EFRA
EXTREME FOOD RISK ANALYTICS

Extreme Food Risk Analytics

D1.1: EFRA Requirements Roadmap

Lead Authors: Ali Hürriyetoğlu (WFSR), Bas van der Velden (WFSR)

Funded by
the European Union

Grant Agreement No.	101093026		
Project Acronym	EFRA		
Project Title	Extreme Food Risk Analytics		
Type of action	HORIZON Research and Innovation Actions		
Call Topic	HORIZON-CL4-2022-DATA-01-05		
Project Start Date	January 1st, 2023	Project End Date	December 31st, 2025
Project URL	efraproject.eu		
Work Package	WP1 | WP Requirements		
WP Lead Beneficiary	STICHTING WAGENINGEN RESEARCH (WR)		
Deliverable type[1] | Dissemination level[2]	Report | Sensitive		
Contractual due date	30 September 2023	Actual submission date	29 September 2023
Lead Author (s)	Ali Hürriyetoğlu (WFSR), Bas van der Velden (WFSR)		
Contributors	Mary Godec (WFSR), Zuzanna Fendor (WFSR), Irene Benedetto (MAIZE), Alessio Bosca (MAIZE), Milad Botros (MAIZE), Francesco Tarasconi (MAIZE), Marco Trevisan (MAIZE), Raffaella Ventaglio (MAIZE), Salvatore Trani (CNR), Francesco Lettich (CNR) Morgane Rumeau (AGRIVI), Jakub Janostik (SGS), Manos Karvounis (Agroknow), Marilena Dimitrakopoulou (Agroknow)		
Internal reviewer(s)	Vasilis Kotsikoris (RAINNO), Eleni Stogiannou (RAINNO), Ourania Ntinou (RAINNO)		

[1]R: Document, report; DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc; DMP: Data management plan; ETHICS: Deliverables related to ethics issues; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2] PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C - EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete	Changes	Contributor(s)
V0.1	01/08/2023	10	Initial Deliverable Structure	Manos Karvounis (Agroknow)
V0.7	18/09/2023	70	Draft content is complete	Ali Hürriyetoğlu (WFSR) Bas van der Velden (WFSR)
V0.8	19/09/2023	75	Internal Peer & QA Review	Vasilis Kotsikoris (RAINNO), Eleni Stogiannou (RAINNO), Ourania Ntinou (RAINNO)
V0.9	27/09/2023	90	Implementing review results	Ali Hürriyetoğlu (WFSR) Bas van der Velden (WFSR)
V1.0	29/09/2023	100	Final Deliverable	Ali Hürriyetoğlu (WFSR) Bas van der Velden (WFSR)

EFRA Consortium			
#	Participant Organisation Name	Short name	Country
1	AGROKNOW IKE	AGROKNOW	EL
2	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR	IT
3	STOCKHOLMS UNIVERSITET	SU	SE
4	STICHTING WAGENINGEN RESEARCH	WR	NL
5	MAIZE SRL	MAIZE SRL	IT
6	AGRIVI DOO ZA PROIZVODNJU TRGOVINUI USLUGE	AGRIVI DOO	HR
7	RAINNO IDIOTIKI KEFALAIOUCHIKI ETAIREIA	RAINNO	EL
8	SGS ROMANIA SA	SGS ROMANIA	RO
9	MOY PARK LTD	MOY PARK	UK

Table of Contents

List of Figures

List of Tables

Glossary of terms and abbreviations used

Abbreviation / Term	Description
AI	Artificial Intelligence
EMM	European Media Monitor
FL	Federated Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
ML	Machine Learning

1 Executive Summary

This deliverable describes a roadmap created by the EFRA consortium and the requirements for realizing this roadmap. We first provide details of the requirements of the proposed roadmap in Chapters 3-6. Next, analysis and requirements of the use cases that are part of EFRA project are provided in Chapter 7. Finally, the roadmap of the project is provided in terms of outcome tables in Chapter 8.

The consortium conducted a literature review on signals and relevant data sources for emerging risk prediction for *Salmonella* and pesticide use. Drivers of change and risk factors are identified and documented in detail in Chapter 3. Next, the consortium identified, documented, and analyzed the relevant data mining and processing challenges stemming from the dispersed, heterogenous, multilingual nature of the data sources. The assessment of the data sources analyzed yield comprehensive information about their characteristics and suitability for use in the scope of EFRA project. This information is reported in Chapter 4 and the related appendix. The computational infrastructure of each EFRA partner and its suitability for realizing EFRA use cases is documented in Chapter 5. The identification and assessment of the recent AI approaches used in early warning and emerging risks systems in the food risk domain was carried out with the aim to provide a roadmap for advancements in their data use, limitation, efficiency and explainability. This analysis describes federated learning, which utilizes data at its source, for creating machine learning models and how it will be utilized in the scope of EFRA in Chapter 6. The needs of the end-users, which are use case owners in EFRA consortium, in terms of decision support, performance, privacy, and security for food risk predictions are reported in Chapter 7.

The requirement analyses and the scientific surveys conducted in the scope of EFRA projects lay the foundation of the EFRA outcomes, which are provided in Chapter 8. These outcomes are based on the following conclusions derived from the work described in this deliverable: i) Continuous monitoring and collaborative utilization of the data generated resulting from this monitoring is the first step of food safety; ii) Utilization of data and modelling for early warning is the best option we have for preventing food risk events, iii) Data is available from open sources and from use case partners in EFRA consortium, iv) The computational infrastructure, storage capacity, and modelling paradigm has been determined for ensuring best possible use of state-of-the-art technology for ensuring food safety.

2 Introduction

The goal of this section is to provide a brief outline of the **objectives** of the specific EFRA Deliverable, how those are aligned and relevant with the overall project, and what was the approach followed in order to achieve them.

2.1 Mapping EFRA Outputs

The purpose of this section is to map EFRA Grant Agreement commitments, both within the formal Deliverable and Task description, against the project's respective outputs and work performed.

Table 1: Adherence to EFRA GA Deliverable & Tasks Descriptions

EFRA GA Component Title	EFRA GA Component Outline	Respective Document Chapter(s)	Justification
DELIVERABLE			
D1.1 EFRA Requirements Roadmap This deliverable will incorporate the outputs of T1.1-T1.5 as a guiding set of requirements for the lifetime of the project.			
TASKS			
Task 1.1 Scientific requirements on short- and long-term food risk prediction	The food supply chains are impacted by a web of drivers (economic, socio-economic, environmental, regulatory) that may pose direct or indirect development of food safety risks as short or long term. This task will search the scientific and grey literature for sources of food safety risks incidents (in addition to the European Rapid Alert for Food and Feed (RASFF)) and drivers of change and associated data sources, as such to determine which drivers and signals should be considered for short (i.e. early warning) and long-term (i.e. emerging) food safety risks.	Chapter 3	Drivers of food risk events are identified and documented for the use cases in the scope of EFRA project.
Task 1.2 Heterogeneous data mining	The focus of this task consists of assessment of the data sources identified	Chapter 4	We identified open data sources related to food risk events and

requirements (sources & types)	in Task 1.1 (e.g., EFSA data reports, food safety RSS feeds, scientific publications, European Media Monitor (EMM)) that should be mined to provide valuable information for food risk predictions. This assessment will include the availability, ownership, quality, reliability, and format of each data source. Challenges that the mining technologies need to address will be highlighted, especially due to the extreme variety, heterogeneity, dispersity, and multilinguality of the data records and sources. This task is expected to provide input and recommendations on the data sources that should be harvested, aggregated and enriched in the pipeline developed in WP2.		analyzed their characteristics in line with EFRA project goals.
Task 1.3 Energy-efficient Cloud/Edge HPC architecture & integration requirements	The focus of the task is to study the available (public & private) data, computational resources, and technologies as currently used and further developed by consortium partners, as relevant to the end-goal of AI-enabled food risk prevention through the EFRA Tools. Emphasis will be given to how the integrated solution can be further enhanced in its green & energy-efficiency aspects, both by balancing	Chapter 5	We described computational infrastructures of the use case partners and their utility in line with the requirements of the use cases in the scope of EFRA project.

	the load between cloud and edge-based computations and by delivering advances in green AI training and deployment. The task is expected to provide guidelines to WP2, WP3 and WP4.		
Task 1.4 Public & private data for AI training and data sharing requirements	Intended to lead the design of the specific use-case-driven solutions and the Data Analytics Powerhouse, T1.4 will focus on collecting the requirements to i) deploy the privacy-preserving AI training approach directly over private/sensitive food safety datasets, ii) access, process, and combine public and private data sources and streams, iii) wherever the privacy-preserving AI training approach cannot be directly deployed, identify appropriate schemes that can enhance private FAIR data sharing (e.g., aggregation or anonymization), iv) facilitate FAIR data 35 interoperability through existing/novel data and metadata standards. This task will provide direct guidelines to the data sharing approaches in WP4 and to the use-cases of WP5.	Chapter 6	A comprehensive review on creation and deployment of privacy-preserving machine learning models is conducted. The application of a federated learning system is planned.
Task 1.5 Industrial	This task will identify the information needs of the	Chapter 7	EFRA use cases are described extensively. Problem definition,

decision support requirements for risk prevention	targeted human/expert decision makers in the context of AI-enabled food risk prevention in industrial settings. Decision support use-cases and scenarios will be further specified, around the use of the TRL3 software tools that will be enhanced further within the project. Resulting requirements will influence and inform work in all EFRA WPs, especially the use-cases of WP5.		data and computational infrastructure requirements and success criteria are documented.

2.2 Deliverable Overview and Report Structure

In this section, a description of the Deliverable's Structure is provided, outlining the respective Chapters and their content.

Chapter 3 provides a survey of drivers of food risk events in relation to *Salmonella* and pest occurrences. The focus of these surveys is on early warning system development. This work is the output of the work performed in the scope of Task 1.1.

Chapter 4, which is the work conducted in the scope of T1.2, describes characteristics of open source data sources related to food safety events. A short list of sources is reported on the basis of criteria such as accessibility, utility, and frequency is reported as well.

Chapter 5 is a documentation of the computational resources available for the EFRA use cases by the use case partners. The focus of this overview is mainly about compute power, storage capacity, and accessibility of these resources. The support of big data and machine learning are the key aspects of our analysis. This report is the output of T1.3.

Chapter 6 describes federated learning paradigm, which is a privacy-preserving machine learning paradigm. The results of our scientific survey is concluded with action items for application of this paradigm for the use case we will implement with a use case partner. Task 1.4 foresees search, analysis, and planning of this work. Chapter 7 is the detailed description of the use cases that will be performed with industrial partners of the EFRA consortium.

Chapter 8 suggests a roadmap in terms of outcome tables that specify the problem each subtask as determined by each partner, the relation of an outcome to tasks defined in EFRA project plan, and the end goal in terms of KPI's of the EFRA project.

Chapter 9 provides a brief summary on the utility of the work we conducted in the scope of this deliverable, lessons learned, and the preparations planned for the implementation of the use cases. This is the final chapter of this deliverable.

3 Scientific Requirements on Short- and Long-Term Food Risk Predictions

The intricate and interconnected global food supply chains are profoundly influenced by a multitude of drivers, spanning economic, socio-economic, environmental, and regulatory domains. These factors, in turn, can give rise to direct or indirect food safety risks, some of which emerge rapidly while others slowly build over time, forming an intricate web of complex challenges. This section will focus on two key areas of concern that demand our urgent attention: poultry pathogens and the use of pesticides in primary produce.

These areas have been selected due to their capacity to introduce significant food safety risks, as evidenced through various sources, including scientific and grey literature, as well as the European Rapid Alert for Food and Feed (RASFF). In our exploration of these topics, we will consider both the short-term pressures – those that warrant immediate early warning systems – and the long-term, emerging risks that may potentially be exacerbated by wider scale issues such as climate change.

By delving into the dynamics and drivers behind these risks, we aim to pinpoint the critical indicators that should be monitored for effective risk management. This assessment is pivotal in defining which factors should be prioritized and addressed to ensure food safety and security. In doing so, we can navigate towards solutions that strike a balance between our immediate necessities and the sustainability of our long-term food supply.

3.1 Unraveling the Pathways: A Comprehensive Analysis of *Salmonella spp*. Contamination Sources in the Poultry Supply Chain

Salmonella spp. is a major foodborne pathogen responsible for a substantial number of infections worldwide, with poultry products being a primary vehicle for its transmission. The poultry supply chain is integral to the global food industry, providing a reliable source of poultry products to consumers worldwide. Despite its significance, the intricate and multifaceted nature of this supply chain presents challenges, particularly concerning potential contamination events occurring at various stages, including production, processing, and distribution. Safeguarding the safety and quality of poultry products necessitates a comprehensive understanding of the potential sources of contamination. This report explores the diverse contamination sources in the poultry supply chain and discusses the implementation of artificial intelligence approaches to mitigate risks and uphold food safety standards.

Analysis of Contamination sources in each production step of poultry supply chain

- **Poultry Farming & Hatcheries**: The supply chain begins at poultry farms and hatcheries where chickens or other poultry birds are raised, and eggs are incubated to produce chicks.
- **Poultry Farm**: The poultry farm is responsible for raising the birds until they are ready for processing. This includes providing feed, water, shelter, and appropriate care to ensure the health and well-being of the birds.
- **Slaughterhouse**: Once the birds have reached the appropriate age and weight, they are transported to the slaughterhouse for processing. At the slaughterhouse, the birds are humanely slaughtered and undergo various processing steps to prepare them for distribution.
- **Processing Facilities**: After slaughter, the poultry is taken to a processing plant where it undergoes further cleaning, cutting, and preparation for packaging. Poultry processing facilities are critical points in the supply chain where contamination can occur:

- **Packaging & Labeling**: The processed poultry is then packaged and labeled for distribution. Proper packaging ensures the safety and hygiene of the product during transportation and storage.
- **Transportation and Distribution**: During transportation and distribution, various factors can lead to potential contamination.

Results

Salmonella contamination in poultry is a multifaceted issue that involves multiple contamination sources throughout the supply chain [1–3]. Contamination sources in poultry supply chain include:

1. **Infected Breeder Flocks**: If the parent breeder flocks are carriers of *Salmonella* or other pathogens, the eggs they lay can be contaminated. Consequently, the chicks hatching from these contaminated eggs will already carry the pathogen [4].
2. **Contaminated Eggshells**: The eggshells themselves can become contaminated during laying or collection, allowing pathogens to enter the egg's internal contents. Improper handling and storage of eggs can exacerbate this risk [5].
3. **Environmental Contamination:** The hatchery environment, including incubators, hatching trays, walls and floors, can become contaminated with *Salmonella* through contact with contaminated eggs or infected chicks [6].
4. **Soil and water contamination in farm:** Irrigation water and type of soil amendment can be risk factors for *Salmonella* contamination.
5. **Inadequate Cleaning and Sanitization**: Improper cleaning and disinfection of hatchery equipment can lead to the persistence of pathogens on surfaces and the potential for cross-contamination.
6. **Personnel and Equipment**: Human handlers and equipment that come into contact with the eggs and chicks can introduce pathogens into the hatchery environment.
7. **Airborne Contamination**: Airborne particles carrying pathogens can settle on eggs and surfaces, leading to contamination.
8. **Poor Biosecurity Practices**: Insufficient biosecurity measures, such as limited access control and inadequate hygiene protocols, can facilitate the entry and spread of pathogens into the hatchery.
9. **Water Contamination**: Water used for egg washing, cleaning, or misting in the hatchery can be a source of contamination.
10. **Inadequate Refrigeration or Storage:** Improper storage conditions can allow the growth and survival of *Salmonella* in poultry products.
11. **Hygiene and Sanitation**: Poor sanitation practices in processing plants can result in the proliferation of harmful bacteria, viruses, and parasites [7].
12. **Employee Hygiene**: Improper handwashing and hygiene practices among processing plant workers can contribute to the spread of pathogens.
13. **Contaminated Ingredients**: If poultry products are further processed and mixed with other ingredients, any contaminated ingredients can introduce *Salmonella*.
14. **Cross-Contamination**: Inadequate separation between raw and processed poultry can lead to cross-contamination of pathogens, such as *Salmonella* and *E. coli*.[8]
15. **Inadequate Pest Control**: Pests, such as rodents and insects, can carry *Salmonella* and spread it within the facility [9].
16. **Temperature Control**: Improper temperature control during transportation can promote the growth of bacteria and compromise the quality and safety of poultry products.

17. **Packaging**: Damaged or contaminated packaging can introduce pathogens or harmful substances to poultry products during transit.

Biofilms are indeed an important contamination source for *Salmonella* in poultry production and processing facilities. Biofilms are complex communities of microorganisms that adhere to surfaces and form protective structures, making them resistant to cleaning and disinfection efforts. In the context of poultry production, biofilms can develop on various surfaces, **including equipment, processing machinery, floors, walls, and even poultry carcasses**. Biofilms can harbor and protect *Salmonella* bacteria, providing them with a survival advantage in harsh environmental conditions. Once established, biofilms can serve as continuous sources of contamination, leading to persistent and recurrent *Salmonella* outbreaks in poultry facilities [10].

