Ref. Ares(2024)2334381 - 28/03/2024

Funded by the European Union

D6.1 HOLiFOOD infrastructure & demonstrators deployment report

Manos Karvounis
28/03/24

D6.1

HOLiFOOD infrastructure & demonstrators deployment report

Document ID	D6.1
Due date	31/03/24
Submission date	28/03/24
Dissemination	Public
Work Package	6
Author(s)	Manos Karvounis, William o' Sullivan, Ali Hurriyetoglu
Document Version	0.4

Grant Agreement	101059813	Duration	September 2026
Start Date	October 2022	End Date	48 Months

Contributors

Name	Organisation
Manos Karvounis	Agroknow
Marilena Dimitrakopoulou	Agroknow
William O Sullivan	Crème Global

Revision history

Version	Date	Reviewer	Modifications
0.1	21/03/24	Ine Van Der Fels	Revisions
0.2	25/03/24	Giulia Petrollini	Revisions
0.3	26/03/24	Debora Serra	Revisions
0.4	28/03/24	Marilena Dimitrakopoulou	Final document

Contents

Index of tables

Index of figures

1 Subject

Leveraging existing technology services and platforms, Work Package 6 (WP6) endeavors to extend, integrate, and deploy an open, distributed, and interoperable HOLiFOOD infrastructure.

At the core of this infrastructure lies the fusion of data and predictive models from Work Package 1 (WP1) alongside external open-source data and prediction models. WP6 is tasked with developing additional services to facilitate knowledge sharing in scenarios where direct data exchange is challenging, such as through federated learning and encryption.

Task 6.1, led by AGROKNOW in collaboration with partners CREME, WR, UVMB, and BfR, focuses on delineating the technology, data, and stakeholder requirements vital for shaping the HOLiFOOD infrastructure. By actively engaging in horizontal meetings with Work Package 4 (WP4) Living Labs (LLs), Task 6.1 aims to meticulously gather, analyze, and process stakeholders' needs regarding data, models, and computation capabilities. Continuous interactions and iterative revisions with all LLs ensure the ongoing validation and refinement of these requirements as WP4 progresses.

2 Summary

The HOLiFOOD Infrastructure combines AI models with robust data management systems to offer predictive analytics for food safety. It's designed to be adaptable, allowing for continuous data integration and model refinement. The architecture consists of several layers:

1. Data Layer: Handles both structured and unstructured data, utilizing databases for efficient storage and retrieval.
2. AI Management Layer: Hosts and executes AI models, ensuring their readiness for deployment.
3. AI Services Layer: Houses components like the Timeseries & Forecasting Engine, enhancing predictive capabilities.
4. Interface Layer: Facilitates interaction with external systems through various modules.

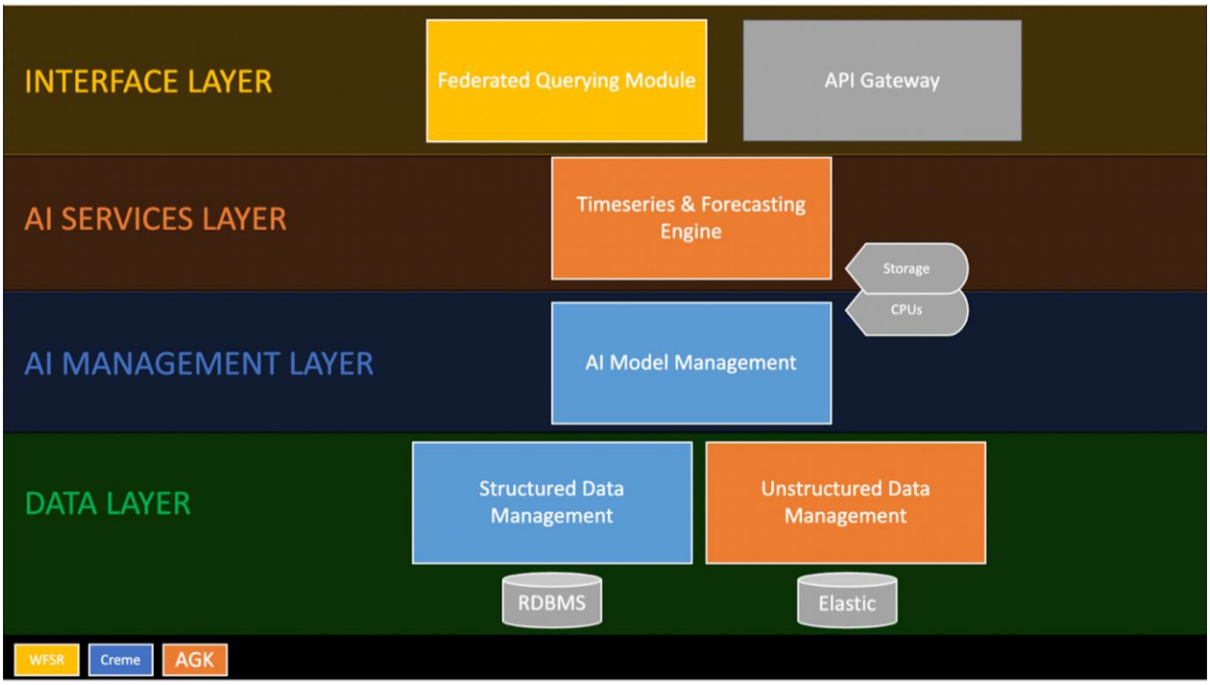
Stakeholder requirements, gathered through internal interviews and events, highlight the importance of high-performance capabilities and data reliability in AI-driven processes.

HOLiFOOD Infrastructure Modules include data management, AI model integration, and forecasting engine components. Data management involves encryption and scraping, employing tools for ingestion, curation, storage, processing, and monitoring. AI model integration offers a platform for hosting models and facilitating collaboration. The forecasting engine analyzes time series data to provide risk assessments and trend forecasts. Overall, the HOLiFOOD infrastructure aims to provide reliable predictive insights crucial for effective food safety management.

