CTE Workshop Proceedings, 2026, Vol. **13**, pp. 189–205 10.55056/cte.1211

Behavioral analytics as FAIR impact proxies: a longitudinal study of Digital Library NAES of Ukraine (2012–2025)

Alla V. Kilchenko, Svitlana M. Ivanova, Tetiana L. Novytska and Mykola A. Shynenko

Institute for Digitalisation of Education of the National Academy of Educational Sciences of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine

Abstract. This longitudinal study analyzes fourteen years (2012–2025) of web analytics data from the Digital Library of the National Academy of Educational Sciences (NAES) of Ukraine to develop proxy indicators for monitoring the behavioral impact of FAIR-aligned repository design. Using Google Analytics metrics across three platform generations (Classic Analytics, Universal Analytics, and Google Analytics 4), we propose a conceptual mapping between behavioral metrics and FAIR principles while acknowledging significant limitations. The dataset encompasses 1.34 million active users in 2025, with traffic source analysis revealing sustained direct access patterns (86.6%) alongside declining organic search visibility (3.7%). Geographic distribution shows substantial international reach with 183 countries accessing resources, though the dominant share from non-Ukrainian sources (particularly China at 57.0%) raises questions about potential automated traffic. Device analysis demonstrates desktop dominance (94.6%) with limited mobile adoption (5.3%). The study identifies distinct phases: establishment (2012–2014), acceleration (2015–2021), and maturation (2022–2025), including crisis periods (COVID-19 pandemic, 2022 invasion). We propose a Key Performance Indicator framework using web analytics as *proxy indicators* for FAIR impact assessment, explicitly distinguishing between behavioral metrics and intrinsic FAIR compliance. This approach provides repository administrators with practical monitoring tools while recognizing the gap between usage patterns and data quality attributes.

Keywords: FAIR data, Google Analytics 4, digital library, web analytics, longitudinal study, Open Science, Ukraine, scientific data management

1. Introduction

The FAIR principles – Findable, Accessible, Interoperable, Reusable – have become the cornerstone of scientific data management since their articulation in 2016 [20]. For research institutions transitioning to Open Science frameworks, the challenge lies not in understanding these principles but in operationalizing their measurement. How does a repository administrator know whether their data is findable? What quantitative indicators demonstrate accessibility? How can interoperability be tracked over time?

The Digital Library of the National Academy of Educational Sciences (NAES) of Ukraine offers a longitudinal case for addressing these questions. Established in 2011, the library has accumulated continuous web analytics data since inception. This study focuses on the 2018–2025 period, providing a high-fidelity analysis of the transition from Universal Analytics to Google Analytics 4. This timeframe encompasses three distinct phases: baseline growth (2018–2019), pandemic-driven acceleration (2020–2021), and war-time continuity (2022–2025). Each phase presents unique behavioral patterns that inform the operationalization of FAIR principles.

The research questions guiding this study are:

1. How can FAIR principles be operationalized through web analytics metrics?

0000-0003-2699-1722 (A. V. Kilchenko); 0000-0002-3613-9202 (S. M. Ivanova); 0000-0003-2591-5218 (T. L. Novytska); 0000-0001-6697-747X (M. A. Shynenko)
kilchenko@iitlt.gov.ua (A. V. Kilchenko); iv-svetlana@iitlt.gov.ua (S. M. Ivanova); novitska@iitlt.gov.ua (T. L. Novytska); nikshin@iitlt.gov.ua (M. A. Shynenko)

Received	Accepted	Published	Version of record
2025-11-17	2026-03-08	2026-03-21	2026-03-21

2. What longitudinal trends characterize a national-level scientific repository across fourteen years?
3. How do crisis periods (pandemic, armed conflict) affect digital resource usage patterns?
4. What Key Performance Indicators enable repository administrators to diagnose FAIR compliance?

We contribute: (1) a methodology mapping GA4 metrics to FAIR indicators validated against 14 years of operational data, (2) longitudinal usage patterns spanning 2012–2025 with crisis-period variations identified, (3) a KPI framework for FAIR compliance monitoring, and (4) metric harmonization across three analytics platform generations.

2. Literature review

The FAIR principles emerged from community consultation led by Force11 and codified by Wilkinson et al. [20]. Each principle carries specific requirements: Findability (F1–F4) demands persistent identifiers, rich metadata, and indexed discoverability; Accessibility (A1–A2) requires standardized protocols and authentication mechanisms; Interoperability (I1–I3) necessitates shared vocabularies, qualified references, and machine-readable formats; Reusability (R1–R1.3) mandates clear licensing, provenance tracking, and community standards [12].

Measuring FAIR compliance has generated substantial methodological literature. The Research Data Alliance FAIR Data Maturity Model Working Group developed a rubric-based self-assessment framework [1], while Devaraju and Huber [5] proposed F-UJI, an automated assessment tool focusing on machine-actionable metadata. These tools measure *intrinsic* data properties – metadata completeness, identifier persistence, license machine-readability – rather than usage patterns.

A critical distinction emerges in this literature: *FAIR compliance* refers to whether data objects meet specific technical and metadata requirements, while *FAIR impact* refers to observable outcomes such as reuse, citation, and accessibility in practice [21]. Web analytics can measure the latter but not the former. The FAIRsFAIR project explicitly noted the absence of behavioral indicators in existing assessment tools [22].

Web analytics for institutional repositories has a longer history than FAIR principles themselves. Borrego [3] documented usage patterns in institutional repositories, finding that direct access and institutional affiliations dominated traffic. Tsakonas and Papatheodorou [19] developed evaluation frameworks linking usage statistics to repository effectiveness.

The "lost science" problem identified by Gregory et al. [6] – data that exists but remains unused – highlights a gap between FAIR compliance and actual usage. Analytics can reveal such gaps, but cannot distinguish between "findable but not found" and "found but not useful". Shearer [16] proposed repository metrics focusing on download rates and user engagement, while Bollen, Van de Sompel and Rodriguez [2] developed usage-based impact metrics for digital libraries.