Table 2: Investigation Analysis of Salmonella Contamination Incidents in the Poultry Supply Chain

Sample ID	Strain	Source	Origin	Poultry operation	Year
1	Newport	irrigation water	south-east USA	primary agricultural practices	2017
2	Enteritis	irrigation water	south-east USA	primary agricultural practices	2017
3	Livingstone	egg transfer area	USA -Maryland	primary breeder farms	2017
4	Thomasville	hatchery rooms	USA-Alabama	primary breeder farms	2017
5	Enteritis	chick sorting area	USA-Mississippi	primary breeder farms	2017
6	Mbandaka	macerator room	USA-Texas	primary breeder farms	2017
7	Typhimurium	ventilation ducts	USA-Wyoming	primary breeder farms	2017
8	Infantis	waste area outside	USA	primary breeder farms	2017
9	Agona	airborne dust	Japan	primary breeder farms	2017
10	Heidelberg	fluff&feces in transport trail liners	Japan	primary breeder farms	2017
11	Kentucky	fluff&feces in transport trail liners	Japan	primary breeder farms	2017
12	Montevideo	incubator temperature	Japan	primary breeder farms	2017
13	Hadar	incubator temperature	Japan	primary breeder farms	2017
14	Kentucky	litter	Bristol-UK	broiler farms	2017
15	Heidelberg	feces	Japan	broiler farms	2017

16	Mbandaka	bedding	USA-North Carolina	broiler farms	2017
17	Hadar	flies	USA-Georgia	broiler farms	2017
18	Enteritis	flooring after poor dectontamination	USA	broiler farms	2017
19	Infantis	flooring after poor dectontamination	USA	broiler farms	2017
20	Anatum	compost	USA	broiler farms	2017
21	Anatum	wastewater	USA	broiler farms	2017
22	Anatum	pest	USA	broiler farms	2017
23	Anatum	parent flocks	USA	broiler farms	2017
24	Anatum	Fresh feed	USA	broiler farms	2017
25	Anatum	Topsoil	USA	broiler farms	2017
26	Typhimurium	Grain drying area	Bristol, UK	feed production	2017
27	Enteritidis	Intake pits	Norway	feed production	2017
28	Newport	Grinder spills	Norway	feed production	2017
29	Ohio	Cooler interior & spillage	Norway	feed production	2017
30	Ohio	Pelleter area	Norway	feed production	2017
31	Ohio	Wild bird droppings	Norway	feed production	2017
32	Ohio	Raw soybean	Norway	feed production	2017
33	Ohio	Ship interior	Norway	feed production	2017
34	Hadar	Duration	Alberta, Qubec, Canada.	transportation	2017
35	Infantis	Flock size	Netherlands	transportation	2017
36	4,12:d	Temperature fluctuations	Hungary	transportation	2017
37	Ohio	Wait time in crates	Iowa, USA	transportation	2017
38	Mbandaka	Fecal shedding	Iowa, USA	transportation	2017
39	Senftenberg	Cross contamination	Iowa, USA	transportation	2017
40	Derby	Feather debris	France	slaugher house operations	2017
41	Derby	Area outside the plant	France	slaugher house operations	2017
42	Derby	No. of workers handling	France	slaugher house operations	2017
43	Derby	evisceration	France	slaugher house operations	2017
44	Derby	Picker fingers	France	slaugher house operations	2017
45	Liverpool	Bone marrow	Georgia, USA	Further processing	2017
46	Kentucky	Neck skin	Poland	Further processing	2017
47	Typhimurium	Mechanically separated meat	Thailand	Further processing	2017
48	Give	Mechanically separated meat	Australia	Further processing	2017

49	Montevideo	Packaging	Australia	Further processing	2017
50	Senftenberg	Wheat flour	Australia	Further processing	2017
51	Agona	Peppers (added after final pathogen reduction)	Australia	Further processing	2017
52	Thompson	Peppers (added after final pathogen reduction)	Australia	Further processing	2017
53	Thompson	Herbs	Australia	Further processing	2017
54	Heidelberg	Retail ground chicken	Washington-USA	Distribution channels	2017
55	Enteritidis	Retail ground turkey	USA	Distribution channels	2017
56	Kentucky	Food handling (raw meat cross contamination)	Ontario, Canada	Distribution channels	2017
57	Infantis	Skin after chilling	Ecuador-USA	Slaughterhouse	2020
58	Infantis	Skin after final washing	Ecuador-USA	Slaughterhouse	2020
59	Amsterdam	Raw feed materials	Ecuador-USA	Feed mill plant	2020
60	Liverpool	Raw feed materials	Ecuador-USA	Feed mill plant	2020
61	Infantis	Overshoes	Ecuador-USA	Broiler farms	2020
62	Uganda	Overshoes	Ecuador-USA	Broiler farms	2020
63	Infantis	Skin after final washing	Ecuador-USA	Slaughterhouse	2020
64	Bargny	knives swab	Egypt	Slaughterhouse	2017
65	Enteritidis	table	Egypt	Slaughterhouse	2017
66	Kentucky	abbatoir wall	Egypt	Slaughterhouse	2017
67	Typhimurium	carcass	Egypt	Slaughterhouse	2017
68	Enteritidis	farm walls	Egypt	Broiler farms	2017
69	Typhimurium	farm walls	Egypt	Broiler farms	2017
70	Kentucky	carcass	Egypt	processed broiler	2017

Conclusion

Addressing potential contamination sources in the poultry supply chain requires a holistic approach involving stakeholders at every stage. Implementing and enforcing strict biosecurity measures at farms, ensuring proper hygiene and sanitation practices in processing plants, maintaining adequate temperature control during transportation, and promoting safe handling practices at retail and consumer levels are vital steps to mitigate the risk of contamination. Regular monitoring, testing, and traceability measures are also essential to identify and address contamination issues promptly, thereby ensuring the safety and quality of poultry products reaching consumers' plates.

References

[1] M. Elsayed, F. El-Gohary, A. Zakaria, & M. Gwida, “Tracing of salmonella contaminations throughout an integrated broiler production chain in Dakahlia Governorate Egypt,” *Pakistan Veterinary Journal*, **39** (2020) 558–562. https://doi.org/10.29261/pakvetj/2019.038.

[2] K. Rajan, Z. Shi, & S. C. Ricke, “Current aspects of Salmonella contamination in the US poultry production chain and the potential application of risk strategies in understanding emerging hazards,” *Critical Reviews in Microbiology*, **43** (2017) 370–392. https://doi.org/10.1080/1040841X.2016.1223600.

[3] A. Wales & R. Davies, "Review of hatchery transmission of bacteria with focus on Salmonella , chick pathogens and antimicrobial resistance," *World's Poultry Science Journal*, **76** (2020) 517–536. https://doi.org/10.1080/00439339.2020.1789533.

[4] R. Rodríguez-Hernández, J. F. Bernal, J. F. Cifuentes, L. C. Fandiño, M. P. Herrera-Sánchez, I. Rondón-Barragán, & N. V. Garcia, "Prevalence and molecular characterization of salmonella isolated from broiler farms at the Tolima region—Colombia," *Animals*, **11** (2021) 1–11. https://doi.org/10.3390/ani11040970.

[5] M. R. Karim, G. Md, M. Samad, M. R. Karim, M. Giasuddin, M. A. Samad, M. S. Mahmud, M. R. Islam, M. H. Rahman, & M. A. Yousuf, "Prevalence of Salmonella spp in Poultry and Poultry Products in Dhaka, Bangladesh Prevalence of Salmonella spp. in Poultry and Poultry Products in Dhaka, Bangladesh," *International Journal of Animal Biology*, **3** (2017) 18–22.

[6] L. E. Hubbard, C. E. Givens, D. W. Griffin, L. R. Iwanowicz, M. T. Meyer, & D. W. Kolpin, "Poultry litter as potential source of pathogens and other contaminants in groundwater and surface water proximal to large-scale confined poultry feeding operations," *Science of the Total Environment*, **735** (2020). https://doi.org/10.1016/j.scitotenv.2020.139459.

[7] T. Obe, R. Nannapaneni, W. Schilling, L. Zhang, C. McDaniel, & A. Kiess, "Prevalence of Salmonella enterica on poultry processing equipment after completion of sanitization procedures," *Poultry Science*, **99** (2020) 4539–4548. https://doi.org/10.1016/j.psj.2020.05.043.

[8] O. A. Adeboye, M. K. Kwofie, & N. Bukari, "Campylobacter, Salmonella; and & Escherichia coli Food Contamination Risk in Free-Range Poultry Production System," *Advances in Microbiology*, **10** (2020) 525–542. https://doi.org/10.4236/aim.2020.1010039.

[9] T. K. Nguyen, L. T. Nguyen, T. T. H. Chau, T. T. Nguyen, B. N. Tran, T. Taniguchi, H. Hayashidani, & K. T. L. Ly, "Prevalence and antibiotic resistance of Salmonella isolated from poultry and its environment in the Mekong Delta, Vietnam," *Veterinary World*, **14** (2021) 3216–3223. https://doi.org/10.14202/vetworld.2021.3216-3223.

[10] L. Merino, F. Procura, F. M. Trejo, D. J. Bueno, & M. A. Golowczyc, "Biofilm formation by Salmonella sp. in the poultry industry: Detection, control and eradication strategies," *Food Research International*, **119** (2019) 530–540. https://doi.org/10.1016/j.foodres.2017.11.024.

3.2 Decision support for optimized, food safety-conscious regional pesticide application

Apples are considered the second most important fruit crop worldwide [1]. In traditional apple orchard management systems, yield loss due to pest damage can be as high as 50% [2]. But pesticide use is associated with sustainability and environmental concerns, driving a global push towards optimization of pesticide use, highlighted by the recent June 2022 adoption of the European Council's Directive on the Sustainable Use of Plant Protection Products [3]. One target of this directive is EU-wide reduction of chemical pesticides by the year 2030, recommending expanded technological monitoring solutions to increase integrated pest management (IPM), a multipronged, ecosystem-aware approach to pest control.

This report intends to provide an overview of planned work for development of a novel pest forecasting model for apple orchard IPM, expanding a decision-support framework developed by AGRIVI [4]. Many ecological, topological, meteorological, and human factors are at play in the transmission and colonization of plant pests, which can be combined with ground-level high-resolution data from smart agriculture sensors like those supplied by AGRIVI. Large-

scale additional datasets – including publicly accessible remote sensing data and meteorological data - will be used to train a final, expanded model for pest-specific activity prediction, which will then be integrated with the AGRIVI ruleset to produce updated region- and phenology-informed pesticide recommendations.

Proposed Methodology

Recent years have seen the rise of "smart" or precision agriculture, with an array of big data, artificial intelligence (AI), and machine learning (ML) methods employed for proactive decision-making. Remote sensing data - i.e., satellite imagery, multi- or hyper-spectral sensing, thermal and IR sensing, and radar detection - can be mined for information on agricultural land morphology and crop phenology, as well as used to monitor disease, damage, and pest detection. On the ground, Internet of Things (IoT) sensor networks can be employed to integrate data from monitoring of factors such as soil moisture, air conditions, and pest activity near sticky or pheromone traps [5]. With these data sources in mind, ML and AI have found widespread application as tools to enhance agribusiness decision support [6]. ML classifiers incorporating various IoT data streams and remote sensing imagery have been previously employed for downstream risk prediction in agriculture [7], [8].

A pest risk prediction model must consider the many biotic and abiotic factors that drive pest spread, transfer, and development; such drivers are specific to each pest and class thereof, as well as the region. Some drivers of pest development are included in Table 3.

Table 3: Drivers of pest development

Category	Examples of specific drivers	Rationale	Reference
Geographical info	Latitude Spatial GIS data	Latitudinal trends have been demonstrated in pest developmental stages – for example, latitude may establish lower/upper bounds for overwintering capability.	Latitude for moths [9] Latitude gradient: bollworm/cotton earworm [10]
Satellite data (remote sensing)	Vegetation indices (ex. NDVI, GI, GCVI)	Monitoring water indices serves as early warning system. But water indices have also been used in forecasting models for spatial predictions of insect and crop phenology. Phenological milestones at both regional and global scales can be established, such as snow-melt or first greening.	Vegetation indices [11] for bollworm
	Water indices (ex. NDWI)		NDWI for modeling insect (moth) phenology [9]
	Soil indices (ex. Soil water index, surface soil moisture)		
Weather (meteorological) station (either ground-based or otherwise)	**Precipitation:** accumulation, intensity, and frequency, relative humidity, dew point	Moisture is a major driver of pest lifecycle development. Moisture may also control pest predators or parasites (for example, parasitic fungi that feed on aphids). Similarly, temperature (and thermal accumulation, expressed commonly as	Rainfall, humidity, temperature, wind speed apple scab [14] RH and precipitation [15] Accumulation of effective temperatures (Codling, Metos by Pessl) [16]

	Temperature: air temperature, land surface temperature	growing degree-days or GDD) drives both pest and crop lifecycle development. This is pest-dependent; for example, temperature and solar radiation are primary drivers in the development of codling moth, whereas precipitation and relative humidity are the major drivers for the development of pests like fire blight [12]. Atmospheric CO2 can alter overwintering habitat decay, increase leaf size/density, raise humidity in foliage and exacerbate pest presence in foliage [13]	Temperature-sum thresholds relating to crop phenological events can be employed rather than full phenological models [12] Precipitation intensity and frequency [12] Atmospheric CO2 [13]
	Atmospheric: wind speed and wind direction, radiation, sunshine hours, atmospheric CO2		
Soil properties	Local soil temperature, conductivity, relative humidity, composition (ex. total organic content and clay percentage), microbiome, moisture/drainage indices, pH, soil horizons, color, texture, classification, erosion, and drainage	Soilborne pathogens and insects find preferable habitat in certain types of soil. Physicochemical soil characteristics are additionally associated with water and pesticide retention. Soil pH and conductivity affect microbial soil communities under the surface, and affect nutrient availability and soil fertility, as well as affecting a plant's innate tolerance of or resistance to pests and pesticides.	Averaged winter soil temperature predictive for overwintering success for corn earworm [10] Soil-pest relationships [17]
Pest-specific	Pest trap activity (measured with computer vision-based monitored smart traps)	Pest trap activity serves as not only a warning system, but also important frequency-based data that can be used to train a forecasting model. Similarly, historical pest presence is useful in assessing current pest invasion risk. Each pest has its own stress thresholds in pest-specific models, reflecting adverse seasonal conditions. As an example of a general model, CLIMEX	Pest trap activity [18] [19] Pest-specific developmental phases (ex larvae stages for codling moth, Metos by Pessl) [17]
	Historical pest presence		Historical presence in orchard or in local area (Metos by Pessl) Maturation process model specific to ex scab (RIMpro)

	Generic or pest-specific lifecycle or distribution simulators (ex CLIMEX [20])	uses soil moisture, air temperature, daylight/length, cold stress, heat stress, dry stress, wet stress, diapause day length and temperature, and hot wet stress, and has been expanded for specific pest application.	
Agronomic inputs	Irrigation	In addition to precipitation, irrigation is another source of field water and influences humidity and atmospheric conditions. Duration of leaf wetness (in combination with temperature) is important in the development of pests like apple scab or fire blight. Foliage density is useful not just as a metric for crop phenology, but also relates to infection risk by providing habitat suitable for invasion or development. Pollinator visit frequency affects crop yield, and interacts with pest predation and damage effects.	Irrigation [15] Leaf wetness for apple scab [21] Duration of leaf wetness [22] Duration of leaf wetness (and humidity) are biggest factors in development of fire blight disease [12] Minimum wetting period (ex for apple scab) [14] Pollinators for crop yield [29], satellite-derived crop phenology [30] Crop phenology (apple leaf spreading to fruit expanding period) considered [23]
	Leaf wetness duration		
	Foliage density-based metrics: leaf area index, canopy height, width between canopy rows		
	Pollinator visit frequency		
	Generic or crop-specific phenological markers or developmental simulators		

Conclusion

Within the last five years, increased global focus has been brought to bear on pesticide use from a sustainability perspective. Development of pest forecasting models has conventionally been motivated by prevention of yield loss, reduction of operating costs, or improvement of operating efficiency - i.e., an economic rationale for the farmer via lowered treatment costs. In addition, EFRA is motivated by a focus on food safety, and the downstream effects of decreasing and optimizing chemical pesticide usage. In recent years there has been increased awareness of pesticide residue risks on primary produce, a hazard not just for consumers but also for workers along the supply chain [31].

References

[1] "FAOSTAT." https://www.fao.org/faostat/en/#home (accessed Sep. 05, 2023).

[2] J. Cross, M. Fountain, V. Markó, and C. Nagy, "Arthropod ecosystem services in apple orchards and their economic benefits," *Ecol. Entomol.*, vol. 40, no. S1, pp. 82–96, 2015, doi: 10.1111/een.12234.

[3] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the sustainable use of plant protection products and amending Regulation (EU) 2021/2115. 2022. Accessed: Aug. 23, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2022:305:FIN

[4] "AGRIVI: Farm Management Software for Digital Agriculture," *AGRIVI*. https://www.agrivi.com/ (accessed Aug. 23, 2023).
[5] "Work ongoing in DEMETER Orchard pilot - Demeter," Aug. 08, 2022. https://h2020-demeter.eu/work-ongoing-in-demeter-orchard-pilot/ (accessed Aug. 23, 2023).
[6] A. Yousaf, V. Kayvanfar, A. Mazzoni, and A. Elomri, "Artificial intelligence-based decision support systems in smart agriculture: Bibliometric analysis for operational insights and future directions," *Front. Sustain. Food Syst.*, vol. 6, 2023, Accessed: Aug. 23, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsufs.2022.1053921.
[7] J. Wang and H. Yue, "Food safety pre-warning system based on data mining for a sustainable food supply chain," *Food Control*, vol. 73, pp. 223–229, Mar. 2017, doi: 10.1016/j.foodcont.2016.09.048.
[8] Y. H. Kim, S. J. Yoo, Y. H. Gu, J. H. Lim, D. Han, and S. W. Baik, "Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey," *IERI Procedia*, vol. 6, pp. 52–56, Jan. 2014, doi: 10.1016/j.ieri.2014.03.009.
[9] J. Pöyry *et al.*, "Predictive power of remote sensing versus temperature-derived variables in modelling phenology of herbivorous insects," *Remote Sens. Ecol. Conserv.*, vol. 4, no. 2, pp. 113–126, 2018, doi: 10.1002/rse2.56.
[10] D. Lawton *et al.*, "Pest population dynamics are related to a continental overwintering gradient," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 119, no. 37, p. e2203230119, Sep. 2022, doi: 10.1073/pnas.2203230119.
[11] O. Nanushi, V. Sitokonstantinou, I. Tsoumas, and C. Kontoes, "Pest presence prediction using interpretable machine learning." arXiv, May 16, 2022. Accessed: Sep. 05, 2023. [Online]. Available: http://arxiv.org/abs/2205.07723.
[12] M. Hirschi *et al.*, "Downscaling climate change scenarios for apple pest and disease modeling in Switzerland," *Earth Syst. Dyn.*, vol. 3, no. 1, pp. 33–47, Feb. 2012, doi: 10.5194/esd-3-33-2012.
[13] W. J. Manning and A. V Tiedemann, "Climate change: potential effects of increased atmospheric carbon dioxide (CO2), ozone (O3), and ultraviolet-B (UV-B) radiation on plant diseases," *Environ. Pollut. Barking Essex 1987*, vol. 88, no. 2, pp. 219–245, 1995, doi: 10.1016/0269-7491(95)91446-r.
[14] M. Wrzesień, W. Treder, K. Klamkowski, and W. R. Rudnicki, "Prediction of the apple scab using machine learning and simple weather stations," *Comput. Electron. Agric.*, vol. 161, pp. 252–259, Jun. 2019, doi: 10.1016/j.compag.2018.09.026.
[15] S. Guo, X. Ge, Y. Zou, Y. Zhou, T. Wang, and S. Zong, "Projecting the Global Potential Distribution of Cydia pomonella (Lepidoptera: Tortricidae) Under Historical and RCP4.5 Climate Scenarios," *J. Insect Sci.*, vol. 21, no. 2, p. 15, Apr. 2021, doi: 10.1093/jisesa/ieab024.
[16] M. S. U. Extension, *Fruit Crop Ecology and Management*. Michigan State University Extension, 2002.
[17] M. A. Altieri, L. Ponti, and C. I. Nicholls, "Soil Fertility, Biodiversity and Pest Management," in *Biodiversity and Insect Pests*, G. M. Gurr, S. D. Wratten, W. E. Snyder, and D. M. Y. Read, Eds., 1st ed.Wiley, 2012, pp. 72–84. doi: 10.1002/9781118231838.ch5.
[18] D. Marković, D. Vujičić, S. Tanasković, B. Đorđević, S. Ranđić, and Z. Stamenković, "Prediction of Pest Insect Appearance Using Sensors and Machine Learning," *Sensors*, vol. 21, no. 14, p. 4846, Jul. 2021, doi: 10.3390/s21144846.
[19] J. Suto, "Codling Moth Monitoring with Camera-Equipped Automated Traps: A Review," *Agriculture*, vol. 12, no. 10, Art. no. 10, Oct. 2022, doi: 10.3390/agriculture12101721.
[20] J.-M. Jung, W.-H. Lee, and S. Jung, "Insect distribution in response to climate change based on a model: Review of function and use of CLIMEX," *Entomol. Res.*, vol. 46, no. 4, pp. 223–235, 2016, doi: 10.1111/1748-5967.12171.
[21] A. Stella, G. Caliendo, F. Melgani, R. Goller, M. Barazzuol, and N. La Porta, "Leaf Wetness Evaluation Using Artificial Neural Network for Improving Apple Scab Fight," *Environments*, vol. 4, no. 2, Art. no. 2, Jun. 2017, doi: 10.3390/environments4020042.
[22] A. Atlamaz, C. Zeki, and A. Uludag, "The importance of forecasting and warning systems in implementation of integrated pest management in apple orchards in Turkey*," *EPPO Bull.*, vol. 37, pp. 295–299, Sep. 2007, doi: 10.1111/j.1365-2338.2007.01129.x.