3 HOLiFOOD Infrastructure

3.1 HOLiFOOD Infrastructure Conceptual Architecture

The HOLiFOOD infrastructure is an integrated platform where AI models are seamlessly combined with robust data management systems to provide predictive analytics for food safety. This infrastructure is agile, allowing for the incorporation of additional unstructured data and continuous model retraining, which is crucial for maintaining up-to-date and accurate predictive capabilities. The figure below presents a conceptual diagram of the HOLiFOOD infrastructure architecture.



The HOLiFOOD infrastructure depicted is a multilayered AI-powered system designed to enhance food safety measures. This system integrates various components across its layers, which are as follows:

- **Data Layer**: This foundational layer manages both structured and unstructured data. Structured Data Management utilizes relational database management systems (RDBMS) for efficient storage and retrieval of structured data from work packages WP1, WP2, and WP3. In contrast, Unstructured Data Management employs Elastic search databases to handle unstructured data such as lab test results and public food safety announcements. This setup facilitates the collection, cleaning, and preparation of data essential for training AI models, particularly those developed in WP1 and WP3.
- **AI Management Layer**: Central to the infrastructure, this layer focuses on AI Model Management. It hosts and executes trained AI models, leveraging the

capabilities of Creme's infrastructure. This layer ensures that the AI models, once trained on data processed at the Data Layer, are readily available for deployment and execution.

- **AI Services Layer**: This layer houses the Timeseries & Forecasting Engine, which is a critical component for predictions. Agroknow's Predictions Engine, which forms part of this layer, employs forecasting algorithms to treat incident data as time series. This enables the AI models from WP1 and WP3 to project future trends and incident numbers based on historical data, thus enhancing predictive capabilities.
- **Interface Layer**: The topmost layer in the infrastructure serves as the point of interaction with external systems. It includes a Federated Querying Module, which allows for a distributed querying mechanism across various data sources, a component that is being discussed for inclusion following WFSR's suggestion in T6.2. Additionally, an API Gateway is present to facilitate secure and managed access to the services and data provided by the lower layers of the infrastructure.

4 Stakeholder requirements

4.1 Methods to elicit Stakeholder Requirements

Task 6.1 coordinates with & participates in the horizontal meetings of the WP4 LLs to collect, analyse and process requirements that stakeholders have in terms of data, models and computation to be able to integrate them into the HOLiFOOD infrastructure througheach pilot. The identified requirements will then be updated and validated through continuous reviews and interactions with all LLs as WP4 activities progress feeding the vertical research.

4.1.1 Internal Stakeholder Requirements

To identify the necessary requirements for the development of the technical HOLiFOOD infrastructure, Agroknow conducted internal interviews with the leaders of Work Package 1 (WUR) and Work Package 2 (CNR). A structured script was prepared for these interviews to gain a comprehensive understanding of the specific needs and requirements. This approach aimed to ensure a systematic and focused exploration of the requirements essential for the successful framework of digital infrastructure for food safety.

4.1.2 External Stakeholders Requirements

Living Labs were held during the Synergy Days event in Thessaloniki on 05/10/23, titled "Novel Digital Infrastructure for Food Safety." The agenda commenced with a warm welcome and introduction, setting the stage for an insightful session. There were more than 20 attendees, including representatives from scientific institutions, innovation SMEs, and technical organizations in the greater food domain.

To foster engagement and break the ice, a Wooclap session featuring icebreaker questions was introduced. Living Lab participants seamlessly accessed the Wooclap website on their smartphones to respond to poll questions. The initial questions (#1-#2) delved into participants' backgrounds and fields of work. During this dynamic session moderated by EUFIC, the facilitator managed an interactive session, with questions and results displayed on the main screen, creating a collaborative and engaging atmosphere.

The Living Lab then delved into a series of follow-up questions (#3-#5) on Wooclap, exploring novel digital infrastructure for food safety. This interactive segment aimed to draw insights from the diverse perspectives of the participants, enhancing the richness of the discussions.

As the Living Lab neared its conclusion, the moderator skillfully wrapped up the polls, summarising key observations and conclusions. This paved the way for an open discussion, allowing participants to share comments or questions with their peers and hosts, fostering a collaborative and inclusive environment.

4.2 Results stakeholder Requirements

4.2.1 Internal stakeholder requirements

Table 1 presents the consolidated results of these interviews with the internal stakeholders.

Table 1. Requirements from internal stakeholders

	ERI model	Chemical Hazards	Biological Hazards
Overview	ML models (Bayesian networks, random forest), text mining techniques, automatic anomaly detection models.	Lateral flow immunoassay and photometric reader for mycotoxins and pesticides (glyphosate).	WGS and qPCR for bacteria in poultry, maize, and lentils.
Type of data required	Textual data from news related to food safety, scientific publication databases and social media	Sampling size, sampling strategy, commodity type, calibration data, test portion. Food consumption data would be complemented.	Sampling size, sampling strategy, commodity type, sequencing data.
Type of data generated	Probability of concentration >MRL. For anomaly detection models maybe probability. Visualization of identified weak signals and food safety-related topics	Reference signals, concentration, name of sample.	Name of sample, CFU of pathogens, sequencing.
Storage and access	Secure drive from WUR, and elastic search. Stored in CSV.	Hardware software - portal reader, not a platform, in excel. CNR has a server. However, they will open access to the data from monitoring. They need to investigate the	Sequencing data will be stored on the HOLiFOOD platform and accessible to all partners.

		anonymity of the data, to discuss it with us.	
Functionality	Interactive dashboard, maybe use of search for contaminants or food products or geographical origin.	No need. Maybe in the future users can find samples by searching for sampling commodities and help traceability.	No need. Maybe in the future users can find samples by searching for sampling commodities and help traceability.
Sensitive data handling	Maybe federated learning can be employed as a solution. If it is too advanced for this project, they can train the model locally, and only move the trained model to the HOLiFOOD infrastructure not the data.	They don't have sensitive data. Maybe if above MRL for example, they anonymise the supplier. Usually, they sign agreements with food companies or within projects. At CNR, they have an ethical committee within the institute, and they handle the sensitivity of data.	They don't have sensitive data. Maybe if above MRL for example, they anonymise the supplier. Usually, they sign agreements with food companies or within projects. At CNR, they have an ethical committee within the institute, and they handle the sensitivity of data.
Reliability and reproducibility of results	Test the model with unseen data and check if the output is what they expected to be. They can share the code so other people can train a model for reproducibility.	Validation according to specific guidelines, and analyze reference material, otherwise they use spiked samples. Comparison to the gold standard method and use ISO protocol for validation.	Validation according to specific guidelines, and analyze reference material, otherwise they use spiked samples. Comparison to the gold standard method and use ISO protocol for validation.

