Vespa.ai

Vespa Case Study: DeviantArt

Delivering AI-powered content discovery and personalization for DeviantArt, one of the world's largest online art communities and a Wix subsidiary.



We Make AI Work

© Vespa.ai Norway AI. July 2026

Vespa.ai

Vespa.ai develops the Vespa AI Search Platform, which unifies retrieval, ranking, machine-learning inference, and real-time serving for large-scale search and AI-native applications.

Document
Vespa Case Study: DeviantArt

08.07.2026

© Vespa.ai Norway AI. July 2026

Contents

↗ About this eBook
How production retrieval workflows are designed and implemented at scale.

Read

↗ Introduction
Wix subsidiary DeviantArt operates one of the largest online communities dedicated to artists and creative expression.

Read

↗ The Challenge
Powering discovery across the world's largest online art community demands more than traditional search.

Read

↗ The AI Search Platform
Recap: the evolution of AI Search is a platform that executes the entire retrieval workflow.

Read

↗ Engineering AI Search at DeviantArt
How DeviantArt applies this workflow to deliver fast, highly personalized content discovery

Read

↗ Summary
DeviantArt's production deployment demonstrates AI-powered discovery in practice.

Read

About this eBook

Who should read this ebook?

This eBook is intended for software architects, search engineers, AI engineers, and technical leaders responsible for building customer-facing AI applications. Whether you're developing AI search, recommendations, RAG, or AI agents, it provides practical insight into how production retrieval workflows are designed and implemented at scale.

AI-powered discovery has evolved far beyond traditional search.

Modern applications must retrieve trusted information, personalize results, apply machine learning, and serve fresh content at internet scale—all within the latency budgets users expect. Success depends not on any single technology, but on how these capabilities work together as a unified retrieval workflow.

Drawing on DeviantArt's production deployment, this eBook explores how the Vespa AI Search Platform enables that workflow. It demonstrates how retrieval, ranking, personalization, and machine learning work together within a single distributed platform to deliver fast, relevant, and scalable AI-powered experiences.

© Vespa.ai Norway AI. July 2026

Introduction



Wix subsidiary DeviantArt operates one of the largest online communities dedicated to artists and creative expression. With more than 100 million registered members and over 650 million pieces of artwork spanning digital art, photography, animation, illustration, sculpture, and emerging media formats, the platform depends on high-performance discovery and personalization systems to help users navigate an enormous and constantly evolving content ecosystem.

Every day, tens of thousands of new works are uploaded by artists around the world, creating a continuous stream of fresh content, changing trends, and highly diverse user interests. Supporting discovery at this scale requires more than traditional search infrastructure. DeviantArt needs to process large volumes of content in real time, understand different artistic styles and terminology, personalize experiences for different users and communities, and deliver responsive search and recommendation experiences under constantly changing traffic conditions.

Using Vespa.ai, DeviantArt built a scalable AI-powered discovery platform that combines machine learning, custom query processing, document enrichment, and real-time ranking within a single serving layer. Vespa's architecture enables the platform to support advanced personalization, custom linguistic processing, and flexible ranking models while maintaining the low latency and operational scalability required by one of the world's largest creative communities.

Results at Scale

- 1B+ documents
- Nearly 4,000 QPS
- 12 ms average query latency
- Sub-millisecond feed latency

© Vespa.ai Norway AI. July 2026

The Challenge

Powering discovery across the world's largest online art community demands more than traditional search.

Every day, tens of thousands of new works are uploaded by artists around the world, creating a continuous stream of fresh content, changing trends, and highly diverse user interests. Supporting discovery at this scale requires more than traditional search infrastructure. DeviantArt needs to process large volumes of content in real time, understand different artistic styles and terminology, personalize experiences for different users and communities, and deliver responsive search and recommendation experiences under constantly changing traffic conditions.

Using Vespa.ai, DeviantArt built a scalable AI-powered discovery platform that combines machine learning, custom query processing, document enrichment, and real-time ranking within a single serving layer. Vespa's architecture enables the platform to support advanced personalization, custom linguistic processing, and flexible ranking models while maintaining the low latency and operational scalability required by one of the world's largest creative communities.