[23] Y. Huang *et al.*, "Forecasting Alternaria Leaf Spot in Apple with Spatial-Temporal Meteorological and Mobile Internet-Based Disease Survey Data," *AGRONOMY-BASEL*, vol. 12, no. 3. MDPI, ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND, Mar. 2022. doi: 10.3390/agronomy12030679.

[24] C. D. S. Ecosystem, "Copernicus Data Space Ecosystem | Europe's eyes on Earth," Jul. 20, 2023. https://dataspace.copernicus.eu/ (accessed Aug. 23, 2023).

[25] "LSA SAF." https://landsaf.ipma.pt/en/ (accessed Aug. 23, 2023).

[26] R. Felber, S. Stoeckli, and P. Calanca, "Generic calibration of a simple model of diurnal temperature variations for spatial analysis of accumulated degree-days," *Int. J. Biometeorol.*, vol. 62, no. 4, pp. 621–630, Apr. 2018, doi: 10.1007/s00484-017-1471-5.

[27] J. R. Marques da Silva, C. V. Damásio, A. M. O. Sousa, L. Bugalho, L. Pessanha, and P. Quaresma, "Agriculture pest and disease risk maps considering MSG satellite data and land surface temperature," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 38, pp. 40–50, Jun. 2015, doi: 10.1016/j.jag.2014.12.016.

[28] Y. Zhu *et al.*, "Identification of Apple Orchard Planting Year Based on Spatiotemporally Fused Satellite Images and Clustering Analysis of Foliage Phenophase," *Remote Sens.*, vol. 12, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/rs12071199.

[29] I. Bartomeus, V. Gagic, and R. Bommarco, Pollinators, pests and soil properties interactively shape oilseed rape yield. Basic and Applied Ecology, 16(8), pp. 737-745, 2015.

[30] S. Skawsang, M. Nagai, N. K. Tripathi, and P. Soni, Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: a case study for the Central Plain of Thailand. Applied Sciences, 9(22), 4846, 2019.

[31] W. Boedeker, M. Watts, P. Clausing, and E. Marquez, "The global distribution of acute unintentional pesticide poisoning: estimations based on a systematic review," *BMC Public Health*, vol. 20, no. 1, p. 1875, Dec. 2020, doi: 10.1186/s12889-020-09939-0.

4 Heterogenous Data Mining Requirements

The focus of this task consists of assessment of the data sources identified in Task 1.1 (e.g., EFSA data reports, food safety RSS feeds, scientific publications, European Media Monitor (EMM)) that should be mined to provide valuable information for food risk predictions. This assessment includes the availability, ownership, quality, reliability, and format of each data source. Challenges that the mining technologies need to address will be highlighted, especially due to the extreme variety, heterogeneity, dispersity, and multilinguality of the data records and sources. This task is expected to provide input and recommendations on the data sources that
should be harvested, aggregated and enriched in the pipeline developed in WP2.

In the task at hand, we delve into the assessment of various food safety data sources. These include EFSA data reports, public food safety authority websites, food safety RSS feeds, scientific publications, and the European Media Monitor (EMM), among others. Our aim is to harness these rich information reservoirs to yield valuable insights for predicting food-related risks.

In our evaluation process, we scrutinize several vital parameters for each data source, including its availability, ownership, quality, reliability, and data format. Our objective is not only to identify the most beneficial and reliable sources but also to understand the potential challenges inherent in extracting and integrating data from these diverse resources.

We acknowledge the complexity presented by the extreme variety, heterogeneity, dispersity, and multilinguality of the data records and sources. Identifying and addressing these challenges, especially in relation to mining technologies, forms a crucial part of our undertaking. The insights and learnings from this task will be instrumental in providing input and recommendations on the data sources that should be harvested, aggregated, and enriched in the pipeline developed in WP2. We aspire to utilize these rich data sources effectively, paving the way for robust and accurate food risk predictions.

4.1 Data Sources Assessment

Methodology

We have examined a number (33) of information resources, ranging from governmental agencies, educational institutions, and social media platforms, primarily from regulatory authorities and organizations related to food safety from around the world.

We carried out our assessment by inspecting publicly available data source documentation, and samples of data whenever possible. For each source we tried to answer questions such as:

- What format is the data? How challenging is it to harvest it?
- Is data subject to copyright and/or database rights? Whose?
- Are there specific terms and conditions governing the use of the data?
- Is it possible to purchase a permit to use the data? How much would it cost?
- How much historical data is available?
- How much future data is expected?
- When was the data last published?

- How rich is the data? Does it provide many details?
- How precise is the data? What is the level of detail?
- If data is annotated, what is the annotation schema?
- How trusted is the data?
- How timely is the data source? Does the data track closely recent events? What is the delay between an event and the publication of data about it?

In general, it was not always possible to give a conclusive answer to all these questions, but our partial answers to these questions and our reasonable judgement form the basis of our assessment.

The result of this work consists in a synopsis, a recommendation, a comparison table, and an Annex with details about specific data sources, such as their URLs and their terms and conditions.

Synopsis

In our quest for data and resources, we have looked into different platforms to compile a comprehensive list of the information available worldwide. The potential data sources encompass an extensive array of domestic and international regulatory entities, knowledge sharing platforms, scientific databases, and social media. The data is predominantly text-based, with a limited number of video sources. The evaluation of each potential data source is based on several crucial parameters, including the size of the document in terms of numbers and type, update frequency, cost, margins of compliance, reliability, and quality of data.

On a national level, key sources include the Canadian Food Inspection Agency, the U.S Food and Drug Administration (FDA), the Food Safety Authority of Ireland and the Abu Dhabi Food Control Authority. The FDA also provides the largest individual sources in terms of document quantity with its Import Refusals and Inspections Citations. In terms of global outreach, we have examined data from the Rapid Alert System for Food and Feed (RASFF), the European Food Safety Authority (EFSA). There are also several databases of scientific literature available as data sources, such as Scopus and PubMed.

Furthermore, aside from authoritative bodies, the research also integrates data from food safety dedicated websites like BarfBlog, Food Safety Dot Com, Food Safety News and Food Safety Tech. There is also active inclusion of information from social media platforms such as X (Twitter) and YouTube, although the exact size of these sources is not specified as they are continuously updated.[3]

The abundance of information found manifests predominantly in text format, making up a staggering 90% of all food safety-related data. However, this is not to undermine the influential role that multimedia plays. YouTube videos, although scarce, comprise roughly 10% of the collection. These platforms allow for a plethora of perspectives and give a voice to different entities such as scientists, health influencers, and consumers.

[3] We have selected the sources that are the most relevant and usable ones for the scope of EFRA project. However, this methodology may have caused a bias in geographical origin of the sources. We argue that adding more sources from more countries may not resolve this issue. Having noted that, EFRA consortium will take this potential imbalance into account in usage of and inference from the data.

Out of all the resources sourced, FDA Import Refusals emerges as the largest provider with an impressive 440 thousand records. On the other end of the spectrum, smaller authorities such as the Abu Dhabi Agriculture and Food Safety Authority contribute significantly to the local scope of food safety with a collection of 20 documents.
We noticed that some data are updated almost daily (like the Brazilian Health Regulatory Agency, EFSA, and others), while others have less frequent updates. This might affect the timeliness and relevance of the information.

Interestingly, different agencies perceive the urgency of updating their database distinctively. Some regulatory bodies, such as the Brazilian Health Regulatory Agency and EFSA, are extensively diligent, providing daily updates to their database. The frequency of these updates ensures the latest developments are captured, thus guaranteeing current and pertinent information. Some other sources, however, are less consistent in their updates.

In terms of data freshness, most sources have data that is days or weeks old compared to the actual event date. Very few sources are truly NRT. Government agencies like the FDA, Canadian Food Inspection Agency (CFIA), and UK Food Standards Agency (UK FSA) tended to maintain the freshest of data available to the public, often publishing recalls and alerts within days. Industry news sites also provided reasonably fresh information, with details on recalls and outbreaks emerging within days of the incidents. Scientific literature sources such as PubMed and Scopus showed more latency, with data availability months after events due to the publication process timeline. Social media like X (Twitter) offered more real-time signals but lacked context around the posts. Some sources were found to be inactive/static websites or contained a mix of content with variable freshness. Based on these findings, it is recommended that timely monitoring of food safety events focus on key government agency sources and industry news sites. Social media mining can provide emerging signals in near real-time to complement this tracking. Expectations should be adjusted around the scientific literature, which will remain a source of retrospective analysis. Inactive or static sources adding little value should be dropped from monitoring. For mixed sources, monitoring should focus on the sections or content types where timeliness is greatest.

In terms of trustworthiness, the Food and Drug Administration (FDA) and European Food Safety Authority (EFSA) are unrivalled, regarded as the most reliable and trusted sources for food safety. Yet, even platforms like X (Twitter) hold their relevance, albeit seen as a lower trust source. In this modern information age, these platforms have shown their propensity to break news faster than traditional media and are often tapped upon for real-time insights.

Looking at some of the resources in more details we can notice that:

- At the core of the Brazilian Health Regulatory Agency's (Anvisa) database are approximately 10,000 documents which are updated nearly daily, making it a consistently relevant resource. However, there is a significant delay in the publication of alerts, which may occur two months post-identification of the problem by the company.
- The European Food Safety Authority (EFSA) hosts around 10,000 documents updated almost daily, providing a rich and current knowledge base.
- The Abu Dhabi Agriculture and Food Safety Authority, despite containing only about 20 documents, still carries high reliability. It updates its materials annually at best, which hints at slower response times.
- The ANSES (French National Agency for Food Safety, Environment, and Labor) provides weekly updated content spanning about 1,500 documents. A pertinent observation is it lacks a specific section dedicated to warnings or recalls, somewhat hampering its utility for immediate food safety issues.

- The Australian Department of Agriculture Imported Food Reports issues food failing reports bi-monthly with a two months' lag and contains 77 documents.
- The Austrian Food Safety Authority maintains an agile system with weekly updates of its 300 documents database. Its recall resources are well updated, enhancing its credibility.
- BarfBlog, with 12,500 documents updated nearly every day, stands as an average trust platform. Meanwhile, the Youtube channel of the Centers for Disease Control and Prevention (CDC) offers a multimedia approach to food safety education.
- The EU Knowledge Centre for Food Fraud and Quality, with 240 documents, updates monthly. Noticeably, the freshness of its news varies greatly.
- The FDA's various sectors–like its Enforcement Reports, Import Alerts, Import Refusals, and Inspection Citations–are extremely well-maintained, with high reliability and freely available information. Its Recall section shines for promptly publishing announcements from companies.
- Websites like Food-Safety.com, Food Safety News, and Food Safety Tech, offering weekly or daily updates, are characterized by their consistent traffic but require additional investigation on their relevance to EFRA events.
- The German Federal Office of Consumer Protection and Food Safety and the New Food Magazine consistently maintain all their resources updated to reflect the most recent events, with a focus on recalls and weekly roundups.
- PubMed and Scopus, including scholarly publications, offer very large databases but bear in mind these materials usually start months after an event to allow for comprehensive investigation and peer-review.
- The Rapid Alert System for Food and Feed (RASFF) stands out for their near real-time update of notifications; yet it should be noted that hazards usually occur weeks before the notification.
- X (Twitter) as a resource has low trust due to the user-generated nature of the content coupled with poor verification methods. Still, its near real-time update might provide valuable insights on emergent situations.
- The UK Food Standards Agency maintains high trust with the earliest recalls starting days prior, making it a reliable resource.
- Finally, the Canadian Food Inspection Agency and the Ministry of Agriculture, Nature, and Food Quality (LNV) in the Netherlands, although not reporting the size of their databases, were highlighted for the freshness and reliability of their alerting systems.

Recommendation

At the present time, our assessment indicates that the following data sources are the best candidates for mining data that might be valuable for the EFRA project. Again, details about these sources are reported in the annex.

- CDC YouTube channel
- EFSA (European Food Safety Authority)
- EFSA (European Food Safety Authority) YouTube channel
- FDA Enforcement Reports
- FDA Import Refusals
- FDA Inspections Citations

- FDA Recalls
- FSIS USDA
- PubMed
- RASFF (Rapid Alert System for Food and Feed)
- Scopus

The basis for this recommendation is the data sources' generally permissive terms of use, data quality and freshness, and ease of extraction. Among these it is worth pointing out that Scopus allows non-commercial use of data for free but commercial use requires payment of a fee, yet to be determined.

Comparison Table

In the following table, we present a high-level view of each data source, with a short description for each the following dimensions:

1. Size: provides the quantity of documents or videos available from each source
2. Cost: indicates whether access to the information from each source is free or paid
3. Compliance burden: indicates how easy or how hard it is to use the data and comply with terms and conditions of each source
    - High: Highly Restricted, Personal Use Only
    - Medium: Default copyright, Non-permissive terms of use,
    - Low: No restrictions, Attribution required, no derivative works allowed
4. Quality: indicates the potential usefulness of the data source for food safety analysis / prediction (e.g., reporting or not food safety incidents) and its level of details (i.e., containing explicit references to products, hazards, locations, companies, etc)
5. Reliability: indicates how reliable the source is and how trustworthy the information it provides is. We marked high reliability for institutional data, medium reliability for news and magazines, low reliability for user-generated reviews.
6. Format: describes the data format, such as HTML, PDF, XLSX, video
7. Ease of extracting information: we classify as easy all the sources that have an RSS feed or a clean and straightforward structure, that use more static HTML markup rather than a heavy use of JavaScript to load content dynamically.
8. Event-Level Freshness: reveals how recent or up-to-date the information is. We measure Event-level Freshness (ELF) by estimating how up-to-date a record is compared to the underlying event. Because a unique definition of the underlying event is hard to produce, ELF is a quali-quantitative measure, albeit a useful one in the context of EFRA Project. Also, ELF analysis is conducted on the most recent data points to quantify the impact the source could have AS IS on the envisioned platform. Generally speaking, we assigned ELF to each source in the following way:
    - days: alerts, notifications and news are published in hours or few days following the event the alert/notification/topic described in the news, typically without in-depth analysis of the underlying phenomenon (e.g. UK Food Standards Agency or FDA Recalls)
    - weeks: reports or detailed alerts that usually involved further verification from the agency, and therefore are separated a few weeks from the original event that is described (e.g. FDA Enforcement Reports); in

this category fall also most educational content, which can be produced with high frequency, but usually incorporate findings and researches produced in the previous weeks or months (e.g. CDC Youtube)

- months: appropriate for index of scientific publications, that usually involve peer-review time on top of the original paper writing time (e.g. Scopus), sources that publish reports lagging months from the original event (e.g. Australian Department of Agriculture Imported Food Reports)

9. Recommendation: high, medium or low level of recommendation based on the factors listed previously.

Table 4: A high-level view of each data source

Source Name	Size	Cost	Compliance burden	Quality	Reliability	Format	Ease of extraction	Event-Level Freshness	Recommendation
Abu Dhabi Agriculture and Food Safety Authority	very small	Free	Medium	Low	High	HTML	Hard	inactive/static	low
Abu Dhabi Food Control Authority	very small	Free	Medium	Source not relevant	High	HTML	Hard	inactive/static	low
ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail)	small	Free	Low	High	High	PDF	Hard	months	medium
Australian Department of Agriculture Imported Food Reports	very small	Free	Low	High	High	PDF	Hard	months	medium
Austrian Food Safety Authority	very small	Free	Medium	High	High	HTML	Hard	days	medium
BarfBlog	medium	Free	Medium	Medium	Medium	HTML	Easy	inactive/static	medium
Brazilian Health Regulatory Agency	medium	Free	Low	Medium	High	HTML	Hard	weeks	medium
BVL German Federal Office of Consumer Protection and Food Safety	small	Free	Medium	High	High	HTML	Hard	weeks	medium
Canadian Food Inspection Agency	medium	Free	Medium	High	High	HTML	Hard	days	medium
CDC (Centers for Disease Control and Prevention) YouTube channel	small	Free	Low	High	High	Video	Easy	weeks	high
EFSA (European Food Safety Authority)	medium	Free	Low	High	High	XML	Easy	weeks	high
EFSA (European Food Safety Authority) YouTube channel	very small	Free	Low	High	High	Video	Easy	weeks	high
EU Knowledge centre for food fraud and quality	very small	Free	Low	Medium	High	HTML	Hard	days	medium

FDA Enforcement Reports	very small	Free	Low	High	High	CSV/API	Easy	weeks	high
FDA Import Alerts	small	Free	Low	High	High	HTML	Hard	days	medium
FDA Import Refusals	large	Free	Low	High	High	CSV	Easy	weeks	high
FDA Inspections Citations	large	Free	Low	High	High	Excel	Easy	weeks	high
FDA Recalls	very small	Free	Low	High	High	Excel	Easy	days	high
Food Safety Authority of Ireland	very small	Free	Low	High	High	HTML/PDF	Hard	days	medium
Food Safety Dot Com	small	Free	Medium	Medium	Medium	HTML	Hard	days	low
Food Safety News	small	Free	High	Low	Medium	HTML	Easy	days	low
Food Safety Tech	small	Free	Medium	Medium	Medium	HTML	Easy	days	medium
FSIS USDA	small	Free	Low	High	High	HTML	Easy	days	high
LNV (Ministry of Agriculture, Nature and Food Quality)	very small	Free	Low	High	High	HTML	Hard	inactive/static	low
New Food Magazine	small	Free	High	Low	Medium	HTML	Easy	days	low
PubMed	very large	Free	Medium	High	High	XML	Easy	weeks	high
RASFF (Rapid Alert System for Food and Feed)	medium	Free	Low	High	High	CSV | XLS | RSS	Easy	days	high
ScienceDirect	large	Mix	High	High	High	PDF	Easy	months	low
Scopus	large	NC	Medium	High (if contents properly filtered)	High	API/CSV	Easy	months	high
TWEET-FID	small	Free	High	Medium	Medium	CSV	Easy	inactive/static	low
X (Twitter)	-	Paid	Low	Medium (if contents properly filtered)	Low	API	Easy	near real-time	medium
UK Food Standards Agency	small	Free	Low	High	High	HTML	Hard	days	medium
USDA youtube	very small	Free	Low	High	High	Video	Easy	months	medium

5 Energy-efficient Cloud/Edge HPC Architecture & Integration Requirements

The focus of the EFRA Task 1.3 is a comprehensive study of existing data—both public and private—as well as computational resources and technologies present in the consortium. These components are currently in use and are being further developed by consortium partners, with the ultimate objective of fortifying AI-enabled food risk prevention through the EFRA Tools.

