Paper ML07

Poseidon: Streamlined Regulatory Screening of HA Websites with AI/ML to Extract Updates and Reduce Manual Efforts

Maria Giovanna Zilli, Merck KGaA, Vevey, Switzerland
Fabian Wahren, HMS Analytical Software GmbH, Heidelberg, Germany

ABSTRACT

Regulatory compliance in the pharmaceutical industry requires continuous monitoring of health authority updates, with subject matter experts at Merck KGaA currently spending considerable time manually screening international regulatory websites for news, guidelines and regulations updates. This manual approach consumes significant resources and risks missing critical updates, as the increasing rate of publications makes it challenging to confidently identify relevant information that could impact the business. We are developing Project Poseidon to address these challenges through an artificial intelligence/machine learning system that automates regulatory surveillance via two core phases: automated crawling for content screening, and intelligent triage for scoring, prioritization and stakeholder assessment. The cloud-based system utilizes a microservices architecture, implementing semantic analysis, large language model-based summarization, and combined rule-based and machine learning approaches for scoring relevance. Early results suggest strong potential for reducing manual screening time and enabling scalable deployment across multiple departments. Project Poseidon establishes a robust foundation for enhanced regulatory compliance automation, enabling SMEs to reallocate efforts toward more meaningful, strategic activities while maintaining the interpretability and traceability requirements essential for regulatory decision-making in pharmaceutical environments.

INTRODUCTION

In the pharmaceutical industry, regulatory compliance depends on timely awareness of policy changes, safety alerts, and guideline updates from Health Authorities (HAs) worldwide. Traditional approaches to regulatory surveillance might consider the establishment of a worldwide network and should rely heavily on manual processes, where domain experts continuously monitor multiple HA websites, assess document relevance, and prioritize updates for organizational response. This becomes even more challenging when different business units monitor overlapping sources, multiplying the manual effort required across the company. With regulatory publications appearing at an ever-increasing rate, these manual screening approaches are becoming difficult to sustain.

This paper presents project Poseidon, a machine learning initiative designed to automate regulatory screening processes at Merck KGaA. The project addresses the technical and practical challenges of replacing manual surveillance while maintaining the accuracy and domain-specific knowledge required for regulatory decision-making. The following sections outline the project's objectives, architectural design choices, and describe its technical implementation. Lastly, we shed some light on the challenges we faced, both on the conceptual as well as on the technical side.

PROJECT DESCRIPTION

Project Poseidon aims to streamline regulatory screening of HA websites at Merck Healthcare, where SMEs currently spend a lot of effort dedicated to manual screening for relevant documents and triage priorities. By developing an artificial intelligence (AI)/Machine Learning (ML) system to automatically screen updates from international HAs published on websites, Poseidon will reduce manual efforts and allow to reallocate these efforts on more strategic and advocacy activities.
This system is expected to enhance efficiency and reallocate SMEs efforts in regulatory activities, ultimately improving regulatory compliance and providing comprehensive assessments of regulatory impacts.