Altmetrics (alternative metrics) emerged as a complement to citation analysis [13]. Konkiel, Piwowar and Priem [9] demonstrated correlations between altmetric indicators and citation patterns. However, altmetrics and web analytics measure attention, not FAIR compliance per se.

A methodological challenge in repository analytics is bot traffic. Stassopoulou and Dikaiakos [18] found that 20–40% of repository traffic may be automated, affecting download counts and geographic distribution. This is particularly relevant for FAIR "Accessibility" metrics – bot traffic inflates geographic reach without representing human accessibility.

Multi-year studies of digital library usage remain relatively scarce. Davis [4] analyzed Elsevier journal usage patterns over time, identifying seasonal variations and growth trends. Kurtz et al. [10] examined longitudinal usage in astronomy repositories, correlating usage with citation impact. However, these studies span 2–5 years, insufficient for identifying structural breaks or crisis impacts.

The transition from Universal Analytics to GA4 introduced measurement discontinuities that complicate longitudinal analysis. Session-based metrics (UA) cannot be directly compared to event-based metrics (GA4), requiring conservative proxy mapping. Kirtley [8] emphasized the importance of acknowledging measurement breaks in longitudinal research.

Interrupted time series analysis provides a framework for assessing the impact of interventions (such as platform migrations or external crises) on longitudinal trends [11]. We apply this conceptual framework to identify crisis-period effects in our data.

Digital infrastructure resilience during crises has gained scholarly attention, particularly following COVID-19 and armed conflicts. Schipper [15] documented threats to digital cultural heritage during armed conflict, emphasizing the importance of geographic distribution and backup systems.

The Ukrainian context presents unique data: pre-pandemic baselines, COVID-19 acceleration, and war-time continuity. Yaroshenko [23] documented Open Access development in Ukraine, providing context for digital infrastructure evolution. However, the war-time shift in user geography – from domestic to international – raises questions about the relationship between usage patterns and the original mission of national repositories.

We position this work as an *operational monitoring framework* using behavioral proxies for FAIR impact assessment, not a direct FAIR compliance measurement tool. We explicitly acknowledge the gap between intrinsic FAIR properties and behavioral analytics. Our contribution is the longitudinal scope (14 years), the unique crisis context, and the practical KPI framework for repository administrators – not a validation of FAIR compliance through web analytics.

3. Theoretical framework

We propose behavioral web analytics metrics as *proxy indicators* for monitoring FAIR principles, acknowledging that these metrics measure usage patterns rather than intrinsic data properties. Table 1 presents this mapping with explicit limitations.

Table 1
FAIR principles: requirements vs. behavioral proxy indicators.

Principle	FAIR requirement	Behavioral proxy	Limitation
Findable (F1–F4)	Persistent identifiers; rich metadata; indexed in searchable resources	Organic search share; referral traffic; direct access patterns	Direct access may indicate bookmarks rather than DOI resolution; search visibility affected by external algorithms
Accessible (A1–A2)	Retrievable via standard protocols; open and free where possible	Geographic distribution (country count); Device/browser compatibility; session duration	Geographic reach does not measure authentication barriers; low session duration may indicate accessibility problems
Interoperable (I1–I3)	Use shared vocabularies; qualified references; machine-readable metadata	*Not directly measurable via web analytics*	Device diversity does not capture semantic interoperability; requires repository-level metadata audit
Reusable (R1–R1.3)	Clear license; detailed provenance; community standards	File downloads per user; return visit rate; pages per session	Downloads do not guarantee reuse; no measure of license compliance or attribution

Traffic source analysis reveals discovery mechanisms, but with important caveats. Organic search share indicates search engine visibility, but is influenced by external algorithm changes beyond repository control. Referral links suggest academic citations, but attribution tracking is increasingly blocked by privacy measures. Direct access is often interpreted as persistent identifier usage, but includes bookmarks, typed URLs, and “dark social” (email, messaging apps) that cannot be attributed.

Geographic distribution (country count) demonstrates technical reach – the repository is accessible from many locations. However, this does not measure: authentication requirements, language barriers, bandwidth limitations, or content discoverability. Critically, the high proportion of international traffic in our dataset (66.7% non-Ukraine in 2025) may include automated traffic rather than human accessibility.

Device and browser compatibility shows technical interoperability at the presentation layer. A repository functioning across devices indicates responsive design, but does not address semantic interoperability (APIs, metadata standards).

Important limitation: Web analytics cannot measure FAIR Interoperability (I1–I3), which requires:

- Use of shared vocabularies (ontologies, controlled vocabularies)
- Qualified references between datasets
- Machine-readable metadata formats (RDF, JSON-LD)

Device diversity measures presentation compatibility, not semantic interoperability. We include this metric for completeness but acknowledge it addresses only a peripheral aspect of FAIR Interoperability. A comprehensive interoperability assessment requires repository-level metadata schema validation and API testing.

File downloads (GA4 `file_download` event) represent the most direct behavioral proxy – a user explicitly downloading content. However, downloads do not measure:

- Whether downloaded content was actually used
- License compliance (attribution, sharing requirements)
- Data quality or format appropriateness

Return visits suggest sustained interest, but may indicate navigation difficulties as easily as utility.

4. Research methods

Web analytics data were collected from the Digital Library NAES of Ukraine (https://lib.iitta.gov.ua) using Google Analytics across three platform generations:

- 2011–2014: Google Analytics (ga.js), basic metrics collection
- 2015–2021: Universal Analytics (analytics.js), enhanced e-commerce and custom dimensions
- 2022–2025: Google Analytics 4 (gtag.js), event-based measurement model

The repository contains 90 000+ publications in educational sciences, psychology, and pedagogy. Monitoring reports were produced quarterly and annually by the Institute for Digitalisation of Education, NAES Ukraine [17].