The ultimate goal is to explore how we can appropriately integrate and improve these existing technologies and approaches to introduce a more green and energy-efficiency approach. This will be achieved through a twofold approach. First, by optimizing the balance between cloud-based computations, which offer vast processing power but consume substantial energy, and edge-based computations, which are typically more energy-efficient and can process data closer to the source. Second, we will focus on delivering advances in green AI, both in terms of training and deployment, which aim to significantly reduce the carbon footprint associated with AI operations.

Our exploration and analysis in this task will provide pivotal guidelines for WP2, WP3, and WP4. In navigating through the intricate landscape of data and rapidly evolving technologies, we aim to create a roadmap towards a greener, more energy-efficient, and ultimately more sustainable AI-enabled food risk prevention system.

Currently, stakeholders in the food safety domain typically rely on their own individual data, platforms and technological solutions. It is very likely that such solutions do not cover all the requirements that the EFRA platform intends to satisfy. Additionally, privacy concerns and strategic business considerations often act as barriers to the sharing of food safety data that could enable the training of more robust and useful food risk predictive models. Moreover, the use of energy-hungry AI in the food safety domain raises concerns about significant energy consumption and its environmental impact. Computational and energy requirements also lead to higher operational costs, limiting accessibility to and sustainability of AI-driven solutions.

The EFRA platform, dedicated to food risk safety analysis, aims to address the above-mentioned shortcomings and will be enforced by federated and micro-service principles. The EFRA platform will have to satisfy the following requirements:

1. **Multi-tenancy**: The EFRA platform should have the capability to support multiple tenants or users simultaneously. This means that various stakeholders should be able to access and utilize the platform without compromising data security or performance. Multi-tenancy ensures that different entities can collaborate on food safety without interference.
2. **Scalability**: The platform should be designed to handle growing demands and increased data volumes. As more stakeholders join and contribute to the food safety efforts, the EFRA platform must be scalable to accommodate additional users, data sources, and computational requirements without significant performance degradation.
3. **Optimized resource usage and power consumption via green and hardware-aware scheduling algorithms**: To minimize energy consumption and operational costs, the platform should employ advanced algorithms that optimize the allocation of computing resources. This could include, for example, using "green" energy sources where possible and being aware of the hardware capabilities to efficiently distribute workloads.

4. **Seamless distribution and management of workloads across different geographically distributed participants**: Given that food safety is a global concern, the platform must efficiently handle workloads that are distributed across various geographic locations. This ensures that data and computing resources are effectively utilized regardless of where they are situated.
5. **Support of distributed and federated AI learning**: To build robust food risk predictive models, the platform should support distributed and federated AI learning. This means that machine learning models can be trained collaboratively using data from multiple sources while respecting privacy concerns and data ownership. This approach enables the creation of more accurate and generalized predictive models for food safety.

In summary, the EFRA platform aims to be a comprehensive solution for addressing the limitations in the food safety domain. It seeks to enable multi-party collaboration, accommodate scalability, reduce energy consumption, optimize resource allocation, and support advanced AI learning techniques to enhance food safety efforts on a global scale.

5.1 Data Collection and Survey Results

In this section, we present the outcomes of a survey conducted in collaboration with EFRA Use Case Partners and Agroknow. The primary objective of this survey was to collect vital insights into EFRA partners' technical platforms, with a specific emphasis on aspects such as data infrastructure, computational resources, and current utilization of AI-based predictive/decision-support models. The survey was designed to provide EFRA technical partners with a comprehensive understanding of the key aspects that underpin the objectives of the EFRA project and their accomplishment through the EFRA use cases. We placed a particular emphasis on exploring ways to seamlessly integrate and enhance existing technologies and methodologies. Our goal is to establish a robust foundation for the EFRA Platform while introducing more green and energy-efficient practices into the project's framework.

Survey Design and Structure

To ensure comprehensive coverage of key aspects, we designed a structured survey comprising two parts: cross-scenario questions and scenario-specific questions.

5.1.1.1 Cross-Scenario Questions

These questions were formulated to collect general information that are applicable to all use case partners:

1. **Overall Architecture**: Partners were asked to provide an overview of their data collection and storage architecture, including logical modules and interconnections.
2. **Computational Resources**: Information on the types of computational resources (server, desktop, laptop) in use was collected, along with approximate specifications such as CPU cores, RAM, and GPUs.
3. **Data Description**: Questions aimed to elicit insights into partners' data, including the number of records, data ownership (private or public), data reliability, and known data-related issues. Data collection methodologies were also explored.
4. **Data Storage**: Partners were queried about their data storage solutions, distinguishing between database system and file system solutions. For database systems, they specified the type (e.g., RDBMS, NoSQL, Graph) and provided rough size estimates.
5. **Data Exposition**: Partners who exposed data through services (e.g., REST APIs) were asked for technical specifications and client reference implementations.

6. **AI Prediction Models**: Partners shared details about AI models in use, including the underlying technologies (e.g., linear models, decision trees, neural networks), information regarding the training process, the model complexity (e.g., number of hyperparameters), and some key performance metrics (latency and accuracy above all). Resource requirements in terms of CPU, GPU, RAM, and Docker image availability for benchmarking were also covered.

5.1.1.2 Scenario-Specific Questions

In addition to the cross-scenario questions, scenario-specific questions were tailored to the characteristics and needs of each use case scenario:

Agricultural Use-Case (Leader: Agrivi)
Questions delved into Agrivi's reliance on cloud computing, the specific cloud provider being used, and resource allocation strategies to meet demand. Considerations for exploring prediction-as-a-service and expected request rates were also explored.

Regulatory Use-Case (Leader: SGS)
Queries revolved around SGS's strategy for serving predictions, including cloud computing reliance and the specific cloud provider being used. Computational needs for Large Language Models and strategies for computational demand reduction were also investigated.

Poultry Use-Case (Leader: MOY Park)
MOY Park's potential to deploy container-based applications for federated learning framework setup was discussed. Details regarding their current computational resources and willingness to share them for distributed AI training were inquired.

Survey Results

In this section, we present the findings obtained from the EFRA Partners participating to the survey in response to the questionnaire. The results are presented on a per-partner basis. The data collected provides valuable insights into key aspects related to data availability, computational resources, and the deployment of AI models within each use case scenario.

5.1.2.1 Agricultural Use-Case (Leader: Agrivi)

Agrivi aims to improve its pest prediction service offered to its clients. To this end, EFRA leverages Agrivi data from weather stations, soil sensors, and inputs gathered through direct scouting activities conducted by the farmers using the service. This EFRA use case aims at developing and integrating more sophisticated predictive algorithms, which would use AI to forecast pest invasions with greater accuracy and suggest optimized responses. This development could lead to more effective pest management strategies and potentially higher crop yields.

1. **Overall Architecture**

   The logical architecture of the Agrivi Pest-Prediction task and the logical flow is depicted in the Figure 1.

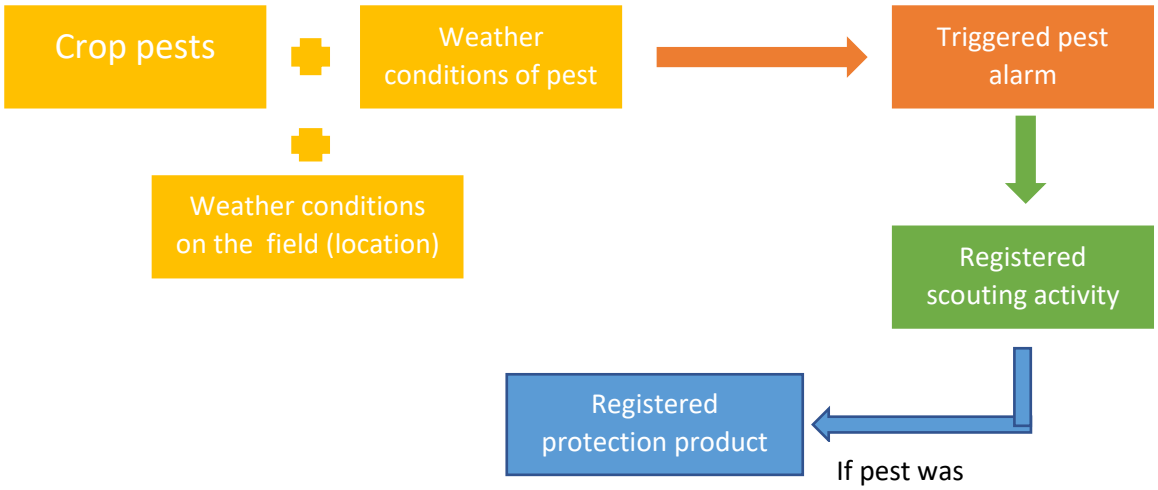


Figure 1: A logical diagram of AGRIVI's computation infrastructure

The modules are part of the Agrivi Farm Management Software (FMS) and it is not a separate entity. It is included in the code of the FMS, and the data is stored into the FMS database.

2. **Computational Resources**
   As it is part of the FMS, Agrivi is not fully aware of the specific computational requirements needed for supporting the pest-prediction use case.

3. **Data Description**
   The datasets currently employed are the following: crop pests, weather conditions for pest occurrence, weather conditions on fields, triggered pest alarms, registered scouting observations, registered protection products usage. All data sets are privately owned. Datasets resulted as farmer's data entry, such as registered scouting observations and registered protection products usage, are less reliable because of inconsistency in data registration from the farmer's side. The weather conditions for pest occurrence data set have slightly lowered reliability due to the changing of weather conditions under which crop pests occur (due to climate change). Datasets such as crop pests, weather conditions on fields, triggered pest alarms, as part of AGRIVI standard database that came from scientific sources, are highly reliable.

4. **Data Storage**
   Agrivi data is stored in the FMS database.

5. **Data Exposition**
   Agrivi is currently not exposing their data to the public through an API nor an extracting tool is provided.

6. **AI Prediction Models**
   Agrivi is currently employing a rule based system for pest prediction. The plan is to use properly AI prediction model in the future. Currently they are working on the solution and this is one of the main Agrivi's objective within the EFRA project.

7. **Cloud Computing exploitation**
   Agrivi is exploiting cloud based resources on Azure.
8. **Matching On Demand needs**

VertexAI is taking care of automatically provision required hardware for the purposes of the AI training. Moreover, to reduce the cost pre-emptible nodes on Google Cloud are exploited.

9. **Predictions on the EFRA platform**
   Agrivi is not interested in running on a prediction-as-a-service basis, with computation performed on EFRA servers, but rather to build an intelligent AI solution to be deployed on their server for the purposes of their clients.

5.1.2.2 Regulatory Use-Case (Leader: SGS)

As part of the EFRA project, SGS is committed to creating an automated regulatory analysis and summarization module that harnesses extensive regulatory data. Currently, interpreting and summarizing this data demands substantial manual effort from users. The primary goal of this use case is to harness advanced AI solutions to significantly alleviate the need for manual intervention. This will be achieved by generating key summaries and extracts from regulatory texts, ultimately leading to a substantial improvement in user experience and operational efficiency measured also as by considering computational complexity end energy consumption.

1. **Overall Architecture**
   The logical architecture of the backend modules of the SGS platform, and how these modules are connected to each other are provided in Figure 2.

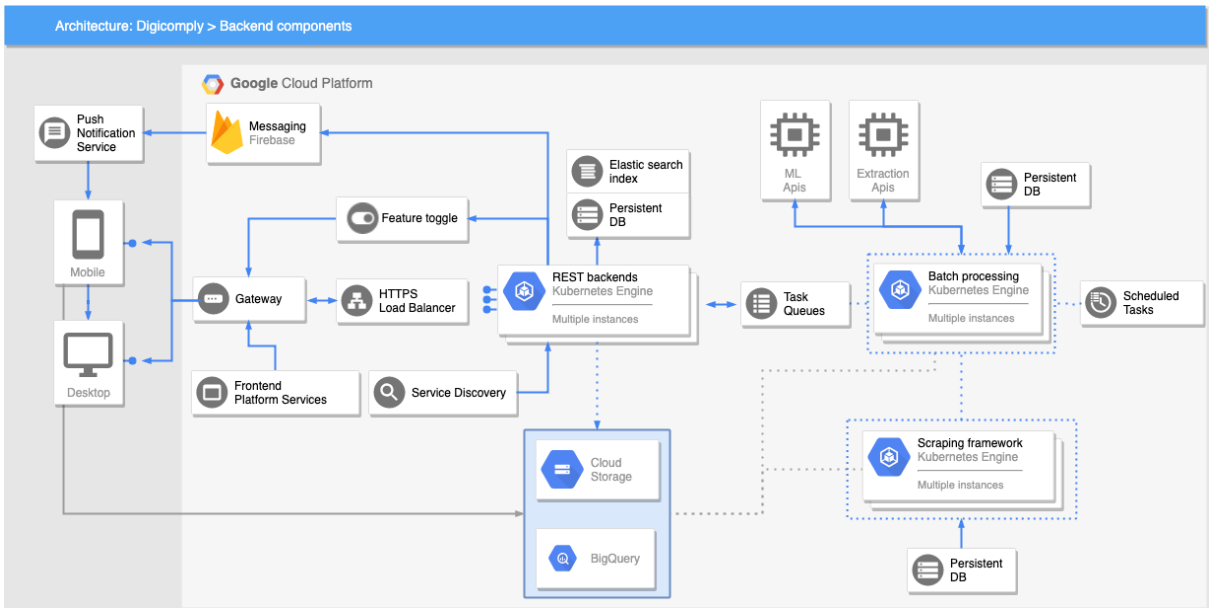


Figure 2: A logical diagram of SGS's compute infrastructure

Among the several modules, the ones responsible for data collection are:

- Scraping Framework: component used to scrape data from public websites. Utilises standard scraping mechanisms such as IP rotation and Rate limiting, following ethical scraping rules.
- Batch processing: Tools and pipelines used to connect all our ML models, extraction models.

2. **Computational Resources**

SGS does not use on-premises hardware, but instead relies exclusively on cloud resources. The type and number of resources employed depends on the current needs.

3. **Data Description**
The table below, which is Table 5, provides an overall description of SGS data, detailing also the number of records for each data type.

Table 5: SGS Data summary

Data Type	Volume (K #records)
Regulatory documents	190
Regulatory Limits	1.339
News	2.157
Scientific papers	2.905
Risk Triggers Food Safety Surveillance	~2-5
Food Safety Incidents	176
Risk triggers Lab Data	7.500
Food Safety Hazards	1.5
Food Safety Products	2
Organization/Brand name	~2-5

4. **Data Storage**
SGS employs a mixture of different technologies for storing data, depending on the specific nature of the data. Specifically, the technologies employed are:
   - Raw data storage - Google Cloud
   - Wide column store - BigQuery
   - Relational Database - Postgres
   - Embeddings storage - custom FAISS / Weaviate
   - Indexing - ElasticSearch

5. **Data Exposition**
SGS is exposing their data through standardized APIs[4]. However, they also have custom integrations with some clients.

The supported data formats are the following:

[4] https://digiciomplypostchanges.docs.apiary.io/

- CSV/XLXS
- JSONL
- Dumps of raw data (raw html, plain text or our custom JSON representing the text in the article)

Formats can widely differ based on what kind of datasets need to be extracted (e.g., articles will be different from maximum residue limits).

6. **AI Prediction Models**
   Depending on the specific needs, SGS is currently employing the following AI technologies:
   - Prediction task: Prophet library with some custom setup.
   - NLP Tasks: classification, multilingual classification, NER, relation extraction (between extracted NER terms). Most of this is done on custom models trained within FLAIR NLP.
   - Image Object Detection/Image classification: Custom models to detect relevant sections of a product artwork or classify images. YOLO for object detection, TensorFlow for the classification.
   - Summarisation/Extraction Pipeline: Different combinations of pipelines are built on embedding data into weaviate store and then model to retrieve valid sections of documents which can be integrated with GPT like services.

   All models are standalone components which can be deployed using docker. Each prediction model contains:
   - REST API for inference
   - REST API for batch inference
   - UI for testing

7. **Cloud Computing exploitation**
   For the training purposes SGS is employing VertexAI jobs on google cloud platform. Some processing pipelines can utilize google dataflow (built on top of airflow framework). For inference SGS have a custom processing pipeline built on Kubernetes.

8. **Matching On Demand needs**
   VertexAI is taking care of automatically provision required hardware for the purposes of the AI training. Moreover, to reduce the cost preemptible nodes on google cloud are exploited.

9. **Computational needs for Large Language Models**
   Running fine-tuned LLMs requires commitment of several GPUs for both training and inference. Currently SGS is trying to reduce as much as possible self-hosted models and fine-tuning pipelines in favor of 3rd party solutions. For lowering processing costs SGS is trying to offload to preemptible machines.

5.1.2.3 Poultry Use-Case (Leader: MOY Park)

The poultry use-case investigates data-driven strategies for preventing *Salmonella* within the supply chain, uncovering its causal relations and provide real-time alerting for hatchery health monitor. These objectives are pursued based on the exploitation of risk assessments, lab test results, WGS analysis in combination with AI techniques.

MOY Park utilizes a software solution, which is MTech Systems[5], across all of their facilities and for all of their tasks such as lab result tracking and business intelligence. The data is stored in local databases and synchronized with their national data center. The MTech system enables MOY Park to track all steps of their production, test, and analysis steps. All of their data can be exported in CSV format and shared with EFRA consortium as an offline copy and providing access to a highly secure virtual machine that contains the data. MOY Park will prepare an API in order to enable real-time data sharing as well. The real-time data sharing will enable efficient updates and operation of machine learning models.

The key points we took from our meeting is that (a) they use the same system, MTech, across all their facilities, (b) data is stored in local databases and can be exported as csv files, or provided as APIs for real-time integration, (c) there is the possibility to deploy local VMs to their national data centers to run limited AI model training and exploitation.

1. **Overall Architecture**
   MOY park utilizes MTech Systems in all their facilities for all of their tasks. This software facilitate entering and storing records of their data regarding its type. This software enables synchronization of the local data with MOY Park's national data centre.

2. **Computational Resources**
   Desktop computers are used to run MTech software. The local copies of the data are synchronized with MOY Park's national data centre.

3. **Data Description:**
   The records in the database are observations such as lab results as containing *salmonella* or not with date, flock, and age. As the MTech software has been used since 2018, it is not possible to obtain data for the period before this period.

4. **Data Storage:**
   The data is stored in local databases of each computer utilized in labs and farms. This data is transferred to MOY Park's national data center regularly.

5. **Data Exposition:**
   MOY Park has the possibility to extract their data in CSV format as an offline copy. A sample of data will be shared with the EFRA consortium as an offline copy. Moreover, they will set up a highly secure virtual machine for providing access to the whole dataset. Finally, the virtual machine will contain access to an API they will setup for real-time data utilization for machine learning scenarios.

6. **AI Prediction Models:**
   MOY Park does not have any capacity for developing or using machine learning models. The company will develop this capacity in the scope of EFRA project using virtual machines that facilitate data access and machine learning model development and utilization.

7. **Cloud Computing exploitation:**
   Not applicable.

8. **Matching On Demand needs:**

[5] https://mtechsystems.io

MOY park has allocated resources for providing sample data and setting up virtual machines that will be accessible to EFRA consortium for data access and machine learning model development and utilization.