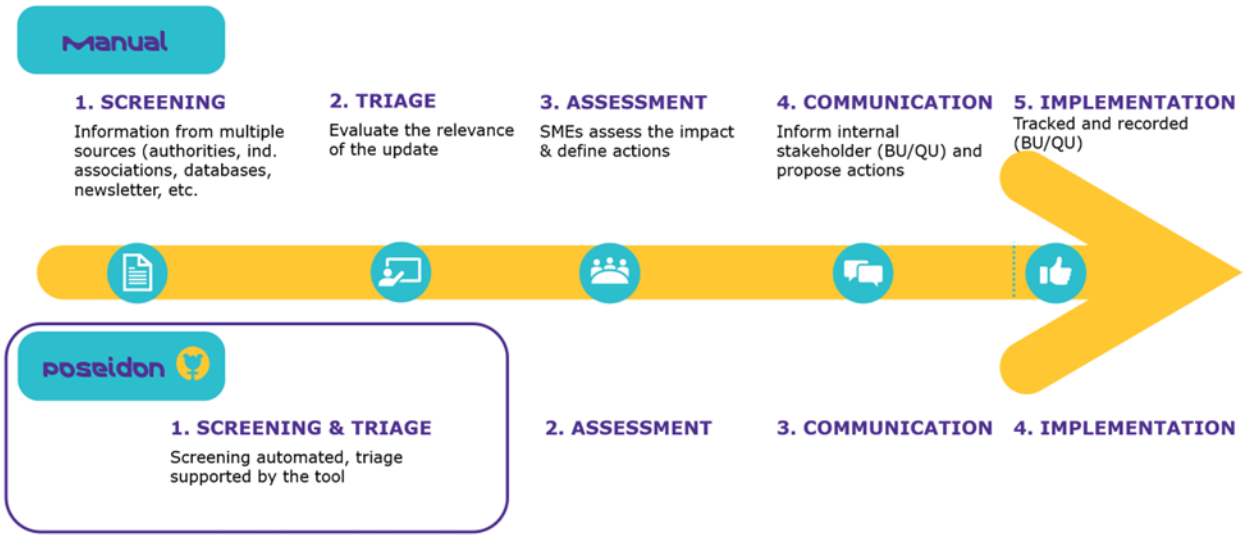

Figure 1: Regulatory screening process: manual vs semi-automated

The regulatory screening process, shown in Figure 1, generally follows a standardized five-step workflow that is mostly applicable across all business units (BU) within the organization, beginning with screening information from multiple sources such as Health Authorities websites, industry associations, external knowledge databases, and newsletters, followed by triage to evaluate the relevance of updates, assessment where subject matter experts (SMEs) analyze impact and define actions, to move to the communication step to inform internal stakeholders and propose actions, and finally implementation where actions are performed, tracked and recorded.

Project Poseidon strategically focuses on automating and supporting only the first two steps—screening and triage—because these processes involve fundamentally similar actions across all functional areas, primarily consisting of information gathering, relevance evaluation, and initial prioritization activities that can be standardized and automated effectively. The subsequent steps of assessment, communication, and implementation remain outside the project scope as they require highly specialized, function-specific expertise and decision-making processes that vary significantly between different business units. By concentrating on the screening and triage phases, Poseidon delivers maximum value through automation of the most time-intensive and standardizable components of the regulatory surveillance workflow, while allowing each functional area to maintain their specialized expertise and customized approaches for the downstream assessment, communication, and implementation activities that define their unique regulatory responsibilities.

The primary objective is to develop two main must-have functionalities. Specifically: crawling of regulatory news from HA website with high level of accuracy, and prioritization of relevant regulatory news through triage.

The Screening phase focuses on data ingestion from diverse Health Authorities websites, preprocessing, and initial filtering to identify relevant updates based on business function's specification and topics of interest.

The Triage phase involves detailed content analysis, categorization, prioritization, and routing updates done by appropriate stakeholders based on criteria defined by the respective business unit. This system reduces manual effort, enhances accuracy, and ensures timely identification of critical regulatory changes, supporting efficient compliance and decision-making processes.

Conditions to consider the system complete were identified as follows:

1. **Development of AI/ML Model**: The project should deliver an AI/ML model developed and deployed effectively, demonstrating the ability to accurately screen and assess regulatory updates from targeted Health Authorities (HAs) websites.
2. **Scalability**: The system should be scalable to other company departments. Hence, the system is built to ensure functional flexibility and customization of information retrieval and prioritization criteria.
3. **Reduction in Time Spent and increase of reliability on Regulatory Screening**: A measurable decrease in the time spent by business SMEs on regulatory screening activities.
4. **User Adoption and Satisfaction**: High levels of adoption and satisfaction among users, including business SMEs and local CRAs, measured through feedback surveys and usage analytics, indicating that the system meets their needs.

We believed that being able of developing a system which delivers the information tailored to our company's interest,

not only on a high-level domain perspective such as safety or manufacturing but going deeper into specific technologies or product-related alerts will decrease the complexity and amount of regulatory news that we are managing.

POSEIDON FUNCTIONALITIES

The Poseidon project includes functionalities designed to revolutionize regulatory surveillance through intelligent automation.