4.2.2 External stakeholder requirements

The figures below display the results of the Wooclap votes. At the end, we summarize our understanding of these answers, and relevant discussion that followed, in terms of external stakeholder requirements for the HOLiFOOD infrastructure, and other similar innovative digital infrastructures.

4. What opportunities do you think AI tools present to your organization for food safety? 23 respondents

5. What are the specific technical requirements your organization needs to implement AI tools for food safety? 19 respondents

The strategic choice of integrating our Living Lab with Synergy Days proved to be fruitful, offering a unique opportunity to receive feedback from experts in the field. Synergy Days, recognised as a cornerstone conference connecting digital innovators in the European agri-food sector, provided an ideal backdrop.

The results from the Wooclap session at the Synergy Days event reveal key insights into the expectations and requirements of external stakeholders for the HOLiFOOD infrastructure and similar digital food safety platforms. Here's a structured summary based on the tag clouds and our follow-up discussions with the stakeholders:

Opportunities Presented by AI Tools for Food Safety:

- Stakeholders perceive AI tools to achieve optimisation in food safety.
- There is an expectation for AI to bring about more efficient standardization, improved integration, and cost-effective solutions.
- Enhancing food safety is seen as a priority, with AI providing the potential for better prediction and prevention rather than reaction to incidents.
- The stakeholders also recognize the value in transparency, faster outcomes, and the potential for AI to contribute to better marketing of produce and timely risk identification.

Specific Technical Requirements for Implementing AI Tools:

- There is a clear demand for high-performance computing infrastructure and data digitalazation to manage and analyze vast amounts of data.
- Qualified staff with training in AI and data handling, indicating a need for capacity building within organizations.

- Infrastructure should facilitate data sharing platforms with appropriate measures for confidentiality and privacy information handling.
- Computing capacity, indicating the need for a robust technical backend to support AI functionalities.

Challenges in Implementing AI for Food Safety:

- A significant challenge highlighted is the digital resources and data availability, pointing to a gap in the necessary technical assets or datasets required to implement AI effectively.
- Financial constraints and a lack of qualified staff are seen as barriers, suggesting a need for investment in both human and technological capital.
- There is an acknowledgment of insufficient data and the need for standardization across different data types and sources.
- Challenges around the sensitivity of data, trust, and understanding AI indicate concerns about data privacy, ethical use, and the need for education around AI's capabilities and limitations.

In summary, stakeholders from the Synergy Days event have identified a range of opportunities that AI presents for food safety, along with technical and operational requirements necessary for its implementation. They also acknowledge a number of challenges that need to be addressed to fully leverage AI in this domain. For HOLiFOOD and similar infrastructures to be successful, they must be designed with high-performance capabilities, incorporate extensive training and capacity building for staff, and include mechanisms to address data availability, standardization, and sensitive data handling while ensuring reliability and trust in AI-driven processes.

5 HOLiFOOD Infrastructure Modules

This section reports on the ongoing work in Task 6.2 which includes the following activities: the development of services for data & knowledge integration/publication (AGROKNOW), services for predictive models' integration/publication (CREME), the deployment of an HPC Infrastructure for hosting & computation (CREME), the creation of a shared encrypted database with private/confidential data (CREME), as well as additional services such as federated querying & learning services (WR).

5.1 Data Layer

Private Data Management & Data Encryption

AWS provides a comprehensive, scalable, and highly secure cloud computing environment. Creme Global's use of AWS S3 for storing HOLiFOOD-related data benefits from AWS's physical security measures, which include multiple layers of operational and physical security to ensure the integrity and safety of data. AWS data centers are staffed 24/7 by security guards, and access is strictly controlled both at the perimeter and at building ingress points by professional security staff utilising video surveillance, intrusion detection systems, and other electronic means.
AWS S3 supports data encryption both at rest and in transit. For HOLiFOOD-related data stored on AWS S3, Creme Global leverages:

At Rest: AWS S3 provides options to encrypt data using server-side encryption with Amazon S3-managed keys (SSE-S3), AWS Key Management Service (KMS) keys (SSE-KMS), or client-side encryption with keys managed by the customer. This ensures that data is unreadable to unauthorized users.

In Transit: AWS uses SSL/TLS to encrypt data in transit between AWS hosted services and AWS S3, ensuring that data is secure as it travels over the internet.
Creme Global utilizes AWS Identity and Access Management (IAM) to control access to AWS resources, including S3 buckets. IAM policies and roles can be defined to enforce who can access specific data within S3, under what conditions, and with what permissions. Multi-factor authentication (MFA) is also required for access, adding an additional layer of security.

By storing data in AWS S3 within the EU, Creme Global ensures compliance with European data protection regulations, such as the General Data Protection Regulation (GDPR). AWS complies with a wide range of industry-specific standards and certifications, including ISO 27001, which helps to ensure that data is managed securely and in compliance with regulatory requirements.

Public Data Management & Data Scraping

This section presents the internal architecture of the Data Management & Scrapping Platform for public data (i.e., non-sensitive information). This Data Management Platform is used for the storage of data coming both internally from HOLiFOOD partners and external stakeholders and from the web at-large, especially from public food safety authority web sites.

Motivation

The escalation of product recalls and border rejections is a significant concern in the food industry, with reports indicating a two-fold increase from 2014 to 2019 based on data from national food safety agencies. Figure 1 encapsulates the trend in the number of recalls and rejections as recorded by these authorities.