9. **Distributed AI training:**
   MOY Park will set up additional virtual machines that contain data and capacity to development machine learning models for additional MOY Park facilities across their supply chain.

5.1.2.4 Agroknow Data Platform

Within the EFRA project, Agroknow's platform plays a pivotal role in all EFRA activities and has the potential to bootstrap the development of a shared infrastructure. With this objective in mind, we extended the survey also to Agroknow, aiming to gather information regarding their platform, computational and storage resources, as well as their utilization of AI prediction models.

1. **Overall Architecture**
   Agroknow's Data Platform is an end-to-end, data-driven ecosystem engineered to aggregate, enrich, and analyze data related to food safety. Utilizing state-of-the-art technologies in crawling, natural language processing (NLP), and machine learning, the platform aims to empower stakeholders with actionable insights. The overall architecture is depicted below in Figure 3.

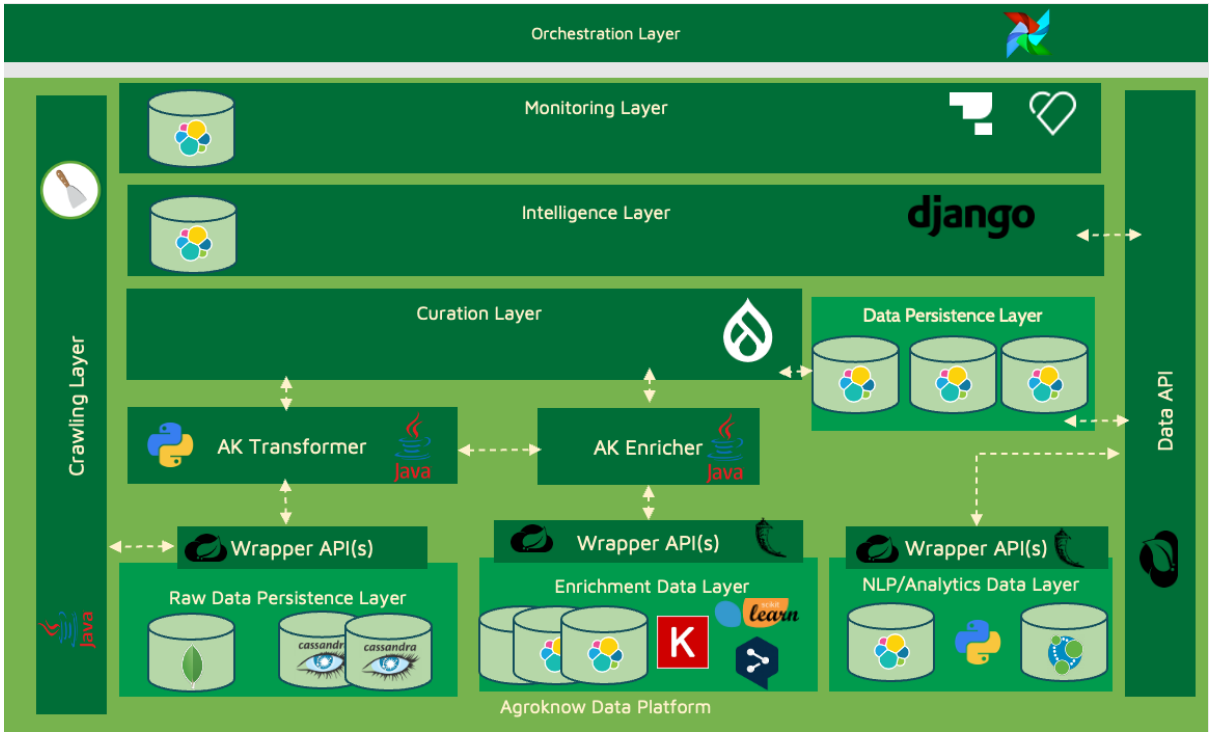


Figure 3: Agroknow's Data Platform

The architecture is partitioned into several layers, each with a distinct role and set of responsibilities:

- **Crawling Layer**: The Crawling Layer employs specialized crawling software and custom scripts to scrape data from public food safety authority websites. It is equipped with anti-rate-limiting mechanisms and IP rotation capabilities to avoid getting blocked. Data is extracted in a variety of formats such as HTML, XML, PDF, and JSON, and then pushed to the Raw Data Persistence Layer.

- **Raw Data Persistence Layer**: This layer is responsible for the initial storage of raw, unprocessed data. This layer uses Apache Cassandra for the primary storage of raw, unprocessed data. Cassandra's distributed architecture is leveraged to handle the high-volume, high-velocity data. Data is indexed with metadata for optimized retrieval and management.
- **Enrichment Data Layer & NLP Data Layer**: The Enrichment Data Layer uses extract-transform-load (ETL) processes to clean and filter the raw data. The NLP Data Layer leverages natural language processing algorithms to convert textual data into machine-readable formats. Both layers cooperate to prepare the data for storage in the Elastic database.
- **Data Persistence Layer (Elastic Database)**: Structured and enriched data is stored in an Elasticsearch database. The database is designed to follow a specific schema that enables efficient querying and data retrieval. It employs sharding and replication strategies for fault-tolerance and high availability.
- **Curation Layer**: The Curation Layer provides an interface equipped with front-facing tools for expert human curators. These tools include annotation interfaces, data validation dashboards, and content management systems (CMS). Curators enrich machine-generated data to eliminate false positives/negatives and improve overall data quality.
- **Intelligence Layer**: The Intelligence Layer houses AI models built on PyTorch. These models execute time-series predictions, anomaly detection, and other advanced analytics over the curated data. The layer is optimized for high computational performance and scalability.
- **Monitoring & Orchestration Layers**: These layers ensure the infrastructure's health and optimal operation. Monitoring tools track performance metrics and alert on system failures, while orchestration tools (Apache Airflow) manage the deployment and scaling of services.
- **Data API**: Finally, the Data API exposes endpoints that allow third-party applications to access the platform's data and insights. Built with RESTful principles, the API supports various authentication methods and rate-limiting capabilities.

2. **Computational Resources**

Agroknow has no on-premises hardware, but instead they rely exclusively on cloud resources. The data platform (everything but the Intelligence Layer), is running on two CPX21 with 4 Intel vCPUs, 4 GB RAM, 80 GB disk space and 160 GB of extra storage shared between the two VMs.

For training purposes and for our research on AI models, Agroknow uses Google Colab. The system is using a Tesla T4 GPU, which is based on Turing architecture. Tesla T4 is a GPU card based on the Turing architecture and targeted at deep learning model inference acceleration. For the exploitation of the AI models (Intelligence Layer), Agroknow uses one CX41 VM with 4 Intel vCPUs, 16 GB RAM, 160 GB disk space.

3. **Data Description**

All data are from public sources and in the public domain. The main data records are food recalls, collected by the Agroknow Data Platform and going back to 1980. There are 838.000 records so far, growing by a rate of approximately 10% per year. They also collect lab test results as announced on an annual basis by public food safety authorities. The relevant data records are more than 200M and are collected using the same process as already described.

Agroknow also possess other data types as well, most of which are not actively scraped any longer. They are presented below:

- Country Risk Indicators (12K)
- Country Corruption Indicators (3K)

- Food Production Data (150K)
- Food Trade Data (34M)
- Companies Data (680K)
- Commodity Prices (390K)
- News (67K)
- Social Media Data (7K)
- Legislation (280)
- Outbreaks Information (57K)
- Product Brands Recipes (26K)
- Sensor Installations & Readings (12M)
- Inspections (227K)

4. **Data Storage**

Agroknow employs Apache Cassandra as the initial data store for raw, unprocessed data, capitalizing on its distributed architecture for high availability and fault tolerance. Data is partitioned across multiple nodes to enhance read and write throughput, and Cassandra's tunable consistency is exploited to optimize data integrity and availability as per the use case requirements.

ElasticSearch on the other hand is exploited for processed and machine-readable information. The Elasticsearch database is engineered to comply with a specific schema designed for efficient querying and optimized data retrieval. It leverages inverted indices and employs sharding and replication strategies to achieve high availability and fault tolerance. Elasticsearch is particularly beneficial for its fast, near real-time search capabilities and analytics.

The total data footprint across both databases is approximately 200GB. However, in a live production environment, the active data set we operate on is around 60GB. This operational data set is managed to fit into Elasticsearch's hot storage tier to ensure low-latency data retrieval and analytics.

Both databases are backed by SSDs to reduce I/O latency, and appropriate backup and snapshot strategies are exploited to protect against data loss. The databases are monitored for performance metrics, and capacity planning is done based on the data growth rate to ensure scalability.

5. **Data Exposition**

Agroknow provides access to their data through a detailed API service[6].

The export format is either in JSON or CSV. The properties are shown below. Data samples are available here.

[6] The documentation of this API is on https://docs.agroknow.com/.

Figure 4: Food Recall Properties

6. **AI Prediction Models**

For Agroknow time series predictions, the Prophet solution is exploited. This AI model can be made available to the EFRA infrastructure through an API.

For NLP, especially for distinguishing between food safety incidents and other food domain or unrelated data (NLP Classification) Agroknow started experimenting with and fine-tuning RoBERTa. The model, once ready, can be made available as a docker image. As of now, the relevant model has the following characteristics:

- **Model type**: RoBERTa, is a state-of-the-art natural language processing (NLP) model and it is based on the Transformer architecture and is designed to understand and generate human language text. RoBERTa is pre-trained on an extensive corpus of text data from the internet and has a remarkable ability to comprehend the nuances of language, making it a powerful tool for a wide range of NLP tasks, including text classification, sentiment analysis, question answering, and language generation.
- **Training process**: a standard procedure using a 10-fold cross validation has been exploited
- **Training requirements**: The resources of Google Colab Pro - V-100 NVIDIA GPU & 32GB of RAM have been exploited for the purpose of training the RoBERTa model.
- **Complexity**: No hyperparameter optimization has been investigated since the model reached a good level of performance (in the standard classification metrics being monitored) so it didn't need further improvement. Therefore Agroknow manually experimented with the original parameters of the model and the complexity was relatively low.
- **Performance**: Accuracy: 93.42%. Training time: 3 hours on a dataset composed by around 20K records, with the average record (i.e., text) being around 1700 characters in length (after removing stop words).

7. **Cloud Computing exploitation**

Agroknow is exploiting cloud based resources provided by Hetzner for inference tasks, and Google Colab for training and research tasks.

8. **Matching On Demand needs**
   Since cloud bases resource are exploited, the cloud provider is taking care of automatically provision required hardware allowing to match on demand needs.

9. **Computational needs for Large Language Models**
   The RoBERTa model is currently being investigated on a Google Colab Pro infrastructure exploiting a V- 100 NVIDIA GPU with 32GB of RAM.

6 Public & private data for AI training and data sharing requirements

Our primary focus in this section is to gather and collate the requirements needed to address four key objectives:

- First, we aim to deploy a privacy-preserving AI training approach directly over private and sensitive food safety datasets, ensuring the confidentiality and security of the data while leveraging its value.
- Second, we plan to devise ways to access, process, and combine public and private data sources and streams. This will help us build a comprehensive and insightful knowledge base.
- Third, in scenarios where the privacy-preserving AI training approach cannot be directly deployed, we will identify alternative strategies to enhance private FAIR (Findable, Accessible, Interoperable, and Reusable) data sharing. These strategies may include techniques such as data aggregation or anonymization.
- Lastly, we aspire to facilitate FAIR data interoperability through the use of existing and novel data and metadata standards. This will ensure our data analytics powerhouse remains versatile and capable of effectively interfacing with various data types and structures.

The insights and directions gleaned from this task will offer direct guidelines for the data sharing approaches in WP4 and the use-case implementations in WP5. The goal is to build a reliable, efficient, and ethically responsible data analytics powerhouse aligning with the specific needs of our use-case scenarios.

6.1 Introduction

Food data is often scattered across various sources, like different organizations, various competing farms or companies, or isolated laboratories, making it difficult to effectively research and develop models. Sometimes, the data can be pooled together and used for conducting statistical analysis, modeling or using machine learning for prediction and assessment. However, in other cases pooling the data is not desirable, for instance if the data is sensitive or confidential. When data is obtained from companies, it is often difficult to convince them to share it with a centralized system as it may potentially hurt their reputation if the data got leaked or provide the competition with an undesirable advantage. Data sharing may sometimes also be difficult due to legal regulations, for instance when it concerns personal or in other ways confidential data. Companies situated in the European union or handling personal data of people inside the European Union must for instance abide by the GDPR regulations. These obstacles also arise for the food sector making food safety research that requires data from various sources challenging [1]. Two methods were recently proposed to improve the data sharing possibilities for food safety research: differential privacy and federated learning [2]. Especially federated learning is appealing because of its privacy by design character. Using federated learning allows the data to stay within its owners' premises. The following section will lay out an introduction to federated learning, its challenges, and concrete requirements for its employment in this project.

6.2 Federated Learning

Federated learning (FL) is a term introduced by Google in 2016 in the paper by McMahan et al. [3]. It was presented there as *a decentralized approach of leaving the data distributed on the mobile devices, and learning a shared model by locally-computed updates*. FL involves a group of client nodes that each contain a portion of data. In the general case the global model is trained locally in a number of communication rounds in which the model uses the local data for training. Then the updated model weights or other parameters are communicated back to the central server. This procedure is repeated until some predefined goal is reached, like the maximum number of communication rounds or some minimum requirement on the performance.

It is however important to note that this initial proposal for federated learning focused on combining data from many mobile devices with the same features. This setting now can be defined as *cross-device*, *horizontal* federated learning with a *centralized architecture*. After the paper by McMahan et al was published, the field expanded to include other variants of the federated learning approach. We will now elaborate on the different categorizations that can be used to describe federated learning, their applications, and more.

In practice, federated learning can involve different objectives and architectures. We can categorize architecture into two broad categories: centralized and decentralized. The next essential division is in the data partitioning: horizontal, vertical and transfer learning. Finally, we can also roughly categorize the federated learning setting by the number of participants and their trustworthiness using the cross-device/cross-silo distinction.

The main federated learning architecture division is between *centralized* and *decentralized* federated learning. The former is the more typical view of federated learning with one central server and multiple client nodes. The data resides at the nodes and the central server communicates with the nodes and aggregates the training results. This type or architecture can also be called a star network with the central server in the middle. It is the most straight-forward way of conducting federated learning. One of its advantages is the central server can carry a considerable portion of computational and storage burdens [4]. This lessens the hardware requirements on the client servers. Furthermore, having a single central server makes it easier to manage, regulate and protect [4]. The alternative decentralized version, with subcategories such as gossip learning [5] or peer-to-peer federated learning [6], omits the central server. Instead, the client nodes communicate directly with one another. One advantage of this approach is that the solution to the communication bottleneck with the central server. Allowing the nodes to communicate with their neighbors reduces that bottleneck. Furthermore, with a centralized system, the server becomes an important single point of failure.

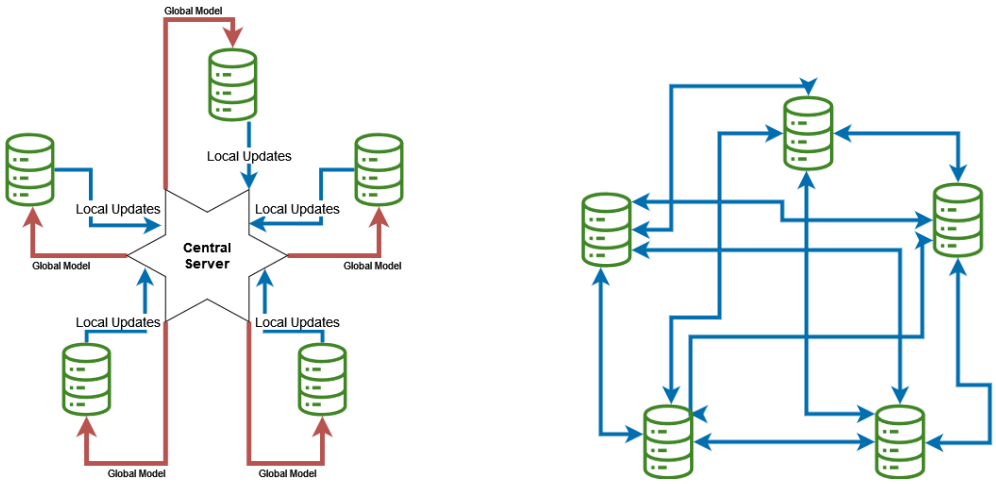


Figure 5: Left: Centralized federated learning Right: Decentralized federated learning

Another important distinction is that between cross-silo and cross-device federated learning [7]. In practice, the same methods can often be applied to both kinds, however the important distinction is in the priorities and assumptions. Participants in cross-silo federated learning are usually different organizations, research institutions, data centers, companies etc. With cross-silo federated learning we may expect more reliable communication, more computational resources and large data sets, called data silos [4]. Since for cross-silo federated learning there are usually a lot less participants and the participants are also carefully selected, it can be expected that there is less risk of a malicious node

present in the federated learning system. Moreover, communication challenges also become lessened when the number of participating devices decreases, reducing the bottleneck at the central node. There will also be fewer nonparticipating client nodes, because it is easier to enforce system and performance requirements and availability requirements on a limited number of participants. On the other hand, the presence of slow or not-responding client nodes might have a larger influence on the training, since dropping a participating node becomes more drastic in the cross-silo setting. The participants of cross-device federated learning may include thousands or millions of mobile devices or internet of things devices. It faces larger hardware heterogeneity by design but offers a large variety and quantity of data.

Federated learning can also be determined by the way the data is partitioned amongst the nodes. We will define those partitions based on the description from a survey on federated learning by Zhang et al. [8]. In *horizontal* federated learning, each client node has different samples with the same, or largely similar features. Horizontal federated learning aims to increase the number of training samples by using datasets from different sources.

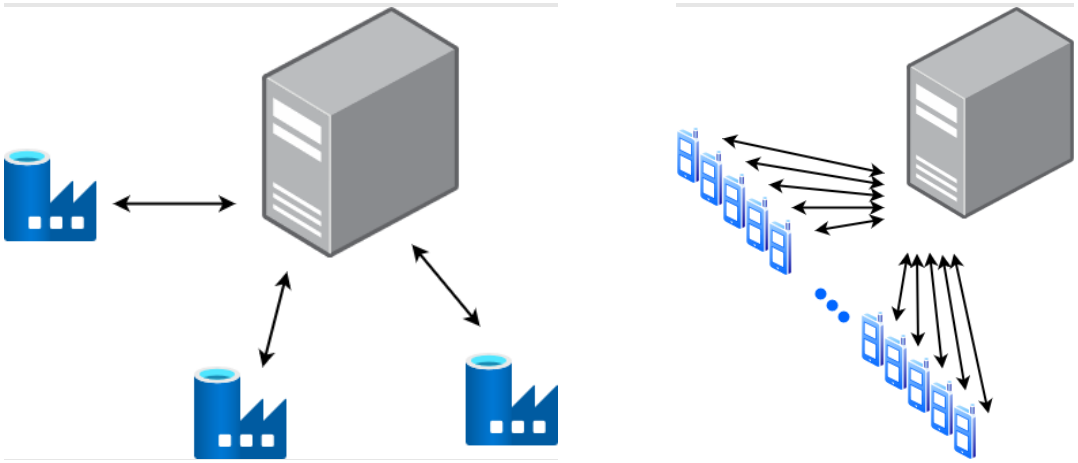


Figure 6: Left: Cross-silo federated learning Right: cross-device federated learning

The second type of federated learning is called *vertical* federated learning where there is a large overlap in samples and a smaller overlap in features. Vertical federated learning is used to increase the feature space of the training data and the training data is limited to the samples that overlap across the client nodes. Often the label is held by one of the clients and the participation of all clients is necessary for inference, because they hold the relevant input information.