To start, the system employs automated **crawling mechanisms** to automatically extract and aggregate regulatory news from multiple Health Authority websites. These crawling mechanisms are scheduled to run overnight, ensuring continuous data updates that maintain current regulatory intelligence. The system's flexibility grants the capability of efficiently managing content's customization, extracting data based on functional scope requirements.
The platform's architecture enables seamless retrieval of regulatory updates from diverse sources and various web platforms. This comprehensive approach culminates in a unified common inbox that consolidates all regulatory information regardless of its original source.

The system's **automation and validation framework** streamline data fetching and aggregation processes while implementing rigorous validation protocols to ensure completeness, consistency, and accuracy of all retrieved data. Information screening capabilities leverage artificial intelligence and machine learning techniques, including natural language processing, to assess the relevance of regulatory updates.
The system analyzes regulatory content based on predefined categories encompassing guideline updates, safety notices, and policy changes. Relevance criteria consider multiple factors including the specific regulatory authority, document status whether draft or final, document titles, publication dates, extraction dates, and screening topics such as CMC, Quality/GMP (Good Manufacturing Practices), or Safety. The system provides customization options, allowing users to define specific relevance criteria that tailor the system's focus to their areas of interest.

Version handling represents a critical functionality that maintains versioning mechanisms to track changes in regulatory documents at the draft stage. The system ensures clear labeling with timestamps and comprehensive metadata for each version, enabling detailed comparison capabilities that highlight differences such as content additions, deletions, or modifications. Users retain the ability to revert to previous versions when required, while comprehensive audit trails log version histories for all documents and updates, ensuring full compliance and traceability throughout the document lifecycle.

Duplicate handling functionality implements algorithms to detect duplicated entries across multiple sources, effectively identifying redundant documents or updates. The system ensures that duplicates undergo minimal processing, storing only essential information such as titles and download locations. Consolidated notification systems maintain detailed logs of duplicate data for comprehensive traceability and reporting purposes in the background.

The **triage and prioritization functionality** implements comprehensive scoring mechanisms that rank and label documents, topics, and updates based on urgency, impact, and relevance factors. This configurable system adapts to different user groups with varying profiles regarding criticality, urgency, and relevance requirements. Manual override capabilities allow users to adjust priority scores or rankings as needed, while topic-specific prioritization enables tailored selection of regulatory news by specific topics with urgency-based assignment that ensures critical news receives immediate attention. The system incorporates feedback loops where users provide input to IT support for continuous AI and ML model training.

Document generation capabilities aggregate selected updates into predefined formats such as newsletters and reports, ensuring inclusion of key details including update titles, publication dates, regulatory authorities, and comprehensive summaries or abstracts. Template customization provides flexible options for consistent branding and formatting while allowing users to save multiple templates for different document types. Automated formatting aligns content with selected templates while including tools for manual review and editing, encompassing text adjustments and hyperlink management. Personalization features tailor documents based on recipient preferences and subscription profiles.

Alert generation functionality creates automated alerts for regulatory updates, including new guidelines, safety notices, and major policy changes. Alert content includes critical information such as update titles, regulatory authorities, publication dates, and document status.

TECHNICAL IMPLEMENTATION

This section outlines the technical implementation of the system described above. After a look at architectural aspects, it presents all parts of the system from data ingestion to the interface with which users can interact with the system. Finally, we will present some challenges as well as potential extensions for the future.

TECHNICAL ARCHITECTURE

Poseidon is a cloud-based application that is embedded into the already existing IT landscape of the company, leveraging its architectural frameworks as well as user management. Due to one of the core requirements for Poseidon to be highly extensible and scalable, the system utilizes microservices to ensure operability even under high workloads. All components are deployed and hosted on Amazon Web Services (AWS). This design choice allows for horizontal scaling as well as dynamic orchestration to both process substantial data volumes generated by continuous monitoring of multiple Health Authority websites as well as serving many users at the same time. Its modular architecture facilitates incremental development which accelerates the release cycles and reduces user feedback latency.

Since in the future other departments and groups within Merck KGaA will use the system, Poseidon's architecture is designed to be adapted and extended to different business needs, domains and websites. This is done while maintaining enterprise-grade performance and reliability standards to support the long-term vision of expanding coverage across multiple domains and organizational functions.

