
# Unlock Historical Web Data with AI-Driven Trend Insights

## Report – WebInsider

Poltergeist: , Georg Hoch , Claudius Kroflin, Tim Bajramaj

Big Data – 14.05.2025

# Context

"There were 5 exabytes of information created between the dawn of civilization through 2003, [...] that much information is now created every two days." (Schmidt, 2005) This staggering acceleration continues, with global data creation expected to reach 175 zettabytes by 2025 (Reinsel, Gantz, Rydning, 2018). This explosion in data availability coincides with the growing recognition that analytically informed choices deliver tangible competitive advantages, with research suggesting that firms effectively implementing data-driven decision-making consistently outperform competitors in profitability and operational efficiency (Brynjolfsson, M. Hitt, Kim, 2011).

As business landscapes evolve rapidly, decision-makers require immediate access to reliable insights. With the web serving as the primary channel for marketing and external communications—accounting for over 65% of customer touchpoints (L'Hostis, 2024) - historical web trends represent an untapped resource for strategic planning to gain strategic advantages.

WebInsider converts historical web data into strategic intelligence, enabling long-term decision-making by revealing trends in the popularity of terms over time and providing AI-driven, contextual insights that reveal emerging market opportunities.

# Relevance: WebInsider helps to leverage historical data

**Business Case**

WebInsider turns massive, unstructured web archives into strategic intelligence by utilizing sophisticated data processing. The main target audience of WebInsider are strategic decision-makers in various organizations that should be enabled to assess their competitors' moves by long-term contextual analysis powered by AI, deeper historical insights. Whereas Google Trends only shows interest from customer side – “demand”, WebInsider shows the “supply” side. The value add of WebInsider lies in insights about the implementation areas and speed of the supply side.

**Usage of Big Data**

WebInsider leverages the entity of Big Data to increase value to its customers across all dimensions: WebInsider processes large volumes of historical web data—like blogs, news, and academic content - from sources such as Arquivo.pt, distilling it into focused visualizations that highlight long-term term popularity trends. As it evolves, it will incorporate real-time data sources to enhance insight speed (Velocity), expand beyond selected formats to include CSV, PDF, HTML, and more (Variety), and implement rigorous quality checks to ensure reliable results (Veracity). Ultimately, it transforms static data into actionable intelligence through AI-driven analysis and visual tools that support marketing, competitive strategy, and product development (Value).

**Monetization**

WebInsider, once launched, offers a transparent pricing module – based on a subscription model:

- **Free**: Basic features
- **Base (€10/mo)**: Faster access, AI insights, support
- **Pro (€25/mo)**: Full access, unlimited searches, advanced AI, premium support

## Innovation: WebInsider transforms web archives into strategic intelligence

WebInsider leverages AI to contextualize historical & soon real data, revealing hidden patterns that Google Trends and the Wayback Machine miss. It combines trend visualization with contextual analysis, going beyond current search interest and static web snapshots.

- **AI-Enhanced Contextual Analysis:** WebInsider leverages state of the art large-language-models, such as Claude, to contextualize relevant patterns in timeseries, linking them to events, cultural trends, or media coverage,

transforming raw data into actionable insights through comprehensive interpretation beyond traditional trend analysis.

- **Cross-platform Data Integration:** WebInsider unifies data from sources like Arquivo.pt, Wayback Machine, Google Trends, JSTOR, and social media (e.g. X, Instagram, TikTok) via LLMs to track term evolution. This holistic view reveals trends driven by media, academia, and social activity.
- **Expert and user-centric interfaces:** WebInsider dual interface empowers non-technical users to extract contextual insights through a user-friendly design while data experts access advanced trend analysis, customizable visualizations, and sophisticated data manipulation tools for deeper insights.
- **Forecasting of trends:** WebInsider plans to implement AI-driven predictive analytics, leveraging historical data patterns to identify emerging trends and anticipate future shifts in the popularity of terms and media coverage.

## Maturity: WebInsider current limitations are tomorrow's growth opportunities

WebInsider takes a deeper approach by combining AI-powered contextual analysis with historical trend analysis. By revealing patterns that traditional tools miss, this dual functionality allows for a deeper understanding of the evolution of digital terms.

