORGANIZATION	121
EVENT	7
WORK	7
Total	**714**

4.10 Extended content from Studia Hierodotica

The *Studia Hierodotica* publication provides extended presentation materials that complement the web-scraped summaries. Analysis of these full-text articles reveals the depth of scholarly engagement in selected sessions. Table 6 summarises the articles included in this publication.

The article "Vampires: Evolution of the Image in Mass Culture" by Dmytro Chyzhyk (Agora #49) exemplifies the club's engagement with popular culture through a historical lens. The 7-page study traces vampire imagery from Eastern European folklore through literary tradition (Bram Stoker's *Dracula*) to contemporary film and television, explicitly noting its origin as a "Halloween-themed Agora" presentation. This framing illustrates how the club integrates seasonal cultural observances with historical analysis.

The Inquisition article (Agora #50) represents historiographical revisionism, challenging what the author terms the "Black Legend" (leyenda negra) of Spanish history.

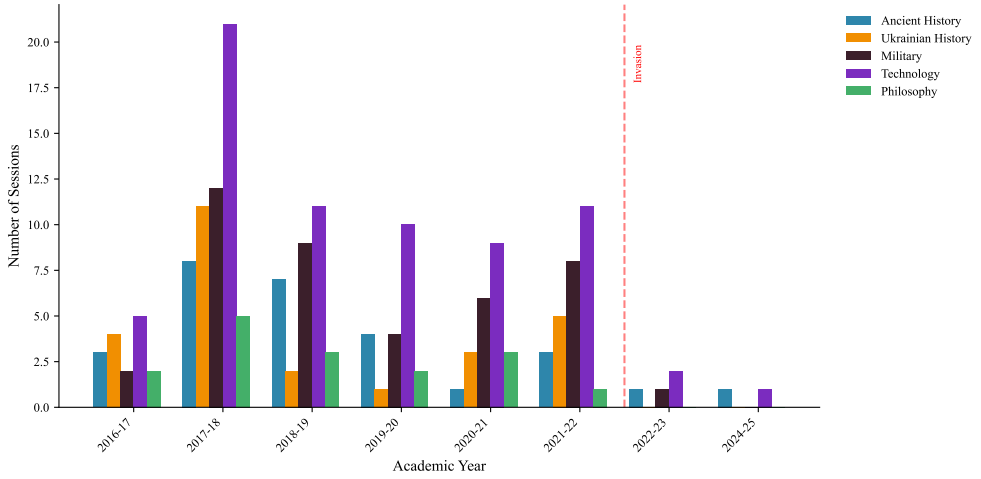


Figure 8: Evolution of discussion topics across academic years (2016–2025). Technology-focused sessions increased notably in 2023–2025, coinciding with broader societal engagement with artificial intelligence. Military and Ukrainian history themes show periodic emphasis corresponding to anniversary commemorations and current events.

Table 4: Most frequent presenters (2+ sessions)

Name in Ukrainian	Name in English	Sessions
Сергій Корнієнко	Serhii Korniienko	14
Дмитро Чижик	Dmytro Chyzhyk	6
Аліна Латипова	Alina Latypova	5
Діана Назаренко	Diana Nazarenko	5
Анастасія Шевченко	Anastasiia Shevchenko	5
Антон Доренко	Anton Dorenko	4
Михайло Сазонов	Mykhailo Sazonov	4
Олександр Холодняк	Oleksandr Kholodnyak	4
Олексій Гетманов	Oleksiy Hetmanov	3
Михайло Мельник	Mykhailo Melnyk	3

Such critical engagement with received historical narratives demonstrates the club's willingness to address controversial topics and encourage evidence-based reassessment of popular historical assumptions.

The "Lipstick: History Through the Ages" article (Agora #48) exemplifies interdisciplinary micro-history, tracing a single cultural artefact – cosmetics – from ancient Mesopotamia and Egypt through to Hollywood's golden age. This approach mirrors the Annales school's focus on everyday life and material culture, suggesting methodological sophistication in the club's historical practice.