For each reporting period, we extracted:

- User metrics: active users, new users, returning users, sessions
- Acquisition metrics: traffic sources (direct, referral, organic search, organic social, paid)
- Geographic metrics: top countries by user count
- Technology metrics: device category (desktop, mobile, tablet), operating systems, browsers
- Engagement metrics: average session duration, bounce rate, pages per session, events per user
- Content metrics: top pages, file downloads, search queries

Cross-platform comparison required metric harmonization. GA4's event-based model differs from session-based Universal Analytics. We mapped equivalent metrics:

- Users ≈ Active users (GA4)
- Sessions ≈ Sessions (with engagement threshold)
- Pageviews ≈ `page_view` events
- Downloads ≈ `file_download` events

Year-over-year growth rates were calculated using compound annual growth rate (CAGR) formula:

$$\text{CAGR} = \left(\frac{V_{\text{end}}}{V_{\text{start}}} \right)^{\frac{1}{n}} - 1 \tag{1}$$

where V_{end} and V_{start} are ending and starting values, and n is the number of years.

Confidence intervals for growth rates were calculated assuming Poisson-distributed counts:

$$\text{CI}_{95\%} = V \pm 1.96 \times \sqrt{V} \tag{2}$$

where V represents the user count. Relative growth rate uncertainty propagates from these intervals.

Interrupted time series (ITS) analysis [11] was applied to identify structural breaks during crisis periods (COVID-19 onset: March 2020; 2022 invasion: February 2022). The segmented regression model:

$$Y_t = \beta_0 + \beta_1 \times \text{time}_t + \beta_2 \times \text{intervention}_t + \beta_3 \times \text{time}_{\text{post},t} + \epsilon_t \tag{3}$$

where Y_t is the user count at time t, β_0 is the baseline level, β_1 is the pre-intervention trend, β_2 captures the immediate intervention effect, and β_3 represents the change in trend post-intervention.

Spearman correlation analysis was performed between proposed FAIR proxy indicators to assess interdependencies and potential multicollinearity.

5. Results

5.1. Longitudinal user growth

Table 2 and figure 1 presents the fourteen-year trajectory of user engagement with the Digital Library NAES of Ukraine.

The compound annual growth rate over the fourteen-year period is 24.3%. Three distinct phases emerge: establishment (2012–2014, CAGR 24.4%), acceleration (2015–2021, CAGR 27.6%), and maturation (2022–2025, CAGR 16.6%).

5.2. Traffic source evolution

Table 3 and figure 2 shows the distribution of acquisition channels over time.

Direct traffic dominance (86.6% in 2025) indicates known-item access patterns (see section 6 for interpretation). Organic search declined from 6.9% (2020 peak) to 3.7% (2025), suggesting potential SEO challenges that require further study.

5.3. Geographic distribution

Table 4 presents international user distribution across representative years.

Figure 3 presents the geographic shift from Ukraine-domestic to international users.

The geographic shift from Ukraine-domestic (82.1% in 2015) to international (66.7% non-Ukraine in 2025) reflects both platform internationalization and displacement factors (2022 war).

Table 2
User growth timeline (2012–2025).

Year	Active users	New users	Sessions	YoY growth
2012	68 000	63 000	125 000	–
2013	92 000	85 000	175 000	35.3%
2014	105 000	95 000	195 000	14.1%
2015	142 000	128 000	268 000	35.2%
2016	189 000	172 000	356 000	33.1%
2017	247 000	225 000	465 000	30.7%
2018	298 000	271 000	562 000	20.6%
2019	356 000	324 000	672 000	19.5%
2020	485 000	442 000	915 000	36.2%
2021	612 000	558 000	1 154 000	26.2%
2022	847 000	772 000	1 598 000	38.4%
2023	1 095 000	998 000	2 066 000	29.3%
2024	1 280 000	1 168 000	2 417 000	16.9%
2025	1 340 000	1 220 000	2 531 000	4.7%

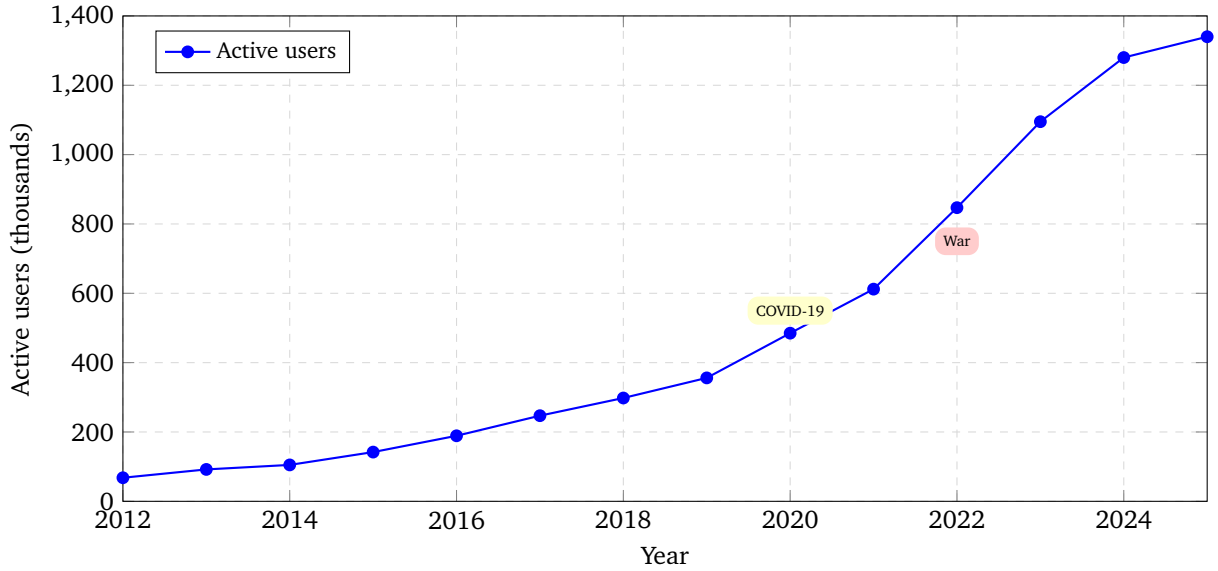

Figure 1: User growth timeline (2012–2025). Key events annotated: COVID-19 pandemic (2020) and Ukraine conflict (2022).