CRAWLING AND PROCESSING PIPELINE

The foundation of our system rests on a crawling engine designed for data extraction across diverse digital environments that implement their own structures and formats. This raises an important question about scalability—how do we maintain performance while handling increasingly complex data sources? Our approach centers on parsing algorithms that adapt to different source formats and extract text while also capturing relevant meta-data.

The system supports multiple file formats, including PDF documents and dynamic HTML content and adapts to different website structures, though this adaptation requires periodic adjustments for sites with particularly complex architectures. One aspect that also deserves attention is our URL tracking capability. Rather than treating web resources as static entities, the system monitors the availability of URLs and articles. This monitoring proves valuable for maintaining data integrity and makes sure that all data within the application is up to date.

The flexibility of the scraping engine proves useful in practice, allowing users to parse content from various entry points—news websites, RSS feeds, email newsletters without rebuilding the entire pipeline. This plug-and-play approach works well for most sources, making it easy to onboard new organizational functions that bring their own set of sources they want to monitor.

To stay compliant, we've implemented full traceability of information sources, which appears essential for regulatory applications. The complexity of maintaining this traceability increases with content that undergoes multiple transformation steps, requiring notification of specific personnel and/or stakeholders.

CONTENT ANALYSIS

Poseidon undergoes a series of processing steps after having scraped a new article to extract information and additional metadata, calculate scores and apply business specific logic. One of the first steps is to detect, whether this new document has already been processed before, either completely equivalent, in a different format or slightly altered, to assess whether it should be treated as a duplicate or a new version of an existing document. Standard diff algorithms miss the semantic significance of changes—a single word modification might completely alter legal interpretation, while extensive formatting changes might be entirely cosmetic. Our semantic analysis evaluates changes based on meaning rather than just textual differences, providing more contextually relevant change detection.

In a subsequent step, the automatic summarization module addresses the challenge of lengthy regulatory documents, leveraging state-of-the-art large language models (LLMs). We've found that effective summarization requires balancing comprehensiveness with accessibility. The results show promise for partial comparison of documents based on summaries as well as creating a quick user understanding of regulatory updates.

The scoring system allows custom score definition incorporating business-specific criteria. We've implemented a plug-and-play approach for scoring algorithms, which facilitates extensibility as requirements evolve. New algorithms can be integrated with reasonable development efforts. Our approach combines rule-based logic with machine learning techniques. Pure machine learning models, while powerful, lack the interpretability that regulatory professionals require. The hybrid approach provides more reliable and explainable results, with the balance between components adjustable based on specific use cases.

USER INTERFACE AND INTEGRATION ARCHITECTURE

To make the data accessible to regulatory screeners, we built a web application offering a wide range of tools to assist the user with staying on top of regulatory updates. Since the assessment of the latter often results from a team effort, Team collaboration features emerged as a requirement. Most regulatory decisions involve multiple stakeholders with different areas of expertise. The platform facilitates this collaborative process by offering a joint view on new documents as well as comment features enabling lively discussions. For new documents, automatically calculated scores can be

annotated and corrected by the domain experts. This creates a continuous improvement loop that enhances system accuracy over time.