Today, the deployment manages more than 1 billion documents and serves nearly 4,000 queries per second, with an average query latency of 12 milliseconds. At the same time, the platform continuously ingests new and updated content at sub-millisecond feed latency, enabling DeviantArt to deliver fast, highly personalized discovery experiences across one of the world's largest and most active online art communities.

© Vespa.ai Norway AI. July 2026

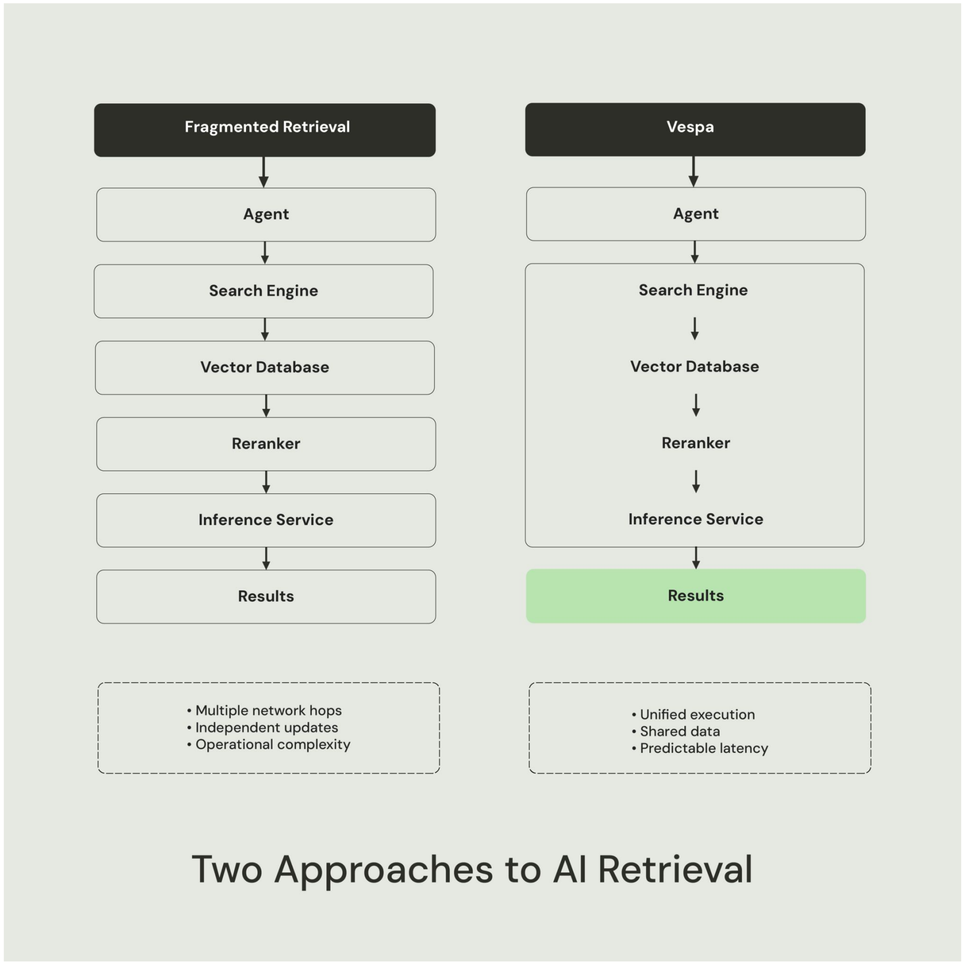
“At DeviantArt, discovery is fundamental to the user experience. With hundreds of millions of artworks, rapidly changing content, and highly diverse user interests, we needed a platform capable of combining machine learning, personalization, and real-time serving performance at a very large scale.

Vespa gives us the flexibility to integrate custom ranking models, user context, and domain-specific processing within a single high-performance system, which is critical for delivering responsive and relevant experiences across our global community.”