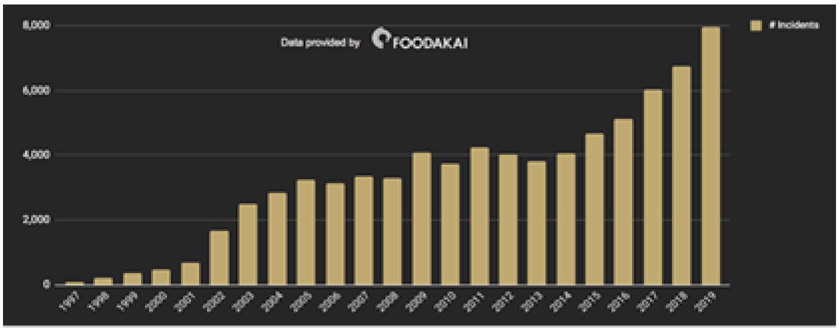


Figure 1: Number of recalls and border rejections reported by National Authorities

The implementation of stringent food safety regulations such as the Food Safety Modernization Act (FSMA) and the guidelines set forth by the Global Food Safety Initiative (GFSI) necessitates comprehensive risk evaluations encompassing supplier, facility, and product integrity. Superficial assessments are no longer adequate in meeting these rigorous standards.

There's a growing need for a sophisticated risk assessment framework, one that adapts to changing consumer demands and the intricate nature of the international food distribution networks. This framework must leverage real-time intelligence on supply chain disruptions, current information on microbiological and chemical hazards, and predictive models. It should also integrate data pertaining to population demographics, consumption patterns, and the various processing conditions such as heat treatment and storage. The objective is to provide practical guidance for risk reduction.

Risk assessments are often relegated to a routine procedural task, overshadowed by the more urgent compliance requirements. This results in the misallocation of resources, with an overemphasis on unlikely scenarios, thereby diminishing the capability of these assessments to pinpoint potential emerging hazards. Figure 2 illustrates the different stages of the food supply chain and the relevant data points that contribute to a comprehensive risk assessment.

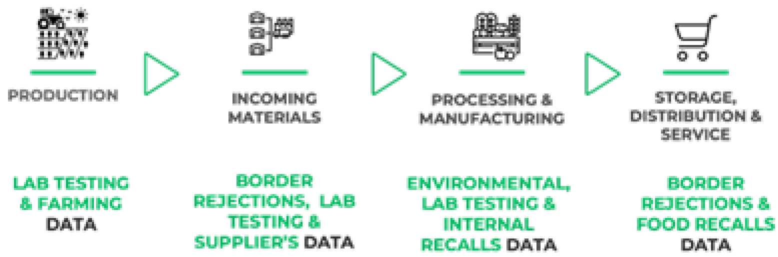


Figure 2: Stages of the supply chain and corresponding data that can be used for risk assessment

Risk management in the food industry is intricate, demanding information gathering from every point in the supply chain—right from production, which includes laboratory test results and agricultural practices, to sourcing, processing, storage, and distribution. Data from the production stage can highlight early signs of chemical and biological contaminants. Information on incoming materials, such as results from border inspections and audit findings, is critical in evaluating the safety of ingredients before their integration into production. Processing data is gathered from laboratory analyses and environmental monitoring, while storage and distribution figures relate to retail oversight and product recall incidents. Market factors, like fluctuating commodity prices and the stability of the country of origin, also bear significance.

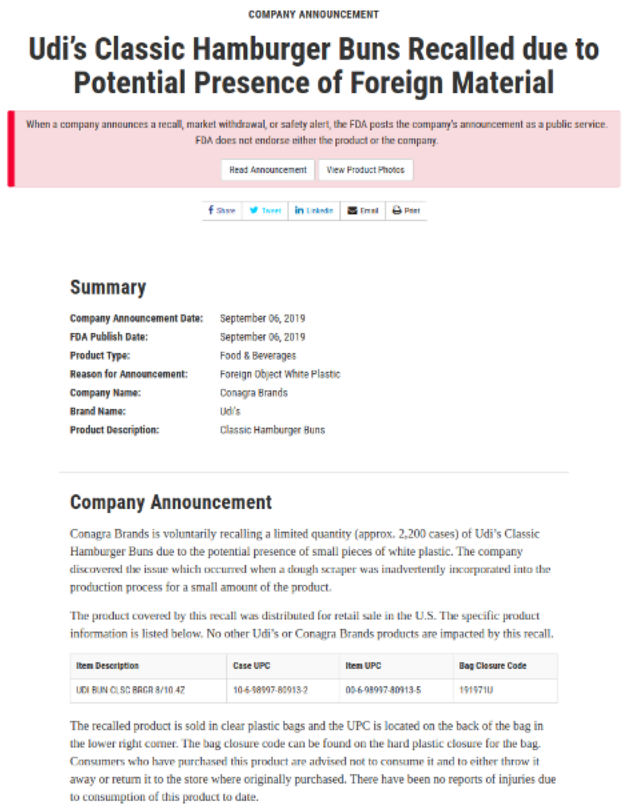
The real challenge is in the assimilation and interpretation of this fragmented, varied, and often multilingual data to construct an efficient risk assessment model. This will be further explored in subsequent sections that discuss the architecture and capabilities of our Advanced Big Data Platform.

To ensure food safety, industry must evolve from a state of passive defense to active hazard management. This transformation is contingent upon the harmonisation of modern analytical tools and methodologies with comprehensive food safety data to facilitate prompt and accurate risk assessments.

From Unstructured Data to Machine Readable Information

The foundation of any tech-driven risk assessment strategy is the efficient collection and refinement of data from extensive global food supply networks. To address this, we have developed a sophisticated Data Platform that incorporates a detailed procedure for managing such data. This system harnesses both computational resources and domain-specific knowledge to proficiently navigate and refine large volumes of data. Built upon a microservices architecture, our platform is adept at gathering, processing, indexing, and disseminating data across the diverse spectrum of the food and agricultural sectors. Each piece of data is subjected to an automated enhancement process and scrutinised by experts to ensure the highest standards of quality.

Our goal is to start with food safety incident announcements, usually originating as textual data in the web (an example in Figure 3), and turn into structured data records, following the schema in Table 2.