Table 5: Historical events discussed

Event (Ukrainian)	Event (English)	Sessions
Скіфська доба	Scythian Era	1
Українська революція	Ukrainian Revolution	1
Реформація	Reformation	1
Інквізиція	Inquisition	1
Голокост	Holocaust	1
Югославські війни	Yugoslav Wars	1
Бій під Крутами	Battle of Kruty	1

Table 6: Articles published in *Studia Hierodotica* (Pechenina & Yashin, 2020)

Title (English)	Pages	Topic
Vampires: Evolution of the Image in Mass Culture	3–9	Halloween special on vampire folklore and media representation
Inquisition: Debunking the "Black Legend"	10–18	Historical reassessment of Spanish Inquisition
Formation of Stereotypes Through Art	19–22	Art's role in creating cultural stereotypes
Role of Women in USSR GRU Structure	23–25	Gender and Soviet military intelligence
Lipstick: History Through the Ages	26–30	Cultural history of cosmetics
History of the Beginning of the Space Race	31–46	Cold War space competition origins

The longest article (16 pages) addresses the Space Race, examining German rocket technology's transfer to both superpowers after World War II and the subsequent competition that culminated in the Moon landings. This topic reflects the club's engagement with Cold War history and the history of science and technology.

4.11 Video transcript analysis

Analysis of seven video recordings from sessions #75, #76, #78, and #80 (totalling approximately 170 minutes of Ukrainian-language content) provided rich discourse data beyond the web-based summaries. Whisper transcription yielded over 1,600 lines of spoken text, enabling named entity recognition on conversational academic discourse.

NER analysis of transcript content identified 204 additional entity mentions not present in the web descriptions, including informal references to presenters, discussion participants, and historical figures mentioned during Q&A exchanges. The transcripts reveal the interactive nature of Agora sessions, with frequent back-and-forth between presenters and audience members.

Session #75, focused on Lesya Ukrainka's *Lisova Pisnia* ("Forest Song"), demonstrated the club's engagement with Ukrainian literary heritage. A participant's comment captures the interdisciplinary spirit:

> "Тобто, в цілому все те, що вам подобається, те, що ви б хотіли обговорити. [...] Це може бути історія помади, історія коміксів."