Chris Nell,
Chief Technology Officer, DeviantArt

© Vespa.ai Norway AI. July 2026

Recap: The AI Search Platform



Two Approaches to AI Retrieval

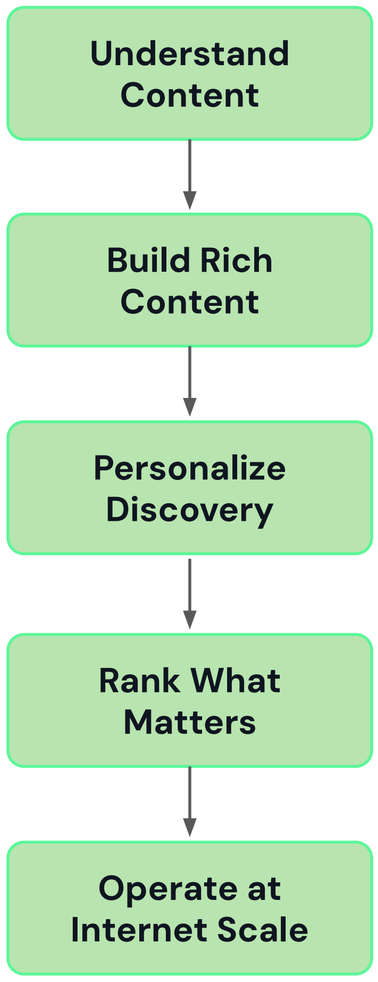
AI Search Platforms have become the execution layer for search applications—from traditional search and recommendations to answer engines and AI agents. Rather than assembling retrieval workflows from multiple specialized systems, they integrate retrieval, ranking, machine-learning inference, and real-time serving into a single architecture.

This unified approach enables organizations to execute the entire retrieval workflow as one system, reducing operational complexity while improving search quality, scalability, and infrastructure efficiency. As AI applications become more sophisticated, the platform, rather than the individual components, becomes the foundation for delivering fast, accurate, and trustworthy results.

Vespa is an AI Search Platform built for Retrieval Engineering, enabling customer-facing intelligent applications to deliver fast, trustworthy, and scalable AI experiences.

© Vespa.ai Norway AI. July 2026

Engineering AI Search at DeviantArt



© Vespa.ai Norway AI. July 2026

Delivering AI-powered discovery requires far more than vector search alone. Every user interaction depends on a sequence of capabilities that work together to understand content, enrich it with context, personalize retrieval, intelligently rank results, and serve them with predictable performance at internet scale. Rather than assembling these capabilities from multiple specialist systems, the Vespa AI Search Platform executes them within a single distributed architecture.

The following sections show how DeviantArt applies this workflow to deliver fast, highly personalized content discovery for one of the world's largest online art communities. The key stages are:

- Understanding Content
- Building Rich Content
- Personalizing Discovery
- Ranking What Matters
- Operating at Internet Scale

Together, these stages illustrate the principles of Retrieval Engineering, implemented through the Vespa AI Search Platform.

1. Understanding Content

Teaching AI search to understand your domain, language, and users.

The Challenge

Every day, new artworks are uploaded using different naming conventions, artistic styles, fandom terminology, and community-generated tags. Helping users discover relevant content requires search to understand this evolving vocabulary rather than relying on generic language processing.

The Vespa AI Search Platform

Vespa supports custom linguistic processing, allowing organizations to tailor how search understands language, terminology, synonyms, and domain-specific vocabularies. Rather than relying on generic language models alone, search behavior can be adapted to reflect the unique language of a particular community or domain.

DeviantArt in Practice

DeviantArt uses custom linguistic processing to improve how users discover artwork across highly diverse artistic styles, fandoms, techniques, and community-driven tags, helping surface more relevant content despite variations in naming, spelling, and artistic terminology.

2. Building Rich Content

Enriching information with the context needed for accurate retrieval and ranking.

The Challenge

High-quality AI search depends on far more than raw documents. Content must be enriched with metadata, relationships, and contextual information that help retrieval and ranking understand how people, artwork, collections, and communities are connected.

The Vespa AI Search Platform

Vespa combines document processing pipelines with flexible document models that support enrichment before indexing. Content can be normalized, classified, tagged, and linked through parent-child relationships, allowing related entities to remain independently searchable while preserving their relationships.

DeviantArt in Practice

DeviantArt enriches newly uploaded artwork with searchable metadata and normalized attributes while modeling relationships between artists, artwork, galleries, collections, and associated metadata. This richer content model improves discoverability and provides additional context for downstream ranking and personalization.