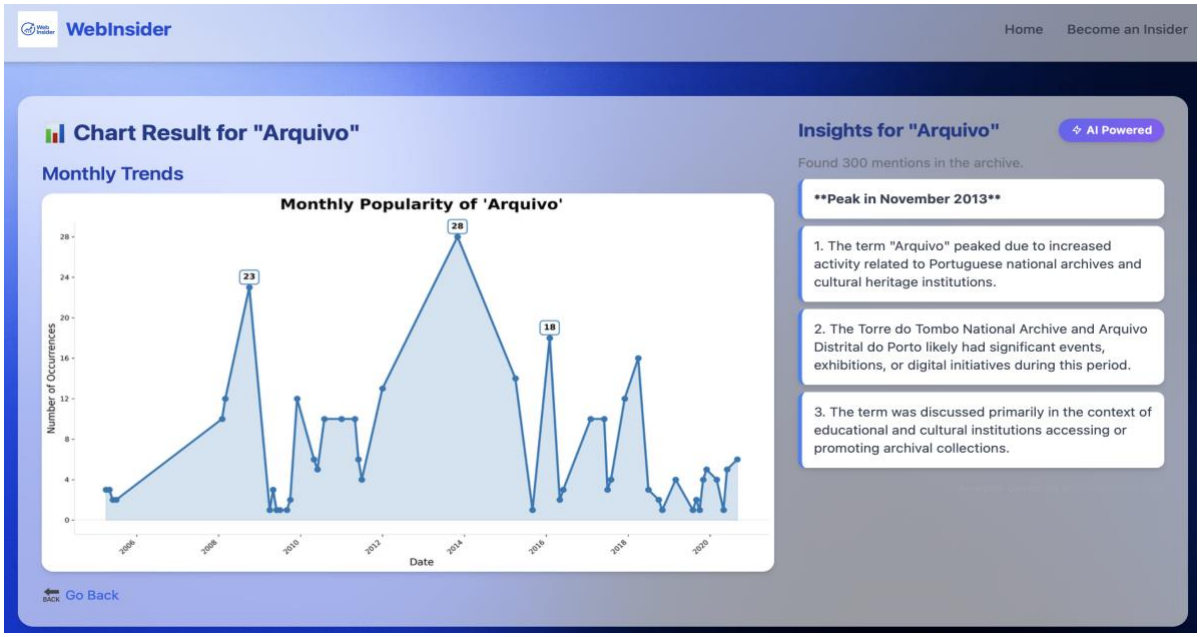


*Figure 1 Demo of the MVP for search term "Arquivo"*

WebInsider uses contextual analysis and top-notch large language models to explain increases in the popularity of terms and their root cause in existing databases. For example, when analyzing “Trump”, the algorithm discovered a notable increase in June 2019 and linked it to a spike in searches for biographical information across different language platforms, revealing trends of global interest.

However, several limitations remain unaddressed in the current early stage of implementation which will be addressed at later stage:

- **Website Type Filtering**: Currently, WebInsider treats all websites uniformly. Introducing website type differentiation (e.g., news, academic, commercial, social media) will unlock more targeted insights for specialized user groups like marketers.
- **Regional Differentiation**: Adding geographical filtering will enhance localized trend analysis, empowering users to uncover regional patterns crucial for targeted campaigns.
- **Historical Context Enhancement**: Present AI explanations offer basic insights. Upcoming task-specific AI models will provide richer context by integrating historical, social, economic, and cultural dimensions.
- **Scalability Improvements**: While high-demand scenarios may strain performance, planned optimizations like data caching and efficient query handling will ensure smooth analysis at scale.
- **ExpandedData Sources**: Currently focused on Arquivo.pt, WebInsider will gain greater analytical depth and regional coverage by incorporating additional archives.
- **Accessible Website**: The platform will evolve to include enhanced accessibility features—such as keyboard navigation and customizable displays—ensuring a more inclusive user experience.

These limitations represent natural opportunities for future development as WebInsider evolves from a functional prototype to a comprehensive business solution.

WebInsider current state can be found in the following GitHub repository

# References


- Anthropic. (n.d.). Claude API – AI-Powered Conversational Models. Retrieved May 13, 2025, from https://www.anthropic.com/claude
- Arquivo.pt. (n.d.). Web Archive of Portugal. Retrieved May 13, 2025, from https://arquivo.pt/
- Brynjolfsson, M. Hitt, Kim, 2011: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486
- L'Hostis 2024: https://www.forrester.com/blogs/2025-the-digital-banking-landscape-is-poised-for-another-transformative-year/?utm_source=chatgpt.com
- Reinsel, Gantz, Rydning 2018: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf?utm_source=chatgpt.com Page 3
- Schmidt, 2005: https://techcrunch.com/2010/08/04/schmidt-data/