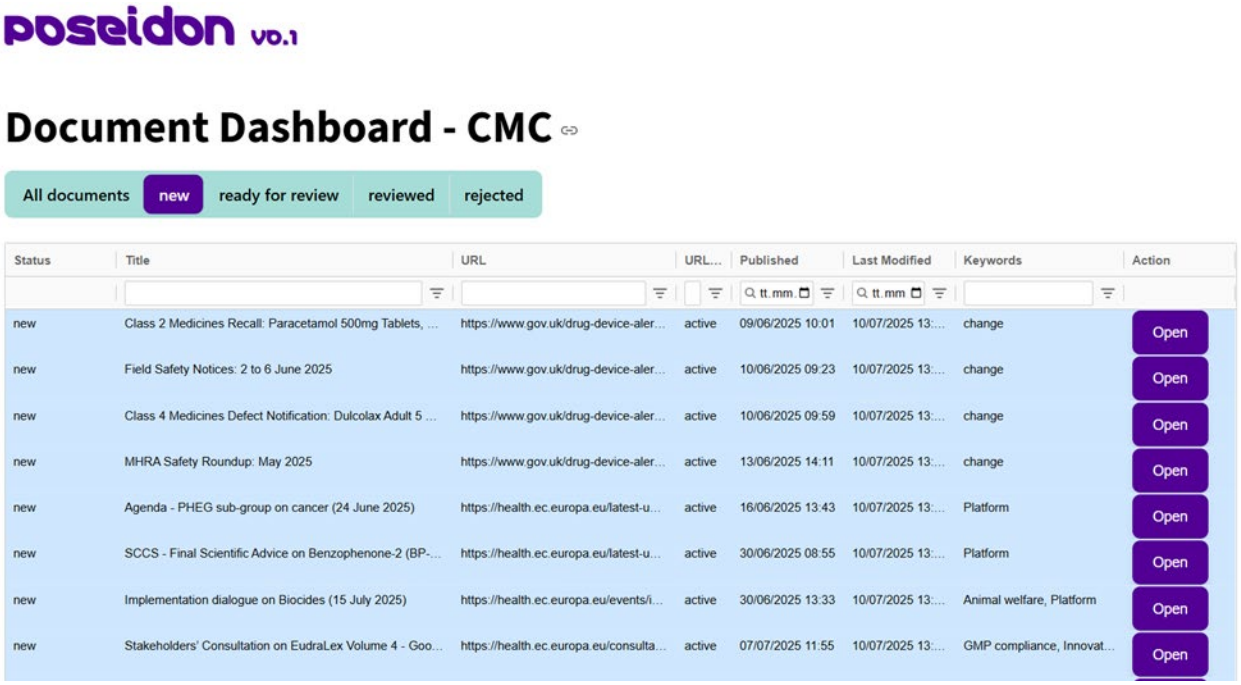

Figure 2: Document Dashboard for organizational function "CMC"

The current architecture aims to serve both power users requiring extensive customization and occasional users needing straightforward and intuitive tools for their everyday tasks. Balancing comprehensive functionality against interface simplicity involves thoughtful progressive disclosure of features based on user roles and experience levels, as shown in Figure 2.

The Poseidon Document Dashboard exemplifies a simple and clean design philosophy that prioritizes immediate access to the most critical regulatory information through strategically organized column headers. The interface presents essential data points, beginning with the Status column that provides instant visual recognition of document processing stages through labels such as "new," "ready for review," "reviewed," and "rejected." The Title column delivers the primary regulatory content identifier, allowing users to quickly understand the nature and scope of each regulatory update without unnecessary complexity. The URL columns provide direct access pathways to source documents, while the Published and Last Modified timestamps offer a temporal context. The Keywords column serves as a categorization system, displaying relevant tags that facilitate rapid content assessment and filtering based on functional expertise areas. Finally, each document can be reviewed by clicking on the respective “Open” button (detailed explanation below after Figure 3: Document Review page). This column structure eliminates information overload while ensuring that regulatory professionals can efficiently identify, prioritize, and access the most pertinent regulatory intelligence, transforming what could be a complex data management challenge into an intuitive, user-friendly experience that supports rapid regulatory decision-making and maintains compliance oversight across multiple regulatory domains.