© Vespa.ai Norway AI. July 2026

3. Personalizing Discovery

Tailoring every search and recommendation to each user's interests and intent.

The Challenge

Every user has different interests, communities, and behaviours. Delivering relevant discovery experiences requires search to dynamically incorporate user context and business logic, rather than returning the same results for every query.

The Vespa AI Search Platform

Vespa supports custom query processing to modify queries, apply business rules, integrate user context, and personalize retrieval before ranking. This enables organizations to build highly tailored search and recommendation experiences without introducing additional processing layers.

DeviantArt in Practice

DeviantArt uses custom query processing to incorporate user profiles, interests, and behavioral context into search and discovery experiences, enabling more personalized and community-relevant recommendations across its creative community.

4. Ranking What Matters

Applying intelligence to identify the most relevant content for every interaction.

The Challenge

Retrieval identifies candidate content, but ranking determines what users actually see. As content volumes grow, ranking must combine semantic relevance, behavioral signals, personalization, and structured metadata while maintaining low latency.

The Vespa AI Search Platform

Vespa evaluates machine learning and ONNX models directly within the serving layer, allowing sophisticated ranking models to execute in real time without moving data between separate systems. This enables AI-driven ranking while maintaining predictable serving performance.

DeviantArt in Practice

DeviantArt uses Vespa's machine learning capabilities to support AI-driven ranking and personalization across rapidly changing content, helping users discover artwork and artists aligned with their interests and engagement patterns.

© Vespa.ai Norway AI. July 2026

5. Operating at Internet Scale

Delivering low-latency AI search reliably as data, users, and workloads grow.

The Challenge

Large-scale AI applications must remain responsive despite changing traffic patterns, rapidly growing content, evolving business requirements, and increasing operational demands. Scaling infrastructure should not increase architectural complexity.

The Vespa AI Search Platform

Vespa combines autoscaling, secure deployment isolation through Enclave, and an extensible component framework within a single distributed platform. Together these capabilities allow organizations to scale efficiently, maintain operational separation where required, and extend platform behaviour without introducing additional serving infrastructure.

DeviantArt in Practice

DeviantArt uses Vespa autoscaling to support highly variable global traffic driven by trending artwork and community activity. Enclave provides secure operational isolation for critical workloads, while custom Vespa components allow the platform to tailor search, ranking, and discovery behavior to the evolving needs of its creative community.

© Vespa.ai Norway AI. July 2026

Summary



Inspired by what DeviantArt has achieved?

Whether you're building AI search, recommendations, RAG, or AI agents, we'd be happy to discuss how the Vespa AI Search Platform can help you deliver fast, personalized AI experiences at production scale.

Contact us.

© Vespa.ai Norway AI. July 2026

DeviantArt's production deployment demonstrates that AI-powered discovery depends on far more than vector search alone. Delivering personalized content discovery at internet scale requires retrieval, ranking, machine-learning inference, custom query processing, and continuously updated content to work together as a unified workflow.

Rather than assembling these capabilities from multiple specialist systems, the Vespa AI Search Platform executes them within a single distributed architecture, reducing complexity while delivering fast, relevant, and scalable AI-powered experiences.

The result is a production AI search platform capable of supporting more than one billion documents, nearly 4,000 queries per second, and highly personalized discovery with an average query latency of just 12 milliseconds. Together, these capabilities illustrate Retrieval Engineering in practice, implemented through the Vespa AI Search Platform.

About Vespa.ai

Vespa.ai develops the Vespa AI Search Platform that brings retrieval, ranking, machine-learning inference, and real-time serving together within a single distributed architecture. Rather than stitching together fragmented search and AI retrieval components, Vespa executes the complete retrieval workflow close to the data, delivering high relevance, predictable latency, and operational simplicity at scale. Organizations including Yahoo, Spotify, Perplexity, and AlphaSense use Vespa to power mission-critical customer-facing applications.

Interested to learn more? We have many different resources and information available through our social platforms

GitHub Twitter LinkedIn YouTube

Vespa.ai

© Vespa.ai Norway AI. July 2026