The third type is called federated *transfer* learning. In transfer learning, neither the samples nor the features overlap or they overlap very slightly [9]. Transfer learning aims to improve both the number of samples and the number of features by combining data from different, yet related domains or tasks. Federated transfer learning trains the models in such a way that the model can be used in the target space using target-specific features after first being trained in the source domain space with other features. It leverages the knowledge acquired from one domain to the other [10]. After the completion of the federated learning training, the model can be finetuned to the local domain to fit the needs of the client better, while still retaining knowledge gathered from the other related domain.

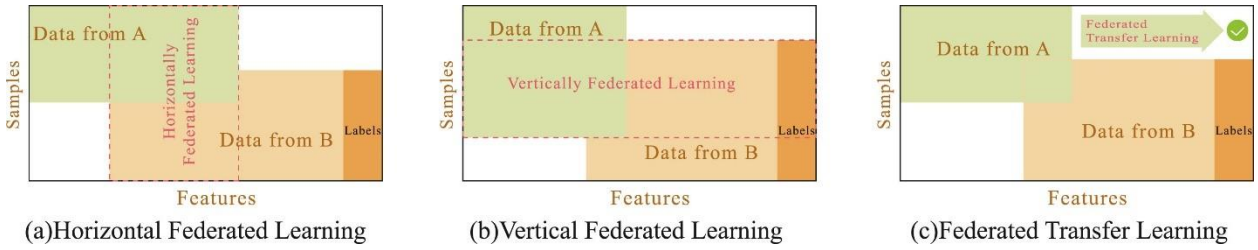


Figure 7: Data partition across client nodes: possible configurations [8]

Finally, each federated learning architecture must implement some way of combining the model updates: *model aggregation.* The model aggregation determines how the updates are combined to update the global model. One very widely spread method called federated averaging or FedAvg introduced by McMahan et al. [3] calculates the weighted average of the local model weights after each communication round. FedAvg takes a weighted average of all model updates after each communication round. Many other aggregation methods like FedProx [11] and RFA [12] exist that try to compensate for some of the FedAvg's shortcomings like its lack of robustness against corrupted updates.

6.3 Federated Learning Challenges

Federated learning is a new field with still a lot of open challenges. One can read more in depth about those challenges in the papers by Li et al [13], Zhang et al. [8], Kairouz et al. [7] and others. The main challenges that as of today do not have one general solution are expensive communication, system heterogeneity, data heterogeneity, privacy and security, and participation incentive generation.

The first challenge is that of expensive communication. High communication costs can form a bottleneck during training, especially in the cross-device scenario. The communication depends on the FL architecture type but also involves choices of synchronous or asynchronous communication, number of communication rounds), number of training passes within each communication round, number of nodes per round, and the size of messages communicated back to the server.

Next, one also needs to consider system heterogeneity, especially in the case of cross-device FL. The participating devices might differ in terms of hardware, connectivity and battery power [13]. This may lead to issues with devices holding up the training process due to failure or slow computation. Federated learning systems must be designed in such a way to tolerate problems arising from heterogenous hardware.

The third challenge involves data heterogeneity, which some might consider a defining feature of federated learning differentiating it from distributed learning [7]. Each local dataset might follow a different distribution because individual circumstances of each data client can introduce biases, leading to data that is not independent and identically distributed (non-IID data). Skewed non-IID data can introduce weight divergence during training, meaning that models with the same initial weights start to deviate too much from one another [14]. The main three strategies to deal with non-IID data for FL training are data based (data sharing and data augmentation), algorithm based (local fine-tuning, multi-task learning and more), and system based (client selection, system level optimization) [15].

Next challenge is the privacy and security of the system. Although FL's main selling point is the preservation of privacy, it is still not guaranteed by the framework. For example, in some cases it may be possible to retrieve data points previously seen by the model through membership inference attacks [16]. To prevent this, some additional steps can be taken like including differential privacy measures, homomorphic encryption of the data and privacy preserving protocols like secure function evaluation and secure multiparty computation [13].

The last challenge is incentive creation. Federated learning is not possible without participating data owners. Each participant shares their valuable collected knowledge, albeit without directly sharing the data. Furthermore, facilitating federated learning may also mean required investment in hardware, power consumption and skilled work hours. Collaboration is only possible if there is enough to gain from it, either in the form of improved models or payment. To improve the incentive, a federated learning scheme should for instance limit freeloaders that only profit from the knowledge of other clients and improve the fairness of the trained model. When fairness is considered, the participating clients will have more motivation to contribute quality data. In a cross-silo setting, freeloaders can also be limited through regulation and contractual agreements between the parties.

6.4 Requirements of a Federated Learning Setting in the scope of EFRA

The EFRA consortium will implement a federated learning system on MOY Park's data. The data from different MOY Park facilities will be stored in separate virtual machines that has the capacity to support the machine learning training and prediction. When the models are ready, they will process real-time data that will be accessible via an API that will be set-up by MOY Park and be accessible via the virtual machines utilized for training machine learning models. This setup will be a cross-silo federated learning.

6.5 References

[1] A. Durrant, M. Markovic, D. Matthews, D. May, J. Enright, and G. Leontidis, "The role of cross-silo federated learning in facilitating data sharing in the agri-food sector," Comput. Electron. Agric., vol. 193,Feb. 2022, doi: 10.1016/j.compag.2021.106648.

[2] C. Qian et al., "A perspective on data sharing in digital food safety systems," Critical Reviews in Food Science and Nutrition. 2022. doi: 10.1080/10408398.2022.2103086.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data." arXiv, Jan. 26, 2023. Accessed: Aug. 14, 2023. [Online]. Available: https://arxiv.org/abs/1602.05629.

[4] L. Yuan, L. Sun, P. S. Yu, and Z. Wang, "Decentralized Federated Learning: A Survey and Perspective." arXiv, Jun. 02, 2023. doi: 10.48550/arXiv.2306.01603

[5] C. Hu, J. Jiang, and Z. Wang, "Decentralized Federated Learning: A Segmented Gossip Approach." arXiv, Aug. 21, 2019. doi: 10.48550/arXiv.1908.07782.

[6] T. Wink and Z. Nochta, "An Approach for Peer-to-Peer Federated Learning," in 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), Jun. 2021, pp. 150–157. doi: 10.1109/DSN-W52860.2021.00034.

[7] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Found. Trends® Mach. Learn., vol. 14, no. 1–2, pp. 1–210, Jun. 2021, doi: 10.1561/2200000083.

[8] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," Knowl.-Based Syst., vol. 216, p. 106775, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.

[9] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A Secure Federated Transfer Learning Framework," IEEE Intell. Syst., vol. 35, no. 4, pp. 70–82, Jul. 2020, doi: 10.1109/MIS.2020.2988525.

[10] S. Saha and T. Ahmad, "Federated transfer learning: Concept and applications," Intell. Artif., vol. 15, no. 1, pp. 35–44, Jan. 2021, doi: 10.3233/IA-200075.

[11] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," Proc. Mach. Learn. Syst., vol. 2, pp. 429–450, Mar. 2020.

[12] K. Pillutla, S. M. Kakade, and Z. Harchaoui, “Robust Aggregation for Federated Learning,” IEEE Trans. Signal Process., vol. 70, pp. 1142–1154, 2022, doi: 10.1109/TSP.2022.3153135.

[13] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, May 2020, doi: 10.1109/MSP.2020.2975749.

[14] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated Learning with Non-IID Data,” 2018, doi: 10.48550/arXiv.1806.00582.

[15] H. Zhu, J. Xu, S. Liu, and Y. Jin, “Federated learning on non-IID data: A survey,” Neurocomputing, vol. 465, pp. 371–390, Nov. 2021, doi: 10.1016/j.neucom.2021.07.098.

[16] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership Inference Attacks Against Machine Learning Models,” in 2017 IEEE Symposium on Security and Privacy (SP), May 2017, pp. 3–18. doi: 10.1109/SP.2017.41.

7 Industrial decision support requirements for risk prevention

In this section, we set out to understand the unique informational needs of the targeted human and expert decision-makers operating within the context of AI-enabled food risk prevention in industrial settings. Recognizing the pivotal role that these individuals play, we aim to illuminate the key information that they require to make effective, informed decisions.

We further specify decision-support use-cases and scenarios, centered around the utilization of TRL3 software tools. These tools, which are already at an experimental proof-of-concept stage, will undergo further enhancement throughout the course of this project.

The requirements that emerge from this exploration will significantly influence and inform the work across all EFRA Work Packages, with particular implications for the use-cases being developed in WP5. By taking a user-centric approach, we aim to ensure that our tools and processes are designed to align with the real-world needs of decision-makers, thereby promoting practical usability and enhancing the effectiveness of AI-enabled food risk prevention in the industrial context.

7.1 Agricultural use-case (Leader: Agrivi)

7.1.1 Scenario AG.1: Enhanced Predictive Capabilities for Pest Alarms

In this scenario, the aim is to boost the pest and disease alarm offered by AGRIVI to its clients.
AGRIVI leverages data from weather stations, soil sensors, and inputs gathered through scouting activities conducted by farmers, enabling improved decision-making processes. These integrated components synergize to provide valuable insights into effective pest and disease management. However, AGRIVI is determined to further advance its existing pest and disease detection algorithm, aiming to provide even more precise and accurate information to farmers.

The enhancement envisioned would involve developing and integrating more sophisticated predictive algorithms, which would use AI to forecast pest invasions with greater accuracy and suggest optimized responses. This development could lead to more effective pest management strategies and potentially higher crop yields.

Software solutions like AGRIVI help farmers manage farm production sustainably by providing valuable insights into crops, crop health, weather, and pests. However, compliance with legal limits goes beyond mere knowledge of regulations.

Current State (before EFRA)

Currently, AGRIVI utilizes data and its proprietary algorithm to enhance farmers' decision-making processes and provide valuable insights for pest and disease management.

Our existing pest and disease algorithm leverages data from weather forecast (no microlocation): temperature, humidity, precipitations, wind...

Our pest and disease alarm further alerts farmers to potential threats, enabling proactive measures for effective management. AGRIVI revolutionizes farming practices through data-driven insights and sustainable agricultural practices.

These existing tools, although efficient, are not as sophisticated or precise as they could potentially be with further enhancements.

Future State (after EFRA)

After the successful completion of EFRA, AGRIVI will offer more advanced predictive algorithms powered by AI.
The enhancement envisioned would involve:

- Weather forecasting data – micro locations
- Detailed feedback from farmers on pests and pest related
- Information from public or private sources regarding changes in pest behaviors / new pest behaviors (caused by climate change, excessive pesticide use, and other factors)

How ?

1. First and foremost, we are considering incorporating information from public or private sources regarding changes in pest behaviors caused by climate change, excessive pesticide use, and other factors. When a specific pest's behavior changes, it implies a modification in the criteria for its occurrence. As a result, the pest alarm and spraying/protection routine need to be updated. Having up-to-date information about pest behaviors ensures more precise predictions and, consequently, improves pest protection protocols.
2. Secondly, through real-life testing with farmers, we will enhance the accuracy of the updated and improved algorithm, specifically in addressing the major risk of **(changing)** pests and diseases in agriculture. By effectively predicting and managing pest and disease outbreaks, the algorithm contributes to better crop protection practices. This, in turn, leads to reduced reliance on unnecessary pesticides, promoting sustainability in agricultural practices. Additionally, by providing farmers with accurate and detailed insights, the algorithm enables them to optimize their pest management strategies, resulting in improved crop yields. The combination of reduced pesticide usage, increased sustainability, and enhanced yields underscores the potential for positive environmental and economic outcomes in farming

To conclude, these enhancements would allow for a significantly more accurate prediction of pest invasions and propose optimized responses based on these predictions. The result would be more effective pest management strategies and potentially higher crop yields due to more precise and timely responses to pest threats.

Envisioned Challenges & Risks

The primary challenges and risks revolve around the development and integration of more advanced AI algorithms. Accurately predicting pest invasions requires in-depth data analysis and a comprehensive understanding of various complex factors. These include climate conditions, specific crop types, regional infestations, and past pest behavior. The algorithm's complexity, combined with the vast amount of data that needs to be processed and understood, presents a significant challenge. Additionally, ensuring the new capabilities smoothly integrate with the existing platform without causing disruptions is another potential challenge.

Relevant Data

Already existing in Agrivi DB	**New data types and potential sources**
AGRIVI can provide historical data about: • pest occurrences (based on sets of criteria) • weather conditions and soil data	Additional data for more accurate predictions include: • New pest behavior data (Public or private sources?) • Confirmation of pest appearance/validity of pest alarm on a micro location • New data stemming from the scouting feature within AGRIVI (farmers feedback, the number of pest per single unit, what part is infected etc...)

• crop types • pesticide usage (report) • Best practices (list of agronomical practices per crop)	
Targeted Users for Piloting Activities	
Users can be farmers who are already using AGRIVI and are open to testing an improved feature. They would ideally have a varied crop portfolio and be in regions with diverse climates and pest populations.	
Support by Agrivi for Scenario Implementation	
AGRIVI can offer support as we integrate this updated version of the algorithm into the platform which will benefit from enhanced capabilities and better prediction accuracy as more parameters are being integrated. AGRIVI can also provide user support to handle any issues or queries, and creating training materials to help users understand and make the most of the enhanced features. Additionally, AGRIVI's team of agronomists and data scientists can provide essential insights and support during the development and testing stages of the new features. Finally, AGRIVI will set up appropriate piloting activities to test the new features with current AGRIVI users. This piloting activities are in the scope of WP5 and will be detailed and reported in D5.3.	
KPIs	
KPI AG1.1	Prediction accuracy > 65% of pest invasions
KPI AG1.2	User satisfaction grade > 7/10 for the new developed features

7.1.2 Scenario AG.2: Regulatory Integration

This scenario aims to provide Agrivi users with a streamlined, user-friendly way to access and understand regulatory frameworks for pesticide usage. With the assistance of Language Model (LLM) chatbots, farmers will be able to retrieve relevant regulations in their native language using natural language queries. This feature, coupled with an automated update system for global pesticide registrations, should enhance Agrivi's usability and foster better compliance with pesticide regulations.
Current State (before EFRA)
Currently, Agrivi allows farmers to manually input regulatory data, such as pre-harvest intervals (PHI) and maximum residue limits (MRL). There are also plans to centralize global pesticide registrations within the system. However, users must be reasonably tech-savvy to navigate this information, and language barriers may pose additional challenges.
Future State (after EFRA)
Post-EFRA, Agrivi will boast a fully integrated system that can automatically update farmers on relevant pesticide regulations. Moreover, LLM chatbots will enable farmers to conveniently access this information in their own language and using natural language inquiries. This will greatly improve the platform's accessibility and user-friendliness, leading to more responsible and effective pesticide application.
Envisioned Challenges & Risks

Challenges could include developing sophisticated NLP capabilities for the chatbot to understand and process natural language queries from various languages accurately. Further, the integration of the chatbot with different regulatory databases for real-time updates, and safeguarding privacy and security within this framework, might be complex tasks.	
Relevant Data	
Already existing in Agrivi DB	**New data types and potential sources**
Agrivi can provide data on: • the current state of regulations as entered by users • interactions between farmers and regulations	In addition to accessing external databases for up-to-date pesticide regulations, the chatbot system would require training data for language understanding and processing. Potential sources might include linguistic databases, regulatory text corpuses, user query logs, etc. (sources?)
Targeted Users for Piloting Activities	
Farmers who frequently use pesticides, operate under strict regulatory frameworks, and possibly face tech or language barriers would be ideal pilot users. A geographically diverse user base would also be beneficial to test the system's efficiency across various regulatory environments.	
Support by Agrivi for Scenario Implementation	
Agrivi can provide technical support for integrating the chatbot and ensuring it operates smoothly with the system's other components. They could also offer resources to train users to interact effectively with the chatbot and navigate the integrated regulatory features. Finally, Agrivi will set up appropriate piloting activities to test the new features with current Agrivi users.	
KPIs	
KPI AG2.1	Reduction in Time Spent on Regulatory Research > 33% over previous approaches used by farmers
KPI AG2.2	User adoption rate > 30% of current user base
KPI AG2.3	User satisfaction grade > 7/10 for the new developed features

7.2 Regulatory use-case (Leader: SGS)

7.2.1 Scenario RG.1: Automated Regulatory Analysis & Summarization Module

Current State (before EFRA)
Currently, SGS DIGICOMPLY provides comprehensive regulatory data, but the interpretation and summarization of these data require significant manual effort from the users.
Future State (after EFRA)
The integration of an Automated Regulatory Analysis & Summarization Module into the SGS DIGICOMPLY platform would drastically reduce manual effort by providing key summaries and extracts from regulatory texts. This could significantly enhance user experience and efficiency.
Envisioned Challenges & Risks

Key challenges might include ensuring the accuracy of automated analysis and summaries, accommodating the diversity and complexity of regulatory texts, especially in a multilingual context and ensuring the module's seamless integration with the existing platform.	
Relevant Data	
Already existing in SGS DB	**New data types and potential sources**
Regulatory texts and amendments, compliance guidelines, product and label regulations.	Specific ontologies and translations dictionaries specialized in the subject.
Targeted Users for Piloting Activities	
The primary users would be product compliance units within food companies and regulators who work with food safety regulations. Existing users in these categories can be enlisted from within the current users of SGS DIGICOMPLY software and participate in the piloting activities.	
Support by SGS for Scenario Implementation	
SGS can provide comprehensive training for pilot users on the new module and collect feedback for continuous improvement. SGS can also collaborate with AI and NLP experts to ensure the model's accuracy and effectiveness. Finally, they can set up and run the relevant piloting activities with their users.	
Outcome at M15 – March 2024	
A summary report for the piloting activities with existing SGS DIGICOMPLY users. The report will include information about the accuracy, the usefulness for the users, the number of documents processed, the most relevant summarisation task performed by the users among the different options tested.	
KPIs	
KPI RG1.1	Accuracy of the automated analysis and summaries > 65%
KPI RG1.2	User satisfaction ratings with the Automated Regulatory Analysis & Summarization Module > 7/10
KPI RG1.3	Reduction in the time spent by users on analysing and summarizing regulatory texts > 40%

7.2.2 Scenario RG.2: Predictions of food safety regulatory changes

The goal of this scenario is to leverage data mining and artificial intelligence to predict changes in food safety regulations. By tracking patterns between early warning risk monitoring, informal media discussions, and expert opinion pieces, we aim to create a model that can predict forthcoming food safety regulatory adjustments.
Current State (before EFRA)
At present, changes in regulations often come as a surprise, causing significant adjustments for food companies. From the regulators' perspective, it's challenging to get a comprehensive view of the regulations that need to change and the reasons behind these changes.
Future State (after EFRA)
After the implementation of EFRA, food companies will be better prepared for potential food safety regulation changes, thanks to predictive analytics. In addition, regulatory authorities will have access to decision dashboards that help them identify intervention areas, enabling more informed decision-making processes.
Envisioned Challenges & Risks

Predicting regulatory changes involves complex intelligence data. The most significant challenge lies in establishing the real-world accuracy of the prediction models. There is also the risk of relying too heavily on predictions, which might not always be accurate due to the inherent uncertainty and variability in the influencing factors.	
Relevant Data	
Already existing in SGS DB	**New data types and potential sources**
• Food Safety Regulatory history • Food safety incident records including early warning risk monitoring • previous changes in regulations related to specific incidents or discussions • Scientific opinions from food safety authorities	There is no need to extend the sources and data types currently available.
Targeted Users for Piloting Activities	
The primary users would be product compliance units within food companies and regulators who work with food safety regulations. Existing users in these categories can be enlisted from within the current users of DIGICOMPLY software and participate in the piloting activities.	
Support by SGS for Scenario Implementation	
SGS can provide a framework for data collection and integration of these new data types. They can also offer assistance in training and validating the AI predictive models, drawing on their extensive regulatory knowledge and data analysis capabilities. Finally, they can set up and run the relevant piloting activities with their users.	
Outcome at M15 – March 2024	
The outcome will be a) dataset of all the triggers and related information from a data set representing the contents collected within the 2023 timeframe within selected jurisdictions, most likely EU + 1 non EU country. On each data point (trigger) we will provide accuracy score. B) an expert review report to outline the potential false negative missed by the prediction model.	
KPIs	
KPI RG2.1	The accuracy of the AI prediction models > 60%
KPI RG2.2	Usability of decision dashboards, as determined by user satisfaction ratings > 7/10
KPI RG2.3	Reduction in time and resources spent by food companies > 30%

7.3 Poultry use-case (Leader: MOY Park)

7.3.1 Scenario PL.1: Advanced Data-driven Good Manufacturing Practices for Prevention of *Salmonella* Cross-contamination

GMP encompasses a set of guidelines and principles that outline the best practices for the manufacturing process. These practices cover various areas, including personnel training, facility maintenance, sanitation, equipment calibration, and documentation. By adhering to GMP, food manufacturers can maintain consistent quality, prevent contamination, and comply with regulatory requirements. For Moy Park, maintaining high hygiene standards and food safety includes conducting regular audits across multiple facilities.