COMPANY ANNOUNCEMENT

Udi's Classic Hamburger Buns Recalled due to Potential Presence of Foreign Material

When a company announces a recall, market withdrawal, or safety alert, the FDA posts the company's announcement as a public service. FDA does not endorse either the product or the company.

Read Announcement | View Product Photos

Share | Tweet | LinkedIn | Email | Print

Summary

Company Announcement Date:	September 06, 2019
FDA Publish Date:	September 06, 2019
Product Type:	Food & Beverages
Reason for Announcement:	Foreign Object White Plastic
Company Name:	Conagra Brands
Brand Name:	Udi's
Product Description:	Classic Hamburger Buns

Company Announcement

Conagra Brands is voluntarily recalling a limited quantity (approx. 2,200 cases) of Udi's Classic Hamburger Buns due to the potential presence of small pieces of white plastic. The company discovered the issue which occurred when a dough scraper was inadvertently incorporated into the production process for a small amount of the product.

The product covered by this recall was distributed for retail sale in the U.S. The specific product information is listed below. No other Udi's or Conagra Brands products are impacted by this recall.

Item Description	Case UPC	Item UPC	Bag Closure Code
UDI BUN CLSC BRGR 8/10.4Z	10-6-98997-80913-2	00-6-98997-80913-5	191971U

The recalled product is sold in clear plastic bags and the UPC is located on the back of the bag in the lower right corner. The bag closure code can be found on the hard plastic closure for the bag. Consumers who have purchased this product are advised not to consume it and to either throw it away or return it to the store where originally purchased. There have been no reports of injuries due to consumption of this product to date.

Figure 3: An example of a food safety incident as announced in a public food safety authority website

Table 2. The main attributes of the EFRA schema for food safety incidents

Data Field	Description
Ingredient	Main ingredient of the recalled brand
Hazard	Reason behind the recall
Supplier	Food company(ies) involved in the recall
Brand	Product brand that was recalled
Date	Date of the recall
Origin	Origin of the recalled ingredient

Architecture and Technical Components

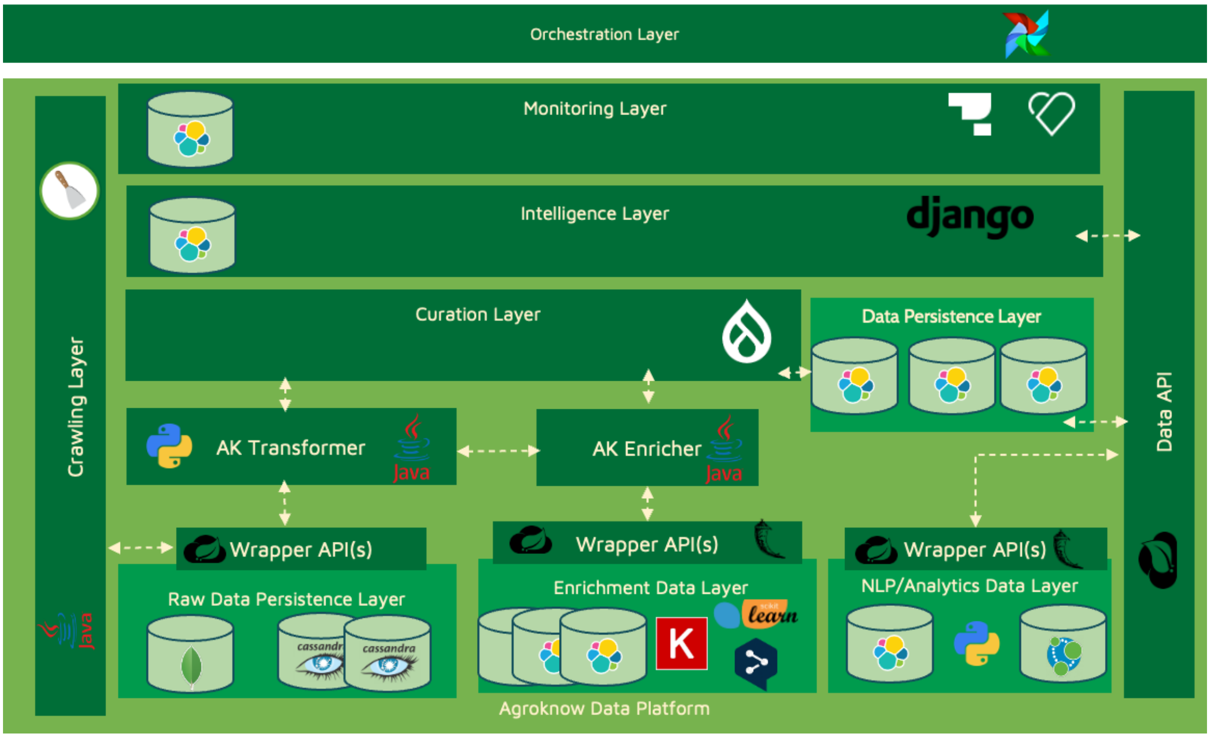


Figure 4: Data Management & Scrapping Platform

The Data Management & Scrapping Platform's goal is to seamlessly gather, process, index, and broadcast a diverse array of food safety data originating from public food safety authority sources, or gathered internally in the HOLiFOOD consortium, or by external stakeholders.

Platform Layers:

- **Crawling Layer**: This is the frontline of data acquisition. It is dedicated to the aggregation of raw, unprocessed data.
- **Raw Data Persistence Layer**: As the name suggests, this layer is where all the acquired raw data is securely stored.
- **Transformer Layer**: Here, the raw data undergoes transformation to align with an internal schema, ensuring uniformity. Additionally, each piece of data is assigned a unique identifier.
- **Enrichment Layer**: This component takes the transformed data and augments it with crucial details such as the origin of products, the associated hazards, and more.
- **Enrichment Data Layer**: This is the hub for all tools, technologies, and frameworks used by the Enrichment Layer.
- **Curation Layer**: An internal Content Management System (CMS), it facilitates manual annotations on food safety data records, ensuring quality and relevance.
- **Data Persistence Layer**: A repository for all production-ready data, it feeds both the application layer and internal ETL workflows.
- **Intelligence Layer**: This is the analytical core of the platform, equipped to calculate predictive outputs and undertake risk assessments.
- **Data API Layer**: A bridge to the outer world, it grants external applications access to the data via JSON endpoints.
- **NLP/Analytics Layer**: This component enhances the Data API by boosting its free-text search capabilities, making data retrieval more intuitive.
- **Monitoring Layer**: A sentinel that ensures all storage engines and API endpoints of the platform are running optimally.
- **Orchestration Layer**: This coordinates and manages the execution of all ETL workflows, ensuring data fluidity across the platform.