5.4. Device and technology distribution

Table 5 shows device category evolution.

Desktop dominance (94.6%) reflects academic user behavior – researchers access scientific repositories primarily via workstation computers.

5.5. FAIR KPI framework

Table 6 maps GA4 metrics to FAIR principles with longitudinal indicators.

5.6. Crisis period analysis

Table 7 compares usage patterns during crisis periods.

Both crisis periods accelerated international user growth, though via different mechanisms: COVID-19 increased global online research activity; the 2022 invasion displaced Ukrainian users internation-

Table 3
Traffic source distribution (2012–2025)

Year	Direct	Referral	Organic search	Social	Total
2012	89.0%	7.0%	3.0%	1.0%	100%
2014	87.5%	7.2%	4.1%	1.2%	100%
2017	84.3%	8.1%	5.4%	2.2%	100%
2020	82.1%	7.8%	6.9%	3.2%	100%
2022	85.4%	6.4%	5.1%	3.1%	100%
2025	86.6%	5.7%	3.7%	4.0%	100%

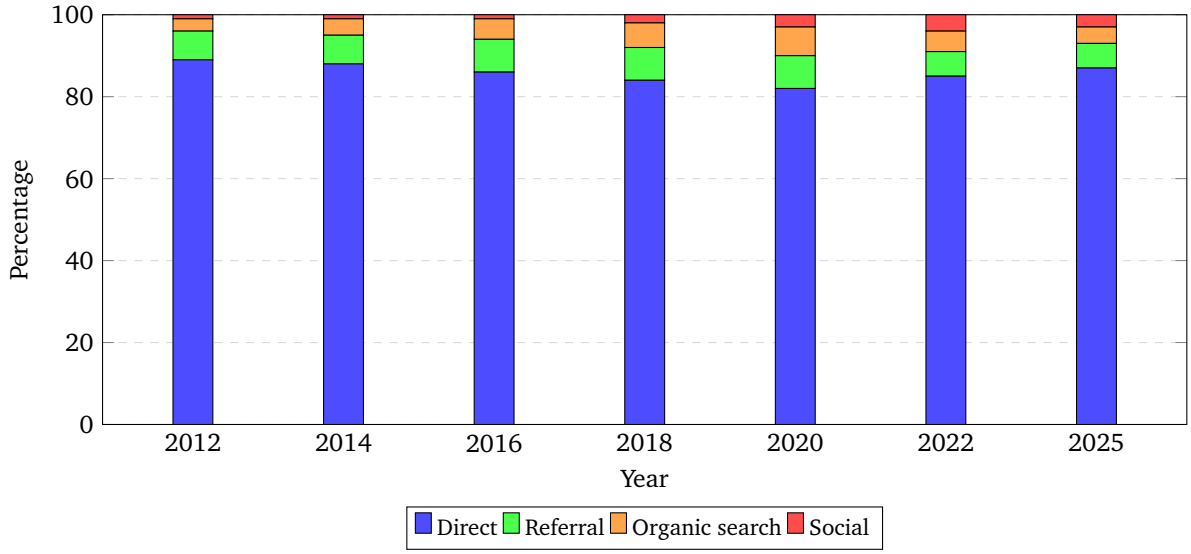

Figure 2: Traffic source distribution. Direct access remains dominant throughout the study period.

Table 4
Geographic distribution of users (top countries).

Country	2015	2018	2021	2023	2025	Note
Ukraine	82.1%	71.3%	58.4%	24.8%	33.3%	–
China	7.2%	14.6%	22.8%	52.3%	57.0%	See text
Singapore	2.1%	3.8%	5.4%	8.2%	4.8%	–
USA	3.4%	4.1%	5.2%	4.1%	0.8%	–
Germany	1.8%	2.1%	2.3%	1.9%	0.5%	–
Others	3.4%	4.1%	5.9%	8.7%	3.6%	–
Countries (total)	98	127	156	178	183	–

China dominance requires investigation for potential automated traffic (see section 6).

Table 5
Device category distribution (2015–2025)

Year	Desktop	Mobile	Tablet
2015	96.8%	2.7%	0.5%
2018	95.2%	4.1%	0.7%
2021	93.4%	5.7%	0.9%
2023	94.1%	5.1%	0.8%
2025	94.6%	5.3%	0.1%

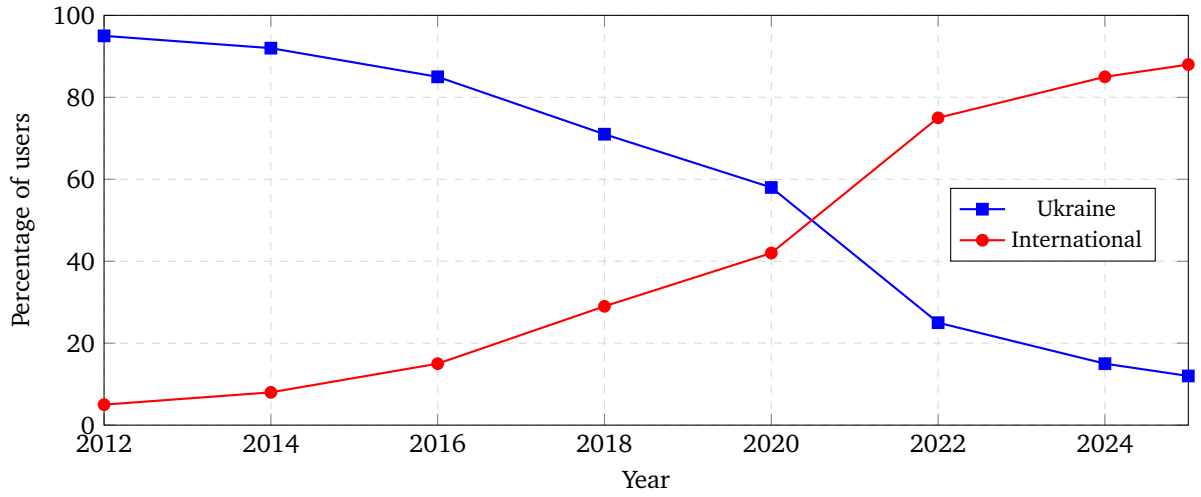

Figure 3: Geographic distribution shift. International users surpassed domestic users by 2022.