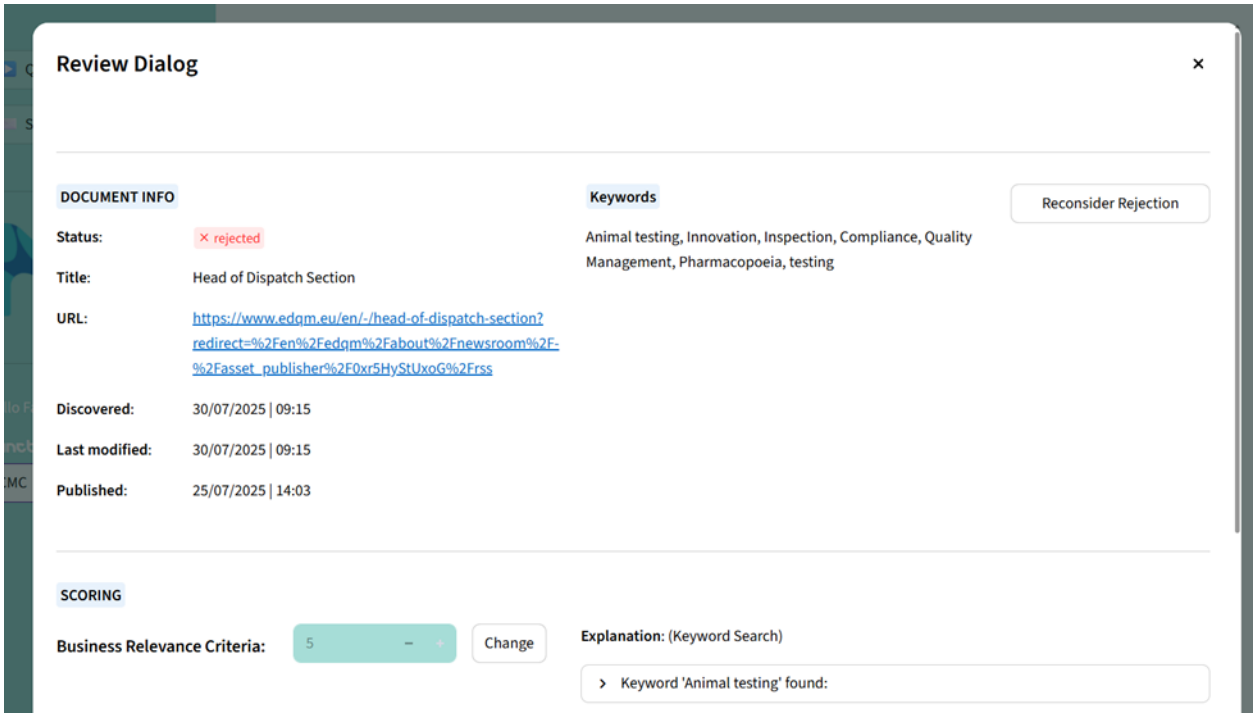

Figure 3: Document review page

The Poseidon Review Dialog interface shown in Figure 3 demonstrates an information-centric design that consolidates all critical regulatory assessment data into a single view optimized for efficient decision-making. The Document Info section presents essential metadata in a clean, structured format, displaying the document's status with indicators, such as the "new," "ready for review," "reviewed," and "rejected" label, followed by the document title that immediately identifies the regulatory content scope.

The clickable URL provides direct access to the source document, as explained in the previous section, while the temporal data fields—Discovered, Last Modified, and Published—offer complete chronological context with precise timestamps that enable regulatory professionals to assess document currency, processing timelines and urgency at a glance.

The Keywords section lists relevant tags which facilitate rapid content categorization and relevance assessment. The Scoring section introduces intelligent automation through the before mentioned scoring system, displaying a numerical score from 1 to 5 with the possibly of manual correction, giving the user the possibility to affect the ML model's behavior by guiding it towards better results. The Explanation field provides transparency into the algorithmic decision-making process.

In addition, hierarchical user management was developed addressing organizational structures typical of regulatory environments. We have implemented knowledge managers for different organizational functions, recognizing that regulatory oversight often follows specific reporting relationships. For example, the right to change automated scoring is reserved to knowledge managers, while still enabling regulatory screeners to see its results.

To conclude, Poseidon's notification and subscription system allows users to configure alerts based on multiple criteria—topic, document type, source, etc., and combinations thereof. This flexibility addresses diverse information needs across different regulatory functions as well as user groups, with filtering capabilities that help users focus on their most relevant content.

CHALLENGES

During the implementation of Project Poseidon we faced several challenges which we want to describe in more detail in this section.

Since several business units will be using Poseidon, the system must accommodate significantly different operational needs across the organization. Each unit monitors distinct HA websites, applies unique logic to calculate scoring metrics, and requires different alert thresholds based on their specific regulatory responsibilities. The challenge becomes more complex when considering that some teams demand full transparency in algorithmic decision-making for compliance purposes, while the workflow of other teams requires advanced machine learning capabilities, even though such models are inherently less interpretable. Successfully deploying Poseidon enterprise-wide requires balancing these competing demands through flexible system architecture that can adapt to diverse business

requirements without compromising core functionality.