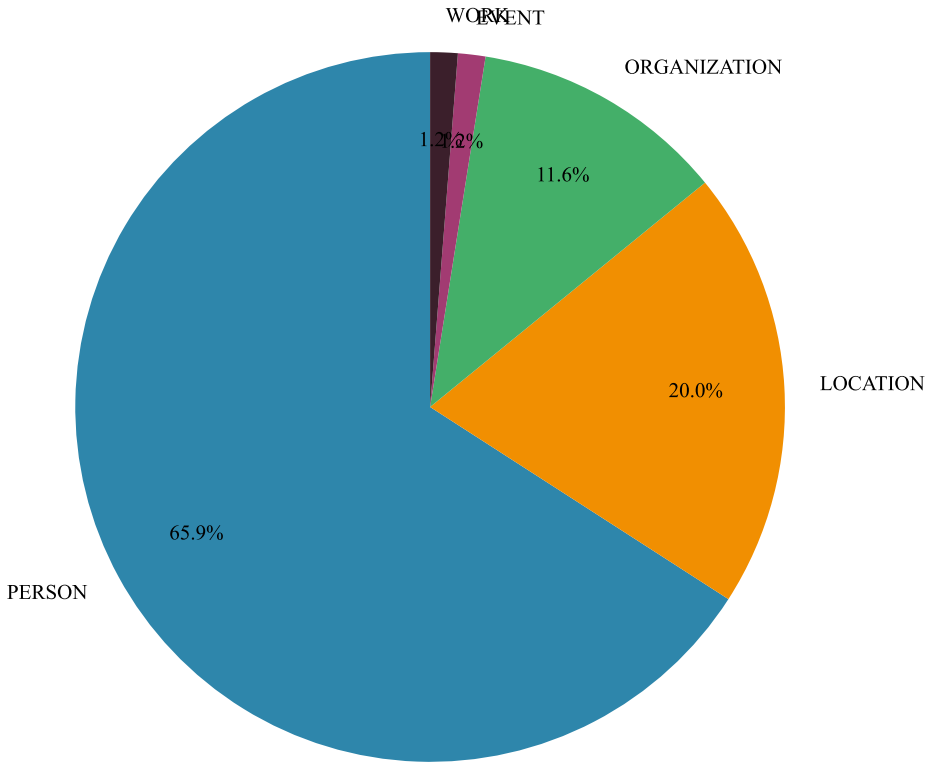


Figure 9: Distribution of named entity types extracted from the Agora corpus. Person entities dominate, reflecting the discussion format's emphasis on individual presenters and historical figures.

> ["Essentially, everything that interests you, what you would like to discuss. [...] It could be the history of lipstick, the history of comics."]

Session #80, addressing artificial intelligence in 2025, revealed discussions of ChatGPT, Claude, and other AI tools in historical research contexts. The transcript shows participants grappling with methodological questions about AI-assisted scholarship – a topic of direct relevance to the present study's methods.

The video transcripts complement the web-based corpus by preserving spontaneous discussion elements typically lost in written summaries, including questions from younger students, clarifications by moderators, and moments of collaborative sense-making characteristic of academic discourse communities.

5 Discussion

5.1 Methodological contributions

This study demonstrates the feasibility of applying digital humanities methods to preserve and analyse regional academic heritage from web-based sources. The combination of web scraping, OCR, machine translation, and NLP techniques provides a reproducible pipeline for similar preservation efforts at other institutions. Our enhanced NER pipeline, combining spaCy's Ukrainian language model with morphological normalization, entity deduplication, and OCR-enriched session data from 38 sessions with unique image content, extracted 714 unique named entities across five categories (Table 3). Key methodological decisions included:

1. Two-phase scraping to handle Joomla pagination
2. LLM-based translation for consistent quality across the corpus
3. Hybrid NER combining spaCy's Ukrainian model with morphological normalization
4. Pattern-based extraction for historical events and cultural works
5. Network analysis to identify participant communities

Several methodological limitations merit acknowledgment. Topic modelling using BERTopic with multilingual sentence embeddings failed to identify coherent clusters, likely due to the relatively small corpus size (80 documents) and the diversity of topics covered within a single discussion series. Future work might address this through document augmentation techniques, hierarchical clustering approaches, or analysis of larger corpora from multiple institutions.

While our spaCy-based NER pipeline achieved reasonable precision for person names, the morphological complexity of Ukrainian posed challenges for entity normalization. The inflectional nature of the language means that a single person may appear in multiple grammatical cases across the corpus, requiring careful deduplication. Future work might benefit from transformer-based NER models fine-tuned specifically for Ukrainian historical and academic texts.

5.2 Insights into Ukrainian academic culture

The Agora Discussion Club corpus provides a window into intellectual life at a regional Ukrainian pedagogical university over nearly a decade. Five patterns emerge from the analysis:

1. *Thematic breadth:* The range of topics – spanning ancient civilisations to artificial intelligence – reflects both the historical training of participants and their engagement with contemporary issues. This breadth suggests that the discussion club serves not merely as an extension of coursework but as a venue for intellectual exploration beyond disciplinary boundaries.
2. *Institutional resilience:* The persistence of activity through the COVID-19 pandemic (with several "Zoom-Agora" sessions documented in 2020–2021) and the ongoing conflict demonstrates institutional resilience. The club adapted its format to remote delivery when necessary while maintaining its core mission of fostering academic discussion.