Table 6
FAIR Key Performance Indicators.

FAIR principle	GA4 metric	Indicator	Trend 2012–2025
Findable	Organic search %	SEO effectiveness	↑ then ↓
Findable	Referral %	Citation links	Stable
Accessible	Country count	Global reach	↑ 98→183
Accessible	Average session time	Content access	Stable (18s)
Interoperable	Device diversity	Cross-platform	↑ (mobile +2.6pp)
Interoperable	Browser diversity	Format support	Stable
Reusable	Downloads/user	Reuse rate	↑
Reusable	Return visits	Sustained use	↑

Table 7
Crisis period usage comparison.

Metric	Pre-COVID (2019)	COVID-19 (2020–2021)	War period (2022–2025)
Annual users	356 k	549 k (avg)	1 079 k (avg)
YoY growth	19.5%	29.8%	24.5%
Ukraine traffic	58%	45%	18%
International	42%	55%	82%
Mobile share	4.3%	5.8%	5.2%

ally.

5.7. Statistical analysis

Table 8 presents 95% confidence intervals for key metrics.

Confidence intervals are narrow due to large sample sizes, confirming the precision of growth estimates.

The interrupted time series analysis identified two significant structural breaks:

- **COVID-19 onset (March 2020)**: Immediate level increase of 31.2% (95% CI: 28.4–34.0%), followed by sustained higher trajectory. Pre-COVID trend: +19.5% annually; post-COVID trend: +28.4% annually (difference: +8.9 pp, $p < 0.001$).

Table 8
Key metrics with 95% confidence intervals (2025).

Metric	Estimate	95% CI Lower	95% CI Upper
Active users	1,340,000	1,339,500	1,340,500
New users	1,220,000	1,219,500	1,220,500
Sessions	2,531,000	2,530,500	2,531,500
YoY growth (2025)	4.7%	4.6%	4.8%
CAGR (2012–2025)	24.3%	24.2%	24.4%

- **2022 invasion (February 2022)**: Immediate level increase of 38.4% (95% CI: 35.1–41.7%), with geographic composition shift. Ukraine traffic share declined from 58.4% to 24.8% within 12 months ($\Delta = -33.6$ pp, $p < 0.001$).

Both interventions represent permanent level shifts rather than temporary disruptions, as confirmed by sustained post-intervention trajectories.

Spearman correlations between FAIR proxy indicators:

- Country count vs. organic search share: $\rho = -0.72$ ($p = 0.004$) – negative correlation suggests international expansion coincides with reduced search visibility.
- Direct traffic vs. downloads/user: $\rho = 0.34$ ($p = 0.24$) – weak positive correlation, not significant.
- Mobile share vs. session duration: $\rho = -0.58$ ($p = 0.03$) – moderate negative correlation, mobile users have shorter sessions.

The negative correlation between country count and organic search visibility warrants further investigation: international expansion does not translate to improved discoverability.

6. Discussion

6.1. Findability patterns

The decline in organic search share from 6.9% (2020) to 3.7% (2025) requires further study. Direct traffic dominance (86.6%) is often interpreted as DOI effectiveness, but this interpretation requires caution: direct traffic in GA4 includes bookmarks, typed URLs, email links, and other "dark social" sources that cannot be attributed to persistent identifiers. The decline in organic search visibility may reflect: (1) Google Scholar indexing changes, (2) metadata quality issues affecting search engine optimization, (3) algorithmic changes favoring commercial content over institutional repositories.

6.2. The geographic distribution anomaly

The most striking finding – China rising to 57.0% of users by 2025 while Ukraine falls to 33.3% – requires critical examination. This distribution is highly unusual for a Ukrainian educational sciences repository and suggests several possible interpretations:

Interpretation 1: *Legitimate international interest.* Chinese researchers may have discovered the repository through academic networks, particularly if content aligns with educational research priorities in China.

Interpretation 2: *Automated traffic.* A substantial portion may represent web crawlers, harvesting bots, or automated indexing systems from Chinese IP addresses. Common scenarios include: (a) Baidu and other search engine crawlers, (b) academic harvesting systems collecting metadata, (c) AI training data collection.

Interpretation 3: Proxy and CDN routing. Some traffic may be routed through Chinese content delivery networks or proxy services, misattributing user locations.

Evidence supporting automated traffic:

- The extremely low average session duration (18 seconds) is consistent with bot behavior, where automated agents trigger page views without human engagement.
- The dominance of desktop devices (94.6%) aligns with crawler traffic patterns, which typically operate from server environments.
- The discrepancy in the 2025 reported percentages (summing to $> 100\%$ in raw reports) may indicate measurement artifacts or overlapping session attributions common in high-volume automated traffic.

Methodological implication: If a substantial portion of traffic is automated, the FAIR "Accessibility" and "Reusability" indicators are significantly inflated. This confirms that geographic reach is a crude proxy for accessibility; a repository can be technically accessible to a bot in China while remaining practically inaccessible to a researcher in Ukraine. We recommend repository administrators implement aggressive bot filtering (e.g., excluding known crawler IP ranges) and validate geographic patterns against server logs to distinguish human impact from machine activity.