Based on the risk assessments, lab test results, WGS analysis and by employing appropriate AI techniques, audits can be prioritized to focus on areas with the highest risk levels. This approach optimizes resource allocation by directing audit efforts towards areas where there is a higher likelihood of cross-contamination or identifying potential sources of contamination.

Current State (before EFRA)

Two important problems for the cross-contamination of *Salmonella* include biofilm formation and Viable but not culturable cells (VBNC). Before EFRA, a data-driven and AI-enabled risk assessment of facility-level *Salmonella* cross-contamination due to these two factors has not been established.

Biofilm formation

GMP practices in the poultry industry include the prevention of biofilm formation through effective cleaning and sanitation. Biofilm formation occurs when bacteria, such as *Salmonella, Campylobacter,* and *Listeria*, attach to surfaces and create a protective matrix of extracellular polymeric substances (EPS). The poultry processing environment provides ample opportunities for biofilm formation due to the presence of moisture, nutrients, and organic matter. Additionally, biofilms can form on live birds, leading to the introduction of pathogens into the processing facility. Biofilms pose significant challenges to food safety in the poultry supply chain. They act as reservoirs for pathogens, allowing them to survive harsh conditions and resist sanitization measures. Biofilms can shelter bacteria from cleaning agents, making it difficult to eradicate them entirely. If biofilms are not adequately controlled, they can lead to cross-contamination, resulting in the spread of pathogens to poultry products during processing, packaging, and distribution.

VBNC

Viable but not culturable (VBNC) cells refer to microorganisms that are alive but cannot be readily cultured using standard laboratory techniques. While they may be metabolically active, they are in a dormant or non-replicative state. VBNC cells can arise due to various factors, including exposure to stressors such as disinfectants, temperature variations, or nutrient limitations. These cells can pose challenges as they may still be capable of causing contamination or reactivating under favorable conditions. By considering the potential for *Salmonella* to enter the VBNC state, food safety protocols can be enhanced to minimize the risk of contamination and ensure the safety of food products. For example, maintaining a clean and hygienic environment reduces the potential for *Salmonella* to enter the VBNC state and persist in the facility.

Future State (after EFRA)

Post-EFRA, Moy Park will utilize an AI model to optimize audit prioritization based on the conditions promoting biofilm formation. Whole Genome Sequencing (WGS) analysis can aid in identifying cross-contamination pathways, and its insights can help enhance facility maintenance practices. AI algorithms will analyze WGS data to identify genetic markers of biofilm-forming strains related to antibiotic resistance, stress tolerance, or adhesion capabilities. Furthermore, integrating environmental data such as temperature, humidity, and nutrient levels with WGS data can identify the factors that promote or inhibit biofilm formation.

Envisioned Challenges & Risks

Potential issues include the availability of sufficient WGS data within the project's timeline, which may require leveraging external resources such as the BLAST genome library. Another challenge could be that multiple factors influencing biofilm formation may not be accounted for entirely in a data-driven approach.

Relevant Data

Already existing in MOY Park DB	**New data types and potential sources**

WGS data can reveal the most likely root cause of cross-contamination, often biofilm formation, by eliminating other sources like worker hygiene. AI algorithms can leverage WGS data to identify genetic markers in biofilm-forming strains linked to antibiotic resistance, stress tolerance, and adhesion capabilities. Simultaneously, integrating this data with environmental monitoring data like temperature, humidity, and nutrient levels helps pinpoint factors that encourage or suppress biofilm formation. This combined knowledge can be instrumental in creating preventive measures and optimizing sanitation protocols. Moy Park's lab results for *Salmonella* over a 5-year period provide a wealth of data, detailing sampling time, location, pathogen tested, analysis method, and results. Data on Moy Park's operational flow and the interconnectedness of processes and facilities across the supply chain are also accessible, providing valuable context. Information about the equipment in use, its maintenance schedule, sanitation practices, and historical audit reports further enhance the data landscape. Each audit, timestamped and tied to a specific facility or supply chain point, is a valuable piece of the puzzle. Lastly, investigation reports following detected *Salmonella* presence offer critical insights.	Numerous factors that can significantly impact the formation of *Salmonella* biofilms have been documented in scientific literature, which includes but is not limited to: Surface Type: For instance, stainless steel, commonly used in food processing, can facilitate biofilm formation. *Salmonella* Strain: Different strains of *Salmonella* have varying capacities to form biofilms, largely influenced by their unique genetic composition. Serotype: Some serotypes have been found to produce robust biofilms, contributing to persistent contamination. Environmental Conditions: Humidity, temperature, pH, and salt concentration can dramatically influence biofilm formation. Nutrient Availability: High nutrient conditions usually favor biofilm formation. Presence of Other Bacteria: *Salmonella* biofilm formation can be enhanced or inhibited by the presence of other bacterial species through interspecies interactions. Antibiotic Resistance Genes: The presence of antibiotic resistance genes can make the biofilms more resilient to antimicrobial treatments, making them harder to eradicate and thus a significant factor in biofilm formation. Certain studies provide quantifiable data linking these factors to biofilm formation, which can be beneficial in building more accurate models for prediction and prevention of biofilm formation. Moreover, the BLAST genome library, a highly curated database containing genomic sequences of various organisms, can be employed to correlate specific genomic sequences with biofilm formation capabilities, thereby aiding in a deeper understanding of biofilm genomics and informing better prevention and control strategies.
Targeted Users for Piloting Activities	
During the first phase, AI models will be primarily trained on a mix of public datasets and proprietary data from Moy Park (MOY). This phase aims to develop and fine-tune the AI models for efficacy and reliability, using MOY Park as the primary environment for model validation and testing. The targeted users at this stage will be food safety experts,	

laboratory technicians, and process managers within MOY Park who are closely involved with the daily operations and have a comprehensive understanding of the existing processes and challenges.

The second phase expands the scope to encompass a larger dataset and a broader range of user experiences. This could involve bringing more MOY Park facilities into the project or collaborating with other poultry companies to implement a Federated Learning approach. Here, the AI model is refined further by learning from multiple datasets owned by different entities while ensuring data privacy. This phase's targeted users would extend to professionals from these participating organizations, fostering a more extensive and collaborative effort in enhancing food safety.

Support by MOY Park for Scenario Implementation

To ensure the successful implementation of the AI models, MOY Park will play a crucial role throughout the project. First, they will contribute necessary datasets for training the AI models, which include laboratory results, Whole Genome Sequencing data, facility and equipment information, audit reports, and investigation reports.

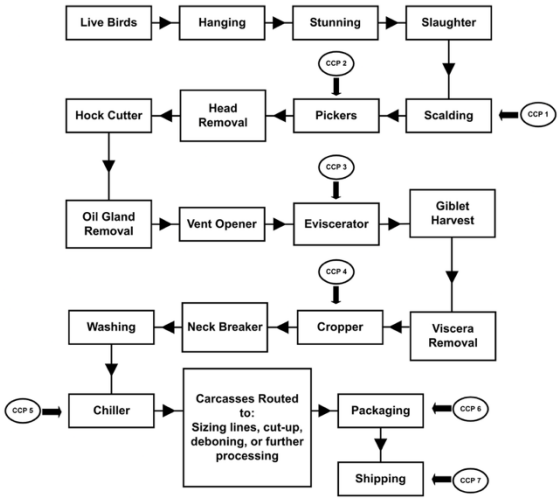
Next, MOY Park's team of food safety experts and researchers will actively participate in the development and testing stages. They will provide invaluable insights to fine-tune the models based on real-world applications and constraints, identify potential gaps or areas for improvement, and validate the model's predictions against their expert knowledge and experience.

Lastly, MOY Park will coordinate piloting activities within their facilities to evaluate the AI model in a real operational environment. These activities will test the model's usability, applicability, and effectiveness in prioritizing audits and managing *Salmonella* cross-contamination risks. MOY Park will also provide feedback and participate in subsequent iterations to continuously refine the model's performance and adaptability.

KPIs	
KPI PL1.1	The accuracy of the AI prediction models > 60%
KPI PL1.2	Usability of decision dashboards, as determined by user satisfaction ratings > 7/10

7.3.2 Scenario PL.2: Uncovering Causal Relationships of *Salmonella* Risk within the Supply Chain

Moy Park manages an extensive production network encompassing feed mills, hatcheries, broiler farms, and fresh poultry processing facilities. Its production approach is comprehensive, overseeing not just broilers raised for consumption but also the management of parent and grandparent flocks. Even with rigorous preventive strategies like Controlled Natural Decomposition (CND), heat treatments, and stringent hygiene controls, sporadic instances of S*almonella* contamination persist. These occurrences, often appearing without overt causes, underline the importance of unravelling potential precursors. These could be conditions within feed mills, hatcheries, farms, or related to biofilm formation and hygiene practices. Moy Park aims to identify and comprehend these causal pathways and transform this understanding into a predictive tool to further fortify their defenses against *Salmonella* outbreaks.



Current State (before EFRA)

Currently, a clear understanding of the intricate causal relationships that contribute to *Salmonella* risk within the complex supply chain remains elusive. Even though reactive measures and routine preventive strategies are in place, a predictive tool to anticipate and mitigate *Salmonella* outbreaks is lacking. Without the ability to predict and strategically act on potential *Salmonella* hotspots, the fight against this pathogen remains reactive rather than proactive, making it challenging to eradicate completely.

Future State (after EFRA)

With the EFRA project's successful implementation, a novel algorithm capable of deciphering causal links for *Salmonella* presence across the intricate supply chain will be launched. By revealing the causal factors contributing to *Salmonella* contamination, this tool allows operators to anticipate and address issues before they escalate, shifting the paradigm from reactive to proactive *Salmonella* control. Consequently, Moy Park can enhance their preventive measures, leading to safer poultry products and an improved public health outlook.

Envisioned Challenges & Risks

Establishing causal links between diverse factors in a complex system such as a supply chain poses substantial challenges, and existing AI methodologies may have limitations in adequately addressing these complexities. If delineating definitive causal connections proves too intricate, the consortium may need to pivot towards identifying likely correlations.

Relevant Data

Already existing in MOY Park DB	**New data types and potential sources**
The existing data at MOY Park encompasses lab test results for *Salmonella* over the past five years, maintained through its proprietary laboratories. Each test result includes crucial details such as sampling time and location, pathogen under examination, analysis method, and test outcome (e.g., pathogen concentration in the tested sample). In certain instances, lab test results may be cross-verified with external reference laboratories for additional accuracy and reliability. Additionally, MOY Park holds a detailed schematic of	As of now, no additional data have been deemed necessary, but this might change as the project unfolds and will be reflected in subsequent iterations of the deliverable.

its operational flow and interconnected processes leading to the final product, which serves as an invaluable source of information. Investigation reports following *Salmonella* detection in lab test results further enrich this dataset. Other significant data include Whole Genome Sequencing (WGS) data, feed types used at different time junctures, and readings from environmental sensors.	
Targeted Users for Piloting Activities	
AI models will be developed and honed using a combination of public datasets and MOY Park's proprietary data. MOY Park's operations will provide a realistic testing and validation environment for these AI models. The targeted users during this phase are MOY Park's food safety experts, laboratory technicians, and process managers, who possess an intricate understanding of the daily operations and prevailing challenges.	
Support by MOY Park for Scenario Implementation	
MOY Park will lend its domain and industry knowledge and understanding of the complex supply chain to ensure accurate data interpretation and practicality of algorithm results. MOY Park will play a pivotal role in the successful implementation of the AI models, contributing necessary datasets - laboratory results, WGS data, facility and equipment information, audit reports, and investigation reports - for training the AI models. Moreover, MOY Park's team of food safety experts and researchers will contribute actively during the development and testing phases, offering invaluable insights to finetune the models based on real-world constraints, identify potential improvement areas, and validate model predictions against their expert knowledge and experience. Finally, MOY Park will facilitate piloting activities within its facilities to test the AI model's practicality, usability, and effectiveness in audit prioritization and S*almonella* cross-contamination risk management. Feedback and participation from MOY Park will be instrumental in refining the model's performance and adaptability through subsequent iterations.	
KPIs	
KPI PL2.1	The accuracy of the AI prediction models > 60%
KPI PL2.2	Usability of decision dashboards, as determined by user satisfaction ratings > 7/10

7.3.3 Scenario PL.3: Real-Time Alerting System for Hatchery Health Monitoring

Moy Park operates numerous hatcheries as part of its extensive production process. Maintaining optimal environmental conditions within these hatcheries is of utmost importance for ensuring bird health, which directly influences the safety and quality of their products. Environmental variables such as temperature, humidity, or air quality, if unbalanced, can adversely affect bird health. Detecting and promptly responding to these fluctuations, especially when overseeing multiple hatcheries, presents a significant challenge. Moy Park aims to establish an advanced alert system that can detect environmental irregularities and predict potential issues before they arise, thereby ensuring ideal hatchery conditions and allowing swift intervention to safeguard bird health.
Current State (before EFRA)

At present, Moy Park does not utilize a specialized algorithm for the detection of anomalies in sensor readings. Nonetheless, they do have a rudimentary alerting system in place that triggers whenever sensor values exceed predefined thresholds.
Future State (after EFRA)
Post EFRA project, Moy Park plans to implement a dedicated algorithm for anomaly detection in sensor readings. This cutting-edge algorithm will be capable of operating within their facilities, at-the-edge, providing real-time insights and alerts, thereby enhancing the promptness and effectiveness of response to potential threats.
Envisioned Challenges & Risks
One of the significant challenges lies in the need for integration at the hardware level. The successful pairing of sensors with edge devices could present difficulties and may require high Technology Readiness Level (TRL) to enable on-premises use through user-friendly deployment and operation. Additionally, maintaining the reliability of the system in a dynamic hatchery environment poses another potential risk.
Relevant Data

Already existing in MOY DB	**New data types and potential sources**
High-frequency readings from environmental sensors, including those monitoring temperature, humidity, and air quality, form a crucial part of the data needed for this project. These readings provide real-time insights into the hatchery conditions and serve as a baseline for anomaly detection.	As of now, the existing data in the MOY database sufficiently cater to the requirements of the proposed AI model. Any potential future needs for additional data types will be evaluated during the development and testing phases.

Targeted Users for Piloting Activities
The primary users involved in this phase will be food safety and Quality Assurance (QA) experts, along with process managers within MOY. Their close involvement with the daily operations and a thorough understanding of the existing processes and challenges make them crucial players in this project. Their feedback and insights will guide the model refinement process, ensuring that it aligns with the practical realities of hatchery operation.
Support by MOY for Scenario Implementation
MOY will provide substantial support to ensure the successful implementation of the AI model. They will grant access to sensor data and assign a dedicated team to set up the edge device. This team will ensure that the model can run effectively within the facilities, offering real-time insights and alerts. MOY's team of food safety experts and researchers will be actively engaged throughout the development and testing stages. They will offer valuable input to refine the model, considering real-world applications and constraints. Additionally, they will identify potential gaps or areas for improvement and validate the model's predictions against their empirical knowledge and experience. Lastly, MOY will orchestrate piloting activities within their facilities to evaluate the AI model in a real-world operational environment. This hands-on testing will assess the model's usability, relevance, and efficacy in monitoring hatchery conditions and detecting potential health risks. MOY will also provide critical feedback and engage in subsequent iterations, allowing for the continuous refinement and enhancement of the model's performance and adaptability.
KPIs

KPI PL3.1	The accuracy of the AI prediction models > 60%
KPI PL3.2	Usability of decision dashboards, as determined by user satisfaction ratings > 7/10

8 Expected Outcomes Roadmap

As we approach the concluding section of this document, we present a comprehensive outline of our expected outcomes by M15 and define a clear roadmap including specific delivery dates. This part serves not only as a summary of our commitments but also as a blueprint for our journey towards achieving the goals we have set.

We will detail each significant milestone in our journey, elaborating on the tasks that need to be accomplished, the timeline for completion, and the key deliverables that we anticipate. This detailed and transparent roadmap is designed to ensure that we maintain our focus, meet our deadlines, and deliver the high-quality results we aspire to.

Crucially, the responsibilities of each partner within the consortium will also be outlined in relation to each outcome. It is through the collective effort and synergies of all our partners that we can turn our vision into reality. By clearly defining roles and responsibilities, we can facilitate smooth collaboration, ensure accountability, and harness the unique strengths and capabilities of each partner.

This forward-looking section encapsulates our collective commitment to the project and serves as a guide for our continued efforts. Our ultimate goal is to execute each task efficiently and effectively, steering our project towards successful and timely completion. This section will be appropriately updated in D1.2 iterations to better reflect the evolving outcomes of the project.