3. *Technological engagement:* The shift toward technology-focused topics in recent sessions (sessions #79–80 on ChatGPT and AI) suggests adaptation to broader societal changes while maintaining historical perspectives. Session #79, specifically addressing "What can historians ask ChatGPT", exemplifies this integration of disciplinary expertise with technological developments.
4. *Publication culture:* The production of *Studia Hierodotica* demonstrates the club's ambition to move beyond ephemeral discussion toward lasting scholarly output. The publication's invocation of Herodotus – "so that nothing done by men may be forgotten over time" – explicitly frames the club's activities within the historiographical tradition. This self-conscious positioning suggests participants view their work as contributing to historical scholarship, not merely student socialisation.
5. *Student development:* The network analysis reveals repeated participation by certain students across multiple academic years, suggesting that the club serves as a venue for developing presentation and discussion skills that complement formal education.

5.3 Implications for digital preservation

Regional academic institutions frequently lack resources for systematic digital preservation, yet their intellectual contributions warrant documentation. This study suggests that computational methods can partially address this gap, enabling retrospective preservation of web-published materials. However, several caveats apply:

- Reliance on publicly accessible content means that materials behind authentication barriers or posted to social media platforms with restrictive access policies may remain unpreservable through these methods
- Video content hosted on third-party platforms (Facebook, YouTube) presents particular challenges due to platform restrictions and content mobility
- The ephemerality of web content necessitates timely preservation efforts before link rot or institutional changes result in permanent loss

The release of the complete corpus under CC BY 4.0 represents a contribution to open scholarship and enables both verification of our findings and extension to new research questions.

5.4 Limitations

Several limitations constrain the present analysis:

1. The corpus reflects only officially documented sessions; informal discussions or sessions that were not announced on the website are not captured. Additionally, the depth of documentation varies across sessions, with some containing detailed descriptions and others limited to titles and images.
2. While most video recordings from Facebook and YouTube could not be systematically retrieved due to platform access restrictions, we obtained and transcribed seven videos from four sessions (sessions #75, #76, #78, and #80) totalling approximately 170 minutes. These were processed using OpenAI Whisper (medium model) for Ukrainian speech recognition, yielding 1,685 lines of transcript text. The

transcripts provide richer discourse data than textual summaries but represent only a subset of recorded sessions.
3. The machine translation pipeline, while producing readable English text, may introduce subtle errors or lose nuances present in the Ukrainian originals. Bilingual researchers should consult the Ukrainian source text for detailed analysis.
4. The absence of interview or survey data from participants means that analysis relies entirely on textual traces rather than first-person accounts of the club's significance to its members.
5. As a single-institution case study, findings may not generalise to discussion clubs at other institutions or in other national contexts.

6 Conclusion

This study has presented a comprehensive computational analysis of the Agora Discussion Club at Kryvyi Rih State Pedagogical University, demonstrating methods for preserving and analysing regional academic heritage through digital humanities approaches. The resulting bilingual corpus of 80 discussion sessions, spanning from October 2016 to January 2025, provides a foundation for future research on Ukrainian academic discourse and digital preservation methodologies.

Key findings include: (1) the remarkable thematic diversity of discussions, encompassing ancient history, philosophy, art, and contemporary technology; (2) the persistence of academic activity through significant external disruptions including the COVID-19 pandemic and Russia's ongoing invasion; (3) engagement patterns correlating with topic novelty and contemporary relevance; and (4) the club's role in fostering sustained intellectual communities among students and faculty as revealed through network analysis.

As universities increasingly rely on web platforms to document their activities, computational preservation methods become essential for maintaining institutional memory. The approaches developed here – combining web scraping, OCR, machine translation, and network analysis – may be adapted for similar efforts at other institutions, particularly those operating in contexts where traditional archiving infrastructure is limited or threatened.

The complete corpus is released under CC BY 4.0 and deposited in Zenodo, supporting replication of this research and enabling future studies on Ukrainian academic culture, digital preservation methodology, and the documentation of intellectual life under conditions of conflict.