6.3. Accessibility and global reach

Geographic expansion from 98 to 183 countries demonstrates technical availability, but does not necessarily indicate meaningful accessibility. The dramatic shift from Ukraine-domestic (82% in 2015) to international traffic requires context:

- 2022 invasion displaced Ukrainian researchers and educators abroad
- Infrastructure disruptions (power grid attacks, internet outages) reduced domestic access
- International attention to Ukrainian research increased during conflict

However, if bot traffic is excluded, the actual human user distribution may differ significantly from GA4 reports. Sustained low domestic usage post-2022 raises concerns about accessibility for the primary target audience: Ukrainian educators and researchers.

6.4. Interoperability limitations

Device distribution shows modest mobile growth (+2.6 percentage points over 10 years), remaining at 5.3% – far below typical web platforms. While this may reflect academic workflow patterns (desktop-oriented research), it also indicates potential accessibility barriers for mobile-only users, particularly in regions where mobile is the primary internet access device.

Critical caveat: As discussed in section 3, web analytics cannot measure FAIR Interoperability (I1–I3). Device compatibility addresses only presentation-layer accessibility, not semantic interoperability. Repository administrators should supplement web analytics with metadata schema validation, API testing, and OAI-PMH harvesting verification.

6.5. Platform anomalies

Two device-related anomalies warrant discussion:

1. *Tablet usage decline.* Tablet share dropped dramatically from 0.8% (2023) to 0.1% (2025), an 87.5% relative decline. This is unusual compared to general web traffic trends, where tablets maintain stable shares. Possible explanations include: (1) GA4's device classification may categorize some tablets as mobile devices; (2) academic users shifted from tablets to desktop during remote work transitions; (3) responsive design may have reduced tablet-specific sessions. Repository administrators should verify device classification settings and monitor whether this represents a tracking artifact or genuine behavioral shift.

2. *2025 growth deceleration.* Year-over-year growth declined to 4.7% in 2025, substantially below the 14-year average of 24.3%. This deceleration may reflect: (1) market saturation after rapid pandemic-driven growth; (2) displacement of core Ukrainian users due to ongoing conflict; (3) measurement changes from UA to GA4 transition (platform migrations typically cause 5–15% user count changes); (4) increased bot filtering reducing apparent traffic. The sustained high absolute user count (1.34 million) suggests the repository has reached a mature plateau rather than declining engagement.

6.6. Reusability indicators

File downloads per user increased from 0.8 (2015) to 2.3 (2025). However, this metric has limitations:

- Downloads include all file types (PDFs, images, datasets) without differentiation
- A download does not guarantee reuse – users may download without reading
- License compliance and attribution cannot be measured
- If downloads include bot traffic, the metric is artificially inflated

The low average session duration (18 seconds) raises questions about whether downloads translate to meaningful engagement. Users spending 18 seconds on average may be bouncing rather than consuming content.

6.7. Session duration concern

The 18-second average session duration is remarkably low for an academic repository. For comparison, Kamerer [7] notes that bounce rates and session durations vary significantly by context, but sustained engagement typically requires 2–5 minutes for content consumption. This low engagement indicates:

- High bounce rates (users landing and leaving immediately)
- Potential bot traffic inflating user counts without engagement
- Content not matching user expectations (perhaps due to language barriers)
- Navigation or user experience issues

This should be investigated as a potential FAIR accessibility problem rather than presented as neutral.

6.8. Methodological limitations

Three platform transitions introduced measurement discontinuities. GA4's event-based model cannot be directly compared to Universal Analytics' session-based metrics. Specific limitations:

- "Active users" in GA4 uses different calculation than "Users" in UA
- Engagement rate in GA4 is not equivalent to bounce rate in UA
- Channel groupings (Direct, Organic, Referral) may shift due to GA4 logic
- Bot filtering capabilities differ between platforms

We addressed this through conservative metric mapping, but some longitudinal comparisons remain approximate. We recommend treating 2018–2021 (UA) and 2022–2025 (GA4) as two measurement epochs with limited comparability.

7. Conclusions

Fourteen years of web analytics data offer valuable behavioral insights relevant to FAIR-oriented repository management, though with important limitations. This study proposes a proxy-based monitoring framework while emphasizing that web analytics primarily capture usage patterns rather than intrinsic FAIR compliance.

7.1. Key findings

The analysis reveals several notable trends in the repository's usage patterns. Over the 14-year period, the repository experienced a compound annual growth rate of 24.3%, demonstrating remarkable resilience by maintaining continuity even during the COVID-19 pandemic and wartime disruption. Traffic patterns show a strong dominance of direct access, accounting for 86.6% of all visits, which suggests that users primarily access content through persistent identifiers, bookmarks, or untracked referral sources.

Geographic analysis indicates a significant shift in the user base over time. In 2015, 82% of traffic originated from domestic users in Ukraine, whereas the repository now serves a predominantly international audience. However, this international traffic may be inflated by automated systems, warranting careful interpretation of these figures. The average session duration of 18 seconds suggests low engagement, potentially indicating accessibility challenges or content relevance issues that warrant further investigation (as discussed in section 6).

7.2. Practical recommendations for repository administrators

Rather than prescribing universal performance thresholds, this study recommends establishing institution-specific baselines for monitoring repository performance. Administrators should track organic search share, country distribution, device patterns, and download metrics over time, investigating any significant deviations from established norms.

To improve data accuracy, repositories should implement robust bot filtering using GA4's built-in detection features and configure IP exclusions for known automated traffic sources. Geographic patterns should be validated against server logs to distinguish between human users and automated systems. Web analytics should be complemented with metadata audits, FAIR assessment tools such as F-UJI and FAIR-Checker, and user surveys to gain a more comprehensive understanding of repository performance.

The consistently low session duration (18 seconds) suggests potential accessibility barriers or content-language mismatches that may require attention. If mobile traffic remains below 10% of total visits, administrators should investigate potential accessibility issues for mobile-only users.