The definition of a document duplicate is highly ambiguous. Besides different formatting, there could be slight changes in the semantic meaning of the text. Duplication detection gets even more difficult when you consider that URLs might change over time, meaning that a duplicate could be found on different URLs as well as different documents appearing behind an already seen URL. Our solution to this problem is a step-based duplicate detection approach, where each step poses stronger criteria on duplicate detection. Furthermore, this allows for the inclusion of certain business definitions of versions/duplicates.

Some of the regulatory news that can be found on the web are very lengthy and do not fit completely into the context window of state-of-the-art LLMs. Since comparing documents is a strong requirement for Poseidon we circumvented this issue by chunking the documents, storing their embeddings in a Vector database, as well as using document summaries as intermediate artifacts that can be used for preliminary analysis.

CONCLUSION

In conclusion, Project Poseidon marks a pivotal step forward in transforming regulatory surveillance at Merck KGaA by effectively addressing the limitations of traditional manual screening methods. Through the integration of AI/ML techniques within a robust, cloud-based microservices architecture, the system delivers enhanced efficiency, scalability, and accuracy in monitoring and processing Health Authority updates from diverse international sources. Key functionalities such as automated content crawling, intelligent triage with customizable scoring, semantic analysis, and comprehensive version and duplicate handling ensure that critical regulatory information is identified and prioritized promptly, while preserving the interpretability and traceability essential for regulatory decision-making.

The modular and extensible design of Poseidon not only supports current regulatory compliance needs but also lays a strong foundation for future expansion across multiple departments and organizational functions, enabling consistent and tailored regulatory intelligence delivery company wide. Early implementation results indicate a significant reduction in manual screening time, allowing subject matter experts to reallocate their efforts toward more strategic, value-added activities such as advocacy and policy analysis. Additionally, the system's collaborative features and user-centric interface foster effective teamwork and continuous improvement through user feedback loops, further enhancing its accuracy and usability.

Overall, project Poseidon exemplifies Merck KGaA's commitment to innovation and operational excellence in the pharmaceutical industry by leveraging cutting-edge technology to meet evolving regulatory challenges. This initiative not only strengthens compliance capabilities but also supports agile and informed decision-making, ultimately contributing to better business outcomes and sustained competitive advantage in a complex and dynamic regulatory environment.

REFERENCES

1. Liu, K., et al. (2023). "Landscape Analysis of the Application of Artificial Intelligence and Machine Learning in Regulatory Submissions for Drug Development From 2016 to 2021." *Clinical Pharmacology & Therapeutics*

2. Boopathi, L.K., Raja, N. (2024). "Artificial Intelligence and Machine Learning: Enhancing the Future of Regulatory Affairs." *International Journal of Drug Regulatory Affairs*

3. Fu, L., et al. (2025). "The applications and advances of artificial intelligence in drug regulation: A global perspective." *Acta Pharmaceutica Sinica B*, 15(1), 1-14

4. Streamlit Documentation: https://docs.streamlit.io

5. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830

6. PostgreSQL Global Development Group. "PostgreSQL Documentation" https://www.postgresql.org/docs/

7. Scrapy Documentation: https://docs.scrapy.org

ACKNOWLEDGMENTS

We would like to sincerely thank all Merck KGaA and HMS companies involved for their invaluable contributions and the excellent collaboration throughout this project. Our companies' culture made this achievement possible, enabling us to test, fail and improve to achieve innovative solutions. Our gratitude also goes to all our colleagues who dedicated their time, expertise, and effort, making this joint endeavor a great success.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:
Author Name: Maria Giovanna Zilli
Company: Merck KGaA
Address: Chemin des champs Neuf, 2
Work Phone: +41 793572417
Email: maria-giovanna.zilli@merckgroup.com
Website: www.merckgroup.com

Author Name: Fabian Wahren
Company: HMS Analytical Software GmbH
Address: Grüne Meile 29, 69115 Heidelberg
Work Phone: 0622160510
Email: fabian.wahren@analytical-software.de
Website: www.analytical-software.de