Data availability

All data and code supporting this study are publicly available:

Primary dataset (Zenodo): https://doi.org/10.5281/zenodo.18329395

This repository contains:

- Raw and processed session metadata (80 sessions, CSV and JSON formats)
- Bilingual corpus with Ukrainian originals and English translations
- Named entity annotations (714 entities across 5 categories)

- Participant co-occurrence network (edge lists and centrality metrics)
- Theme classifications and sentiment scores
- OCR-extracted text from 428 images across 38 sessions
- Video transcripts from 7 recordings (Whisper medium model outputs)
- Full codebook with variable definitions and coding schemes

Analysis code (GitHub): https://github.com/ssemerikov/agora-corpus
This repository contains Python scripts for:

- Web scraping (Selenium, BeautifulSoup)
- OCR processing (Tesseract with Ukrainian language model)
- Machine translation (Claude API)
- NLP pipeline (spaCy uk_core_news_lg, pymorphy3)
- Network analysis (NetworkX)
- Video transcription (OpenAI Whisper)
- Figure generation (matplotlib, seaborn)

Both repositories are released under CC BY 4.0 licence.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors. All data were collected from publicly accessible webpages documenting institutional academic events. No personal data were collected beyond names of individuals participating in public academic presentations, which were already publicly documented on the institutional website.

Informed consent

This article does not contain any studies with human participants performed by any of the authors. The study analyses publicly available institutional records documenting public academic events rather than private communications or personal data.

References

Barcelona, L., & Dempsey, P. (2024). Student Life, Diverse Institutional Histories, and Archival Policy: Inclusion of Student Life Archival Material in University Archives' Collecting Statements. *College & Research Libraries*, *85*(6), 930, https://doi.org/10.5860/crl.85.6.930

Becker, J. (2017, December). Bringing Student Voices into the University Archives: A Student Organization Documentation Initiative Case Study. *In the Library with*

the Lead Pipe, , Retrieved from http://www.inthelibrarywiththeleadpipe.org/2017/bringing-student-voices/

Burnhill, P., Mewissen, M., Wincewicz, R. (2015, Jul). Reference rot in scholarly statement: threat and remedy. *Insights: the UKSG journal*, *28*(2), 55-61, https://doi.org/10.1629/uksg.237

Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., ... Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185, https://doi.org/10.1126/science.aao0185

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, *abs/2203.05794*, , https://doi.org/10.48550/ARXIV.2203.05794 2203.05794

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H., Ginis, V. (2025). Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *CoRR*, *abs/2501.11623*, , https://doi.org/10.48550/ARXIV.2501.11623 2501.11623

Kramarenko, T.H., & Semerikov, S.O. (2025, sep). A century of mathematics education at Kryvyi Rih State Pedagogical University: Honoring the past, shaping the future. *Journal of Physics: Conference Series*, *3105*(1), 012006, https://doi.org/10.1088/1742-6596/3105/1/012006

Liashenko, R.O., & Semerikov, S.O. (2024). Training Specialised Chatbots on Ukrainian Scientific Text Corpora Using Transfer Learning. *2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT)* (p. 1-4). Retrieved from https://doi.org/10.1109/CSIT65290.2024.10982573

Pechenina, N., & Yashin, V. (Eds.). (2020). *Studia Hierodotica: Materialy deiakykh tsikavykh zasidan Dyskusiinoho klubu «AHORA» [Studia Hierodotica: Materials from selected sessions of the Agora Discussion Club]*. Kryvyi Rih: History Department, KSPU.

Risam, R. (2019). *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press.

Slobodianiuk, A.V., & Semerikov, S.O. (2024). Advances in neural text generation: A systematic review (2022-2024). S.O. Semerikov & A.M. Striuk (Eds.), *Proceedings of the 7th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2024), Virtual Event, Kryvyi Rih, Ukraine,*

December 27, 2024 (Vol. 3917, pp. 332–361). CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-3917/paper59.pdf