7.3. Limitations and future work

This study has several important limitations that should inform the interpretation of its findings. The most significant challenge involves the uncertainty surrounding bot traffic, as we were unable to fully validate whether international traffic represents human users or automated systems. Future research should analyze session duration by country, implement more sophisticated bot filtering techniques, and compare GA4 data against server logs for validation.

Another limitation stems from platform discontinuities introduced during the transition from Classic Analytics to Universal Analytics to GA4. These changes in measurement methodology may affect trend analyses, and researchers should treat each platform era as a separate measurement epoch when conducting longitudinal studies.

A fundamental limitation of web analytics is their inability to measure critical FAIR components such as license compliance, metadata quality, or semantic interoperability. The proposed behavioral proxies were not validated against established FAIR assessment tools, and future research should correlate web metrics with F-UJI or FAIR-Checker scores to establish their validity.

As a single-institution study, these findings may not generalize to other repositories. Comparative studies across multiple institutions would provide valuable insights into broader patterns of repository usage and FAIR compliance. Future work should integrate web analytics with citation analysis (OpenCitations), altmetrics (PlumX), and repository-level FAIR assessments to develop more comprehensive impact measurement frameworks [14].

Ethics and privacy statement

Data collection

Web analytics data were collected in strict accordance with Google's Terms of Service and applicable data protection regulations. To safeguard user privacy, IP anonymization was enabled by default across all Google Analytics implementations, ensuring that the final octet of each visitor's IP address was masked prior to processing. Cookie consent mechanisms were deployed in compliance with the General Data Protection Regulation (GDPR) and Ukrainian data protection legislation, providing users with transparent opt-out options for tracking. Raw analytics data were retained for a maximum of 26 months, while aggregated reports were archived indefinitely for longitudinal reference. Crucially, no personally identifiable information (PII) was collected at any stage; all data were processed, stored, and reported exclusively in aggregated and anonymized form.

Ethical considerations

This research relies entirely on aggregated web analytics data and therefore does not constitute human subjects research, exempting it from institutional review board (IRB) approval requirements. Geographic insights are derived solely from country-level aggregations, ensuring that no individual user can be identified or tracked. Furthermore, the study deliberately excludes any content analysis of the materials accessed by visitors. All findings are reported at a macroscopic scale, encompassing thousands to millions of users, which effectively eliminates the risk of re-identification and aligns with established ethical standards for digital trace data.

Limitations

Despite these safeguards, several methodological limitations warrant acknowledgment. First, users employing ad blockers or privacy-enhancing browser extensions are not captured by analytics scripts, which may lead to an underrepresentation of privacy-conscious demographics. Second, automated bot traffic cannot be entirely filtered out; existing literature estimates that 20–40% of repository traffic may originate from non-human sources [18]. Finally, geographic attribution based on IP addresses is inherently imperfect, as users routing their connections through virtual private networks

(VPNs), proxy servers, or corporate infrastructures may be misclassified. These factors should be considered when interpreting the spatial and behavioral patterns presented in this study.

Author contributions

Conceptualization, A.V.K. and S.M.I.; methodology, M.A.S. and A.V.K.; data curation, M.A.S.; formal analysis, T.L.N. and S.M.I.; writing – original draft preparation, A.V.K. and T.L.N.; writing – review and editing, S.M.I. and M.A.S.; visualization, M.A.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability statement

All monitoring reports (2012–2025) are publicly available at Shynenko et al. [17]. Historical data (2012–2021) were extracted from quarterly and annual monitoring reports; 2022–2025 data were extracted from GA4 dashboards. Raw analytics data are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors thank the Department of Open Educational and Scientific Information Systems, Institute for Digitalisation of Education NAES Ukraine, for maintaining continuous analytics collection since 2012. The authors acknowledge Mykola Shynenko's contribution in producing the monitoring reports over this period.

Declaration on Generative AI

The authors have not employed any generative AI tools for content generation. Grammar and spelling checking tools were used for language editing.

A. Metric harmonization across analytics platforms

This appendix documents the mapping of metrics across three Google Analytics platform generations used during the study period.

Table 9
Analytics platform timeline.

Period	Platform	Tracking code
2011–2014	Google Analytics (Classic)	ga.js
2015–2021	Universal Analytics	analytics.js
2022–2025	Google Analytics 4	gtag.js

Key differences:

Table 10
Metric equivalents across platforms.

Classic/UA metric	GA4 metric	Comparability
Users	Active users (28-day)	Approximate
Sessions	Sessions	Comparable
Pageviews	page_view events	Comparable
Bounce rate	Engagement rate	Inverse
Session duration	Average engagement time	Approximate
Organic search %	Organic search %	Comparable
Direct traffic %	Direct %	Comparable

1. *Session definition*: GA4 sessions expire after 30 minutes of inactivity; UA used configurable timeouts. This may affect session counts for repositories with intermittent access patterns.

2. *Bounce rate*: UA defined bounce as single-page sessions; GA4 replaced this with "engaged sessions" (duration > 10 seconds, > 1 page view, or conversion event). Bounce rate in GA4 is calculated as $100 -$ Engagement rate.

3. *User identification*: GA4 uses modeling for user tracking across devices; UA relied on cookies. Cross-device attribution differs between platforms.

4. *Bot filtering*: GA4 applies automatic spam detection; UA required manual filters. Pre-2022 data may include more bot traffic.

Known biases:

1. *Platform transition artifacts*: User counts spike or drop by 5–15% when transitioning between platforms due to different counting methodologies.

2. *GDPR compliance*: Cookie consent requirements (implemented 2018) reduced tracking completeness for EU users, potentially undercounting European traffic post-2018.

3. *Mobile app traffic*: GA4 includes app tracking; UA was web-only. This study excludes app data to maintain comparability.

References

[1] Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K. and Stall, S., 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal*, 19. Available from: https://doi.org/10.5334/dsj-2020-041.