Technical Components Overview:

- **Data Ingestion & Collection Components**: Big Data Ingestion is the platform's mechanism for building comprehensive data pipelines. With the challenge of handling varied data sources ranging from images to JSON files, the ingestion layer uses a suite of tools like Scrapy, Python scripts, and a custom Java project to collate data. The raw data is subsequently harvested and transformed using both Python and Java.

- **Data Curation Component**: The harmony between automation and human expertise is beautifully manifested here. Human experts play a pivotal role, ensuring data accuracy by overseeing automated processes, especially during the mapping and enrichment processes of suppliers and ingredients.
- **Storage Components**: Focusing on performance, this layer employs schema-less persistence technologies to optimise data storage and retrieval operations.
- **Data Processing Components**: The platform relies on Flask, a Python-based micro web framework, for wrapping machines and deep learning models. Furthermore, Django is employed for the development of APIs, concentrating on incident prediction and risk assessment.
- **Monitoring Components**: Ensuring the health and uptime of the platform is paramount. Tools like Logstash, Kibana, and MetricBeat over the Elastic stack offer real-time insights, analytics, and visualiszation into the platform's performance and health.

5.2 AI Management Layer

AI Model Integration & Publication

The Creme Global Platform, historically combining the strengths of two previous services (Expert Models and Data Foundry), offers a robust infrastructure for integrating and hosting AI models, specifically tailored to the needs of data-driven projects across various industries, including food safety, nutrition, cosmetics, and agriculture. These services are designed to streamline the process of managing complex datasets and integrating predictive models, thus enabling a more efficient and accurate analysis of data.

The Creme Global Platform

The modelling services provided by Creme Global provides a scalable and secure environment for hosting and running AI models. This service allows users to:

- Integrate various predictive models seamlessly into the platform. These models can range from simple linear regressions to complex deep learning models, making it versatile for different types of data analysis and prediction tasks.
- Automate the execution of these models, facilitating the generation of output results. This automation is crucial for projects requiring regular data updates or real-time analysis, ensuring that the latest data feeds directly into the models for up-to-date insights.
- Visualise model outputs through specially designed dashboards. These dashboards are crafted to present complex data and analysis results in an easily understandable format, enabling stakeholders to make informed decisions based on the latest insights.

The data management services complement the modelling services by providing a comprehensive data management solution. This service facilitates:

- Secure data storage with robust encryption methods, ensuring that sensitive or private data is securely managed throughout its lifecycle on the platform.
- Efficient data integration from multiple sources, allowing for a more comprehensive dataset that can enhance the accuracy and reliability of model predictions.
- Collaboration and data sharing among partners, enabling the exchange of model output data and insights in a secure and controlled environment. This feature is particularly beneficial for projects involving multiple stakeholders or requiring input from various data sources.

Integration and Collaboration

For projects like WP1, where AI model integration plays a crucial role, the Creme Global platform offers a streamlined process for both hosting models and sharing their outputs. Partners can choose to host their models directly on the platform, leveraging the automated execution and visualization capabilities, or share relevant model output data for integration. This flexibility supports a collaborative approach, enabling various partners to contribute to and benefit from the collective insights generated by the integrated models.

Moreover, the emphasis on security, data privacy, and compliance ensures that all operations within the platform meet the highest standards, which is especially critical for projects handling sensitive or confidential data.
In summary, the Creme Global platform provides a comprehensive suite of tools for AI model integration, data management, and visualisation. These services not only enhance the efficiency and accuracy of predictive modelling tasks but also foster collaboration among partners, all within a secure and compliant framework.

5.3 AI Services Layer

Time series Management & Forecasting Engine

This section describes Agroknow's Predictions Engine, which employs forecasting algorithms to treat incident data as time series. This enables the AI models from WP1 and WP3 to project future trends and incident numbers based on historical data, thus enhancing predictive capabilities.

The Time Series Management & Forecasting Engine is a sophisticated software system designed to enhance the predictive capabilities within the HOLiFOOD platform, particularly in the realm of Trends Forecasting. This Engine stands as a pivotal component in managing and analyzing time series data related to food safety, focusing on various elements like ingredients, hazards, and their combinations, as well as geographical origins.

The core functionality of this Engine is to generate predictions over a span of one year for a wide range of categories, including individual ingredients, combinations of ingredients with hazards, and hazards with origins. Additionally, it provides a nuanced risk scoring mechanism, offering both historical and future perspectives on the risk associated with specific ingredients against top hazards.

This Engine adopts a comprehensive approach, encompassing multiple stages such as data preparation, model training, model evaluation, prediction generation, and risk scoring. Each stage plays a crucial role in ensuring the accuracy and reliability of the forecasts, making it a vital tool for managing and predicting food safety incidents. The Engine's architecture facilitates batch predictions, storing them in persistent storage for on-demand access, aligning with the architectural decision to prioritise batch processing over online inference.

Overall, the Time Series Management & Forecasting Engine serves as a cornerstone for HOLiFOOD, providing robust and reliable predictive insights essential for managing food safety risks effectively.

Methodology

The Time Series Management & Forecasting Engine, integral to the HOLiFOOD platform, employs a structured methodology to provide advanced predictive capabilities in food safety. The Engine's primary function is to forecast trends and assess risks in the domain of food safety by analysing time series data associated with ingredients, hazards, origins, and their interplay.

The Engine's methodology is structured into several distinct components, each contributing to the overall predictive process: Data Preparation, Model Training, Model Evaluation, Prediction, Risk Scoring.