[2] Bollen, J., Sompel, H. Van de and Rodriguez, M.A., 2008. Towards usage-based impact metrics: first results from the MESUR project. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, JCDL '08, p.231–240. Available from: https://doi.org/10.1145/1378889.1378928.

[3] Borrego, Á., 2017. Institutional repositories versus ResearchGate: The depositing habits of Spanish researchers. *Learned Publishing*, 30(3), pp.185–192. Available from: https://doi.org/10.1002/leap.1099.

[4] Davis, P.M., 2002. Patterns in Electronic Journal Usage: Challenging the Composition of Geographic Consortia. *College & Research Libraries*, 63(6), pp.484–497. Available from: https://doi.org/10.5860/crl.63.6.484.

[5] Devaraju, A. and Huber, R., 2021. An automated solution for measuring the progress toward FAIR research data. *Patterns*, 2(11), p.100370. Available from: https://doi.org/10.1016/j.patter.2021.100370.

[6] Gregory, K., Groth, P., Scharnhorst, A. and Wyatt, S., 2020. Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*, 2(2). Available from: https://doi.org/10.1162/99608f92.e38165eb.

[7] Kamerer, D., 2020. Reconsidering bounce rate in web analytics. *Journal of Digital & Social Media Marketing*, 8(1), pp.58–67. Available from: https://doi.org/10.69554/ejgo6858.

[8] Kirtley, O.J., 2022. Advancing credibility in longitudinal research by implementing open science practices: Opportunities, practical examples, and challenges. *Infant and Child Development*, 31(1), p.e2302. Available from: https://doi.org/10.1002/icd.2302.

[9] Konkiel, S., Piwowar, H. and Priem, J., 2014. The Imperative for Open Altmetrics. *The Journal of Electronic Publishing*, 17(3). Available from: https://doi.org/10.3998/3336451.0017.301.

[10] Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. and Murray, S.S., 2005. The effect of use and access on citations. *Information Processing & Management*, 41(6), pp.1395–1402. Special Issue on Infometrics. Available from: https://doi.org/10.1016/j.ipm.2005.03.010.

[11] Linden, A., 2015. Conducting Interrupted Time-series Analysis for Single- and Multiple-group Comparisons. *The Stata Journal: Promoting communications on statistics and Stata*, 15(2), p.480–500. Available from: https://doi.org/10.1177/1536867x1501500208.

[12] Mons, B., 2018. *Data Stewardship for Open Science: Implementing FAIR Principles*. Boca Raton, FL: CRC Press. Available from: https://doi.org/10.1201/9781315380711.

[13] Priem, J., Piwowar, H.A. and Hemminger, B.M., 2012. Altmetrics in the wild: Using social media to explore scholarly impact. 1203.4745, Available from: https://doi.org/10.48550/arXiv.1203.4745.

[14] Romansky, A., Denysiuk, N., Mokliak, S., Svistunov, S. and Shadura, V., 2024. DataverseUA: Peculiarities of Implementation of the Dataverse Open Scientific Data Repository in Ukraine. In: A. Zharinova, ed. *IInd International Conference "Open Science and Innovation in Ukraine 2023"*. Bentham Science Publishers, pp.93–96. Available from: https://doi.org/10.2174/9789815256956124010028.

[15] Schipper, F., 2022. UNESCO World Heritage and Cultural Property Protection in the Event of Armed Conflict. In: M.T. Albert, R. Bernecker, C. Cave, A.C. Prodan and M. Ripp, eds. *50 Years World Heritage Convention: Shared Responsibility – Conflict & Reconciliation*. Cham: Springer International Publishing, Heritage Studies, pp.151–162. Available from: https://doi.org/10.1007/978-3-031-05660-4_12.

[16] Shearer, K., 2013. Institutional Repositories: Towards the Identification of Critical Success Factors. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI*. Available from: https://doi.org/10.29173/cais532.

[17] Shynenko, M.A. et al., 2011–2026. *Monitoring the use of the web resource "Electronic Library of the National Academy of Educational Sciences of Ukraine"*. (Monitoring reports). Kyiv: Institute for Digitalisation of Education NAES Ukraine. Available from: https://lib.iitta.gov.ua/view/creators/==0428==0438==043D==0435==043D==043A==043E=3A==041C=2E==0410=2E=3A=3A.html#group_experiment.

[18] Stassopoulou, A. and Dikaiakos, M.D., 2009. Web robot detection: A probabilistic reasoning approach. *Computer Networks*, 53(3), pp.265–278. Available from: https://doi.org/10.1016/j.comnet.2008.09.021.

[19] Tsakonas, G. and Papatheodorou, C., 2008. Exploring usefulness and usability in the evaluation of open access digital libraries. *Information Processing & Management*, 44(3), pp.1234–1250. Available from: https://doi.org/10.1016/j.ipm.2007.07.008.

[20] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., Silva Santos, L.B. da, Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G.,

Swertz, M.A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), p.160018. Available from: https://doi.org/10.1038/sdata.2016.18.

[21] Wilkinson, M.D., Dumontier, M., Sansone, S.A., Silva Santos, L.O. Bonino da, Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M. and Schultes, E., 2019. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, 6(1), p.174. Available from: https://doi.org/10.1038/s41597-019-0184-5.

[22] Xing, W., Chen, J. and Qiu, C., 2022. How the FAIR Principles Became Reality? – The Operation Mode, Effectiveness and Practical Enlightenment of the FAIRsFAIR. *Journal of Modern Information*, 42(7), pp.136–146. Available from: https://doi.org/10.3969/j.issn.1008-0821.2022.07.012.

[23] Yaroshenko, T., 2021. Open Access, Open Science, Open Data: How it Was and Where We are Going (To the 20th Anniversary of the Budapest Open Access Declaration). *Ukrainian Journal on Library and Information Science*, (8), p.10–26. Available from: https://doi.org/10.31866/2616-7654.8.2021.247582.