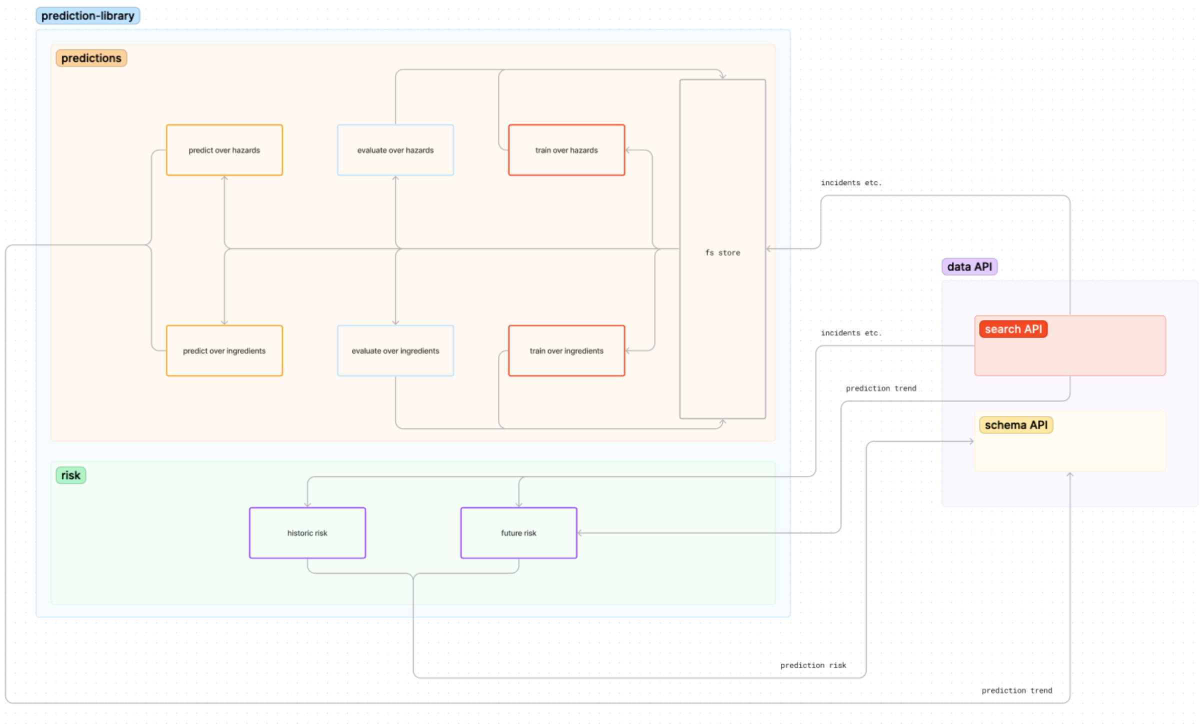


Figure 5: The Time Series Management and Forecasting Engine data flows

<u>Data Preparation</u>

Data preparation involves collecting historical data related to food safety incidents from the Data Platform. This data includes incidents like border rejections and recalls, categorised by ingredient, hazard, and country of origin. The incidents are grouped monthly, forming a pool of time series for each category.

<u>Model Training</u>

For model training, time series with at least three years of data and ten incidents are selected. Various forecasting models are employed, each suited to specific time series profiles. The Engine supports a suite of models rather than a single model to accommodate the diverse nature of the data:

- **Prophet Model**: Used for time series with more than 200 incidents, Prophet is effective for data with strong seasonal trends and handles irregularities well.
- **Ridge Regression**: Applied to time series with fewer than 200 incidents, Ridge Regression is used for its robustness in linear modeling with regularisation.

More specifically, model training involves the following sub-steps and considerations:

- **Selection of Trainable Time Series**: Before training begins, time series data undergoes a filtering process to ensure only quality data is used. The criteria for selection are: (a) a minimum of three years of historical data, and (b) at least ten individual incidents recorded in the time series. These criteria help in selecting time series that provide sufficient information for meaningful analysis and prediction.
- **Categorization of Time Series Data**: The Engine recognizes the distinct nature of each time series. They are categorized based on their 'profile', such as: (a) the density of incidents (sparse vs. dense data), (b) the length of the time series (short-term vs. long-term data). This categorization is vital for selecting the appropriate forecasting model for each time series.
- **Forecasting Model Selection**: Based on the categorized profiles, specific forecasting models are chosen for each time series: (a) Prophet Model, ideal for time series with more than 200 incidents and strong seasonal trends. Prophet's ability to handle non-linear trends, missing data, and outliers makes it suitable for complex and dense time series, (b) Ridge Regression Model, suited for time series with fewer than 200 incidents. Ridge Regression's strength lies in its linear least squares function and regularisation, which is effective for sparser time series.
- **Customisation of Forecasting Models**: The Engine does not use a one-size-fits-all approach. Each model is customized based on the specific characteristics of the time series it will be applied to. This customization ensures that the models are finely tuned to capture the unique patterns and trends of each dataset.
- **Training Process**: During training, the models are fed with the prepared time series data. The Engine ensures that each model learns from the patterns, trends, and anomalies present in the historical data. This process involves iterative adjustments and refinements to improve the model's accuracy and reliability.
- **Addressing Overfitting and Underfitting**: Special attention is given to prevent overfitting, where a model becomes too tailored to the training data and fails to generalise. Similarly, underfitting, where a model is too simplistic to capture the data's complexity, is also avoided. Regularisation techniques, cross-validation, and other statistical methods are employed to maintain a balance.
- **Model Performance Monitoring**: Throughout the training process, the performance of each model is continuously monitored. This involves assessing how well the model predicts known outcomes within the training data. Adjustments are made as needed to improve performance, ensuring that each model reaches its optimal predictive capability.

<u>Model Evaluation</u>

The Engine employs a rolling window cross-validation scheme for model evaluation. This approach is particularly suitable for time series forecasting as it involves training the model on a sequence of historical data and testing it on the subsequent data window. This mimics how the model will be used in real-world scenarios and provides a realistic assessment of its performance.

sMAPE (Symmetric Mean Absolute Percentage Error) is pivotal in evaluating the models. sMAPE is advantageous because it is scale-independent and provides a clear understanding of the percentage errors, making it easier to interpret the accuracy across different time series. It is defined in the range [0, 2], where values closer to 0 indicate higher accuracy.

While sMAPE is the primary metric, the Engine also considers a custom accuracy metric. This metric, distinct from sMAPE, is based on the ratio of the sum of predicted values to the sum of actual values, offering a different perspective on model accuracy.

The Engine's methodology recognises and addresses the flaws in relying solely on summary statistics like the sum of predicted versus actual values. Such approaches can sometimes yield misleading accuracy figures, especially if the model predictions are not temporally aligned with actual events.

Finally, models are not evaluated in isolation. Comparative analysis is conducted to benchmark the performance of different models against each other. This helps in understanding the relative strengths and weaknesses of each model.

<u>Model Tuning</u>

Automated hyperparameter tuning is conducted to optimize each model. Different settings are used for Prophet and Ridge, with the objective function being the minimization of the sMAPE score. The tuning process results in an `hparams.json` file containing the optimal parameters for each model. Key aspects of the model tuning process include:

- **Hyperparameter Optimization**: Utilises automated hyperparameter tuning to identify the most effective settings for each model. Employs an 'offline mode' approach, where the tuning process is conducted independently of the model's operational environment.

- **Prophet Model Tuning**: MAX_EVALS set to 100, indicating the number of hyperparameter combinations to be tested. Objective function based on minimizing the sMAPE score, ensuring a focus on improving forecast accuracy. Utilizes the Tree of Parzen Estimators (TPE) from the hyperopt library for optimization.
- **Ridge Regression Model Tuning**: MAX_EVALS increased to 150, providing a wider range of parameter combinations due to the linear nature of the model. Similar to Prophet, the objective function is to minimise the sMAPE

score. Employs the same TPE algorithm for efficient hyperparameter space exploration.

- **Output of Tuning Process**: The tuning process generates an hparams.json file, which contains the optimal hyperparameters for each model. The structure of this file is designed for ease of interpretation and implementation in the training process.
- **Empirical Validation**: Once optimal hyperparameters are identified, they undergo an empirical validation process. This step ensures that the new parameters indeed enhance the model's performance based on predefined Key Performance Indicators (KPIs).
- **Updating Model Parameters**: After validation, existing models are updated with new hyperparameters to improve their forecasting capabilities. This update process is integral to maintaining the relevance and accuracy of the models over time.

Forecasting & Trend Calculation

The forecasting and trend calculation stages are pivotal in the Time Series Management & Forecasting Engine, translating the refined models into actionable insights. These stages involve generating future predictions and analyzsing trends from both historical and forecasted data.

Concerning forecasting, the following key points are observed:

- **Time Horizon**: The Engine generates forecasts for up to one year ahead, providing a comprehensive view of potential future incidents.
- **Data Utilization**: Trained models use the latest available data to predict future incidents for each month in the forecast period.
- **Model Application**: Each time series is forecasted using its designated model, optimized through the model tuning process.
- **Forecast Granularity**: Forecasts are detailed, providing month-by-month predictions to capture seasonal and short-term variations.
- **Output Integration**: The forecasts are integrated into the Data Platform, making them accessible for further analysis and decision-making within the HOLiFOOD platform.
- **Realism and Practicality:** Forecasts focus on practical utility, aiming to provide realistic estimates that can inform proactive food safety measures.

Concerning trend calculation, the following key points are observed:

- **Simple Moving Average (SMA)**: The Engine employs SMA for calculating trends over different time windows (6, 12, and 18 months), providing a smoothed perspective on the data.

- **Historical and Predicted Data**: Trends are calculated using both historical incidents and forecasted data, offering a comprehensive view of the evolving risk landscape.
- **Trend Direction Analysis**: For a 12-month window, the Engine assesses whether the trend is increasing, decreasing, or unchanged, offering insights into the direction of food safety risks.
- **Trend Communication**: Both the trend value and direction are communicated to the Data Platform API, integrating these insights into the Trends Forecasting module of HOLiFOOD.
- **Contextual Relevance**: By analysing trends over multiple timeframes, the Engine provides contextually relevant insights, allowing for more nuanced understanding and planning.

- **Adaptive Trend Analysis**: The trend analysis adapts to new data, ensuring that trends reflect the most current information and forecasts.

5.4 Interface Layer

API Gateway

This requires further discussion - this is understood to be outside of scope for Creme Global's involvement in HOLiFOOD.

Federated Querying & Learning Services

The platform consists of multiple open and private databases. Benefiting from a combination of these databases is crucial for early warning. With this aim in mind, we will develop services such as federated querying and federated learning that utilise multiple databases for generating a result.

First a federated querying system will enable retrieval and harmonise records from multiple databases. These results will be presented in three different ways that are i) grouping results by source, ii) ranking results by relevance, iii) combining results in case records are complementary. This service will be implemented only on open datasets.

Next, federated learning will be applied on both open and private data that is incorporated in the platform. This is a machine learning method that creates a predictive model in a privacy-preserving manner. Aggregated information is shared with the central model from each data provider.

The roadmap for implementation of the federated learning services is planned to be more concrete in D6.2 and the implementation of the roadmap that will yield running services will be delivered in the scope of D6.3.

6. Conclusions and recommendations

This digital infrastructure, epitomized by the HOLiFOOD framework, represents a collaborative effort to harness the power of technology in safeguarding public health and ensuring the integrity of our food supply chain. By seamlessly integrating data and predictive models from various sources, including Work Package 1 and external open-source repositories, the HOLiFOOD infrastructure provides a comprehensive solution for identifying, assessing, and mitigating emerging food safety issues. The platform's open, distributed, and interoperable nature not only enhances accessibility but also fosters collaboration among stakeholders across Europe and beyond. Moreover, the development of additional services to facilitate knowledge sharing, particularly in scenarios where direct data exchange is challenging, underscores the platform's commitment to inclusivity and innovation. Through initiatives like federated learning and encryption, the HOLiFOOD infrastructure empowers stakeholders to collaborate effectively while safeguarding sensitive information. In essence, the integrated European data and knowledge exchange platform represents a testament to the power of technology in addressing complex challenges facing the food industry.

holi food

Funded by
the European Union