8.1 Outcomes lead by Agroknow

Outcome AGK1: AI-assisted Crawler	
The outcome involves the development of an advanced AI-assisted web crawler that leverages a trained Natural Language Processing (NLP) Classifier. This crawler is designed to start from a select group of pertinent seed web pages, following web links to discover and classify web pages containing information on food safety incidents. The trained NLP Classifier allows the crawler to discern between three types of content: (a) public food safety incidents, (b) food-related content not concerning incidents, and (c) content unrelated to food. This categorisation allows the crawler to only move on promising parts of the web.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
SU	In future iterations, SU will explore the feasibility of enhancing the core NLP Classifier to support multilingual models, allowing for broader data collection and more inclusive information gathering.
Related Tasks	
T2.1 EFRA Data Hub population & heterogeneous data source registry	
Related Deliverables & Content	
D2.1 (due in December 2023)	The model's design along with the approach for training and utilization (English language only) will be prepared.
D5.3 (due in December 2023)	The initial experimental results and a report detailing accuracy and performance will be provided.
Current State-of-the-art (before EFRA)	
At present, Agroknow and similar food intelligence companies usually employ experts to manually scan the internet for valuable sources of food safety incidents. The search is concentrated on discovering public food safety authorities in the targeted country for each specific case.	

Future State-of-the-art (after EFRA)	
Post-EFRA, an AI-assisted web crawler that automatically detects and classifies web pages containing information about food safety incidents will be created. This advancement will enhance the breadth and depth of data collection, improve efficiency, and potentially boost the quality and speed of responses to public food safety incidents.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Collect the relevant datasets	June 2023
Design the NLP Classifier	July 2023
Train the NLP Classifier & onboard in a crawler	September 2023
Conduct experiments & measure performance	December 2023
Envisioned Challenges & Risks	
Some potential challenges could include gathering sufficient and relevant training data, ensuring the accuracy of the NLP Classifier in differentiating between content types, and addressing privacy or security issues that might arise during the crawling process. There's also a risk of false positives or negatives in the classification of food safety incidents. Finally, it is possible that verifying the accuracy of the produced results will require the active involvement of experts.	
Required Data & Potential Sources	
This outcome requires three types of datasets: • Public food safety incidents and the relevant webpage payloads • General food-related articles and the relevant webpage payloads • Articles not related to food at all and the relevant webpage payloads	
Required Domain Expertise Support	
The required support will be provided by Agroknow expert data curators already undergoing the current, manual process.	
KPIs	
KPI AGK1.1	Decrease in time spent searching and classifying information about food safety incidents > 30% compared to current Agroknow processes
KPI AGK1.2	Accuracy of Classification > 70% of correctly classified web pages by the NLP Classifier
KPI AGK1.3	Relevant scientific publication in a peer-review venue = 1

Outcome AGK2: Comprehensive Report Identifying *Salmonella* Contamination Sources	
This outcome involves conducting an extensive literature review to pinpoint potential sources of *Salmonella* contamination, focusing primarily on the processes within a poultry company. The objective is to establish a more comprehensive understanding of possible contamination pathways, which could inform more robust safety measures and response strategies.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
MOY	Moy Park will provide domain expertise, assisting in interpreting the findings and validating the potential sources of *Salmonella* contamination identified in the study.
Related Tasks	
T1.1 Scientific requirements on short- and long-term food risk prediction	
Related Deliverables & Content	

D1.1 (due in September 2023)	Initial report presenting the findings of the study, based on a review of approximately 20 significant publications in the field of food safety, specifically *Salmonella* contamination.

Current State-of-the-art (before EFRA)

Currently, information on *Salmonella* contamination sources is dispersed across various publications, requiring a systematic literature review to gather and analyze the data comprehensively.

Future State-of-the-art (after EFRA)

After EFRA, there will be a deeper understanding of potential *Salmonella* contamination pathways, informed by a consolidated report summarizing the findings from numerous relevant publications. This enhanced knowledge could lead to improved risk management strategies and preventive measures within the poultry industry.

Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Selection of appropriate publication aggregators and query formulation	June 2023
Review of top 10 most relevant papers	July 2023
Review of the subsequent 10 most pertinent publications	August 2023
Compilation of consolidated findings and identification of contamination pathways	September 2023

Envisioned Challenges & Risks

Given the extensive nature of the literature, sorting and analysing the data for relevancy may pose a significant challenge. Additionally, the results presented in different publications may be unclear or may not be directly applicable across different cases, requiring expert interpretation.

Required Data & Potential Sources

This outcome requires the use of appropriate publication aggregators and tailored search queries to extract relevant data. The specific sources and queries used will be documented in D1.1.

Required Domain Expertise Support

Expert support will be provided by Agroknow and Moy Park food safety experts, who will lend their expertise to analyse the literature, interpret findings, and validate potential contamination pathways.

KPIs

KPI AGK2.1	Number of relevant publications studied > 40
KPI AGK2.2	Number of contamination paths identified > 10
KPI AGK2.3	Relevant scientific publication in a peer-reviewed venue = 1

Outcome AGK3: Comprehensive Report and White Paper on EFRA Summit

This outcome involves organizing, executing, and documenting the proceedings of the EFRA summit. The summit's primary objective is to foster a focused discussion on challenges and opportunities related to food intelligence sharing in light of novel technologies. A detailed report will be compiled summarizing the insights and ideas generated during the brainstorming sessions at the summit. This report will be designed as a white paper and disseminated to key stakeholders in the EFRA project and beyond.

Lead Partner

Agroknow

Contributing Partners & Roles

RAIN	RAIN will assist in designing the white paper and promoting it to relevant networks, leveraging their expertise in communication and network engagement.

Related Tasks

T 5.1 Focus Groups on AI Risk Predictions Challenges & Real-world Adoption
Task 6.2 Building & engaging a Public-Private Network on risk prediction

Related Deliverables & Content

D5.1 (due in August 2023)	A comprehensive report outlining the planning, execution, and outcomes of the EFRA summit and associated brainstorming sessions.
D6.1 (due in December 2023)	A white paper based on EFRA summit brainstorming sessions

Current State-of-the-art (before EFRA)

Prior to EFRA, no dedicated forum existed to facilitate focused discussions on food intelligence sharing using emerging technologies.

Future State-of-the-art (after EFRA)

With the successful execution of the EFRA summit, the aim is to establish this event as an annual gathering, bringing together experts from food industry, artificial intelligence, and legal sectors to collaboratively address key challenges and explore opportunities in food intelligence sharing.

Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Planning and promoting the EFRA summit	May 2023
Execution of the EFRA summit	June 2023
Analysis of the brainstorming results	August 2023
Compilation and design of the white paper	November 2023

Envisioned Challenges & Risks

Ensuring a critical mass of diverse participants for the summit may be a challenge. Diversity in terms of backgrounds and expertise could affect the quality of brainstorming sessions. Open-invite focus groups may pose a risk of misaligned discussions due to the broad spectrum of participants.

Required Data & Potential Sources

Insights from the brainstorming session, combined with relevant academic and industry resources, will be used to enrich the content of the white paper.

Required Domain Expertise Support

Agroknow will provide expert support in conducting the summit and analysing the results. RAIN will offer support in designing the white paper and promoting it to appropriate networks.

KPIs

KPI AGK3.1	Number of participants in summit > 40
KPI AGK3.2	Number of participants in brainstorming sessions > 10

Outcome AGK4: New Data Sources, Data Package, and APIs

Outcome AGK4 focuses on the enrichment of the EFRA Data Hub by incorporating new data sources. This enhancement process involves crawling or scraping information from four specific categories of data sources - public food safety authority sites, food safety news sites, regulatory document sources, and agricultural-related datasets. The relevant data records extracted from these sources will be meticulously curated and packaged, which will then be disseminated to the consortium. Moreover, the packaged data will be made accessible through Application Programming Interfaces (APIs) to the EFRA Data Hub, allowing for efficient, structured, and real-time interaction with this valuable information.

Lead Partner

Agroknow	
Contributing Partners & Roles	
SGS	Similar report and data package for new regulatory sources
Agrivi	Similar report and data package for new agricultural datasets
Related Tasks	
Task 2.1 EFRA Data Hub population & heterogeneous data source registry	
Related Deliverables & Content	
D2.1 (due in December 2023)	Report of new data sources added and data volume increase
Current State-of-the-art (before EFRA)	
Prior to the implementation of EFRA, partners Agroknow, SGS, and Agrivi were already active in incorporating new data sources and records into their respective platforms. However, these records were not available via an integrated API, nor were they consolidated for use in more complex, upstream Artificial Intelligence (AI) tasks. This previous state of affairs represented a fragmented approach to data integration, where each partner maintained their own individual dataset, thus leading to a potential under-utilization of the combined value that the data from all these different sources could offer.	
Future State-of-the-art (after EFRA)	
Upon the successful realization of the EFRA initiative, the newly obtained data records will be made available through an integrated API of the EFRA Data Hub. This unified access point will allow for streamlined interaction with the data and facilitate their consolidation for upstream AI tasks. With EFRA's more coordinated and cohesive approach, stakeholders will be able to access a broader and richer array of food safety and agricultural data. This could potentially enhance decision-making processes, support new AI-driven solutions, and foster innovation in the sector.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
List of new targeted data sources	October 2023
Report on new data sources and data package	December 2023
Envisioned Challenges & Risks	
While the initiative is ambitious and has the potential to revolutionize how food safety and agricultural data are accessed and used, several challenges and risks could impede its realization. One significant challenge could be the integration of the new data sources. Depending on the format, structure, and complexity of the original data sources, the process of crawling and scraping the relevant information might prove more difficult than initially estimated. This could lead to an unexpected increase in effort, time, and resources, possibly impacting the planned inclusion of other sources and delaying the whole process.	
Required Data & Potential Sources	
The data needed will come from various online resources, including public food safety authority sites, food safety news sites, regulatory document sources, and agricultural-related datasets. Each of these resources presents potential sources of vital data for the EFRA Data Hub.	
Required Domain Expertise Support	
The intricate process of data mining and curation requires deep domain expertise, which will be provided by the partners Agroknow, SGS, and Agrivi. Their collective knowledge and experience in the field will ensure that the relevant data sources are thoroughly explored, and the data records accurately curated.	
KPIs	
KPI AGK4.1	New data sources added in the EFRA Data Hub > 6
KPI AGK4.2	Data volume increase > 1M records

Outcome AGK5: Novel modules for data linking

Outcome AGK5 focuses on the development of two groundbreaking methodologies for intelligent data linking, intended to enhance the efficiency and utility of data in the field of food safety. The first of these approaches involves the execution of rapid two-hop traversals in large-scale graphs to compute similarity measures between indirectly connected nodes. This would facilitate a deeper understanding of potential correlations in data that aren't immediately apparent. The second approach entails the design of a clustering algorithm specifically tailored for food safety incidents that may share similar root causes, thus improving the capacity to identify patterns and prevent future incidents.

Lead Partner

Agroknow

Contributing Partners & Roles

-	At this point no other partners will need to contribute to this outcome

Related Tasks

Task 2.4 Intelligent Linking, Multilingual Semantic Enrichment and Data Fusion

Related Deliverables & Content

D2.2 (due in December 2023)	Description of the relevant algorithms and approach

Current State-of-the-art (before EFRA)

Prior to EFRA, there was no existing scalable approach for conducting two-hop linking in large graphs, limiting the ability to detect potential connections between indirectly related nodes. Furthermore, the process of deduplicating food safety incidents was tedious and challenging, involving manual curation by domain experts. This not only required significant time and effort but was also prone to human error.

Future State-of-the-art (after EFRA)

Post-implementation of the EFRA initiative, a scalable methodology for two-hop linking in large graphs will be established. This will provide a more efficient means to detect and analyse relationships between indirectly connected nodes within a vast data graph. Furthermore, the design and application of a clustering algorithm for food safety incidents sharing similar root causes will substantially aid domain experts, eliminating the need for tedious manual curation and improving the accuracy and efficiency of incident identification and management.

Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Gathering & curation of the relevant data	September 2023
Design of appropriate algorithms	October 2023
Experiments & validation of approach	November 2023

Envisioned Challenges & Risks

Due to the complexity of the task and potential lack of necessary contextual information, the accuracy of the clustering algorithm might be lower than expected. Another anticipated challenge is that the volume of real food safety data available for two-hop detection may be insufficient to showcase significant performance improvements, thereby hindering the demonstration of the effectiveness of the newly developed methods.

Required Data & Potential Sources

Outcome AGK5 will rely on lab test results of substantial volume and food safety incident data with appropriate metadata (including date of announcement, hazard, pathogen, country reporting issue, and sourcing country). These data will be provided by Agroknow through crawling public food safety authority sites. Such comprehensive data will be instrumental in training and optimizing the proposed algorithms for intelligent data linking.

Required Domain Expertise Support

Agroknow will provide the necessary domain expertise to evaluate the results of the clustering and intelligent linking algorithms. Their proficiency in data science, food safety, and AI will be crucial in ensuring the successful implementation and refinement of the proposed methodologies.

KPIs	
KPI AGK5.1	Accuracy of clustering algorithm as determined by domain experts > 60%
KPI AGK5.2	Scientific publication in a peer-reviewed venue = 1

Outcome AGK6: Novel module for extreme event forecasting	
Outcome AGK6 centers around the development of a pioneering approach to forecast extreme events in food safety incidents at least one month in advance. This innovative undertaking aims to leverage advanced algorithmic techniques to predict potentially catastrophic incidents, providing valuable lead time for preventive and mitigation measures.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
SU	SU can explore time series forecasting/extreme event forecasting explainability
Related Tasks	
Task 3.1 Process to design, train & test explainable food risk prediction models	
Related Deliverables & Content	
D3.1 (due in March 2024)	Description of the relevant algorithms and approach
Current State-of-the-art (before EFRA)	
In the current state of affairs, the food safety domain lacks a forecasting algorithm for extreme events. This gap in predictive capabilities presents a significant risk, as reactive measures to extreme events can often be insufficient and costly, emphasizing the need for more proactive strategies.	
Future State-of-the-art (after EFRA)	
With the successful realization of the EFRA initiative, a validated and accurate algorithm for extreme event forecasting will be developed. This innovative tool would represent a major advancement in the food safety domain, shifting the paradigm from reactive responses to a more proactive and preventive approach. This transition could significantly mitigate the adverse impacts of extreme events, potentially saving resources, efforts, and lives.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Gathering & curation of the relevant data	September 2023
Design of appropriate algorithms	October 2023
Experiments & validation of approach	November 2023
Envisioned Challenges & Risks	
Extreme event forecasting is inherently a complex problem, which may necessitate accounting for a multitude of variables that could be challenging to link with the relevant data sources. This complexity could pose significant obstacles in the design, training, and validation of the proposed forecasting algorithm.	
Required Data & Potential Sources	
Agroknow is anticipated to provide ingredient and hazard time series data, which will be used for training and evaluating the proposed forecasting approach. This kind of high-quality, targeted data is key to developing a reliable and effective predictive model.	
Required Domain Expertise Support	
Agroknow will also provide the essential domain expertise required to guide this ambitious project. With their deep knowledge and understanding of the food safety domain, coupled with their experience in data science and artificial intelligence, they will ensure the development of a robust and effective algorithm for extreme event forecasting.	

KPIs	
KPI AGK6.1	Accuracy of extreme event forecasting (one month prior) > 60%
KPI AGK6.2	Scientific publication in a peer-reviewed venue = 1

Outcome AGK7: Causal analysis of lab test results across a poultry supply chain	
Outcome AGK7 aims to leverage machine learning algorithms to elucidate causal relationships between the occurrence of *salmonella* at a specific stage in a poultry supply chain and indicators from preceding steps. By treating timestamps as primary data points, correlations and potential causal relationships can be identified. Moreover, the processes across the supply chain can be depicted as a graph network to further hint at potential causal connections. This breakthrough could enhance our understanding of *salmonella* outbreaks and their sources, leading to more effective control strategies in poultry production.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
MOY	Provide relevant data (lab test results, WGS data, flow diagram of operations). Previous investigation reports can act as golden truth.
Related Tasks	
Task 3.1 Process to design, train & test explainable food risk prediction models	
Related Deliverables & Content	
D3.1 (due in March 2024)	Description of the relevant algorithms and approach
Current State-of-the-art (before EFRA)	
At present, there is no existing algorithm capable of determining causal links for the presence of *salmonella* across a complex supply chain. This leaves a significant gap in understanding how this harmful pathogen might propagate through poultry production, thereby hampering effective preventive measures.	
Future State-of-the-art (after EFRA)	
Post-implementation of the EFRA project, a pioneering algorithm will be introduced that can establish causal links for the presence of *salmonella* across a complex supply chain. This innovation would revolutionize the field of food safety, providing an enhanced tool for understanding, predicting, and ultimately controlling *salmonella* outbreaks in poultry production.	

Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Gathering & curation of the relevant data and flow charts	October 2023
Design of appropriate algorithms	December 2023
Experiments & validation of approach	February 2024
Envisioned Challenges & Risks	
Establishing causal links is a complex task, and there might be limitations in the available AI approaches. If causal links prove too challenging to establish, the consortium may need to fall back on identifying plausible correlations instead, which, while still useful, might not provide the same depth of insight.	
Required Data & Potential Sources	
To effectively build and train the proposed algorithm, various sources of data will be needed. These include historical lab test results for the presence of *salmonella*, investigation reports of previous incidents, operational flow diagrams, and Whole Genome Sequencing (WGS) data. Such comprehensive data sets would offer a holistic view of the supply chain and the various factors at play, facilitating more accurate and robust causal analysis.	

Required Domain Expertise Support	
Moy Park will provide necessary domain expertise to interpret the data and algorithm results. Their knowledge of the poultry industry and understanding of the complexities of the supply chain will be crucial in ensuring the validity and practicality of the results.	
KPIs	
KPI AGK7.1	Cross-validation of causal links with in-the-field investigation proves satisfactory (satisfaction as measured by domain experts > 7/10)
KPI AGK7.2	Scientific publication in a peer-reviewed venue = 1

Outcome AGK8: Mock-ups and API listing for front-facing EFRA components	
Outcome AGK8 is primarily focused on the creation of high-fidelity mock-ups and the establishment of relevant specifications for the EFRA Data & Analytics Marketplace. Additionally, it involves compiling a comprehensive listing of all APIs presently utilized by Agroknow, SGS, and Agrivi, with a view towards future integration within the EFRA Platform. This initiative is key in laying the groundwork for a seamless, unified data analytics platform within the food safety domain.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
SGS	Provide listing of APIs. Co-design the mock-ups.
Agrivi	Provide listing of APIs.
Related Tasks	
Task 4.4 API Gateway and EFRA Data & Analytics Marketplace	
Related Deliverables & Content	
D4.3 (due in December 2023)	Specifications and mock-ups for Marketplace and listing of APIs
Current State-of-the-art (before EFRA)	
In the current scenario, no dedicated marketplace exists for food safety analytics and AI models. This gap implies a lack of a unified, user-friendly platform where stakeholders can access, analyse, and leverage data for more informed decision-making in the field of food safety.	
Future State-of-the-art (after EFRA)	
Upon successful completion of the EFRA project, a dedicated marketplace for food safety analytics and AI models will be introduced. This novel platform will streamline access to vital analytics and AI models, fostering more efficient, data-driven practices in food safety management. This could lead to more accurate risk assessment, more effective preventive measures, and ultimately safer food products.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Identification of the specifications	October 2023
Creation of the high-fidelity mock-ups	November 2023
Envisioned Challenges & Risks	
The project's lead and contributing partners do not foresee any significant risks in the execution of this task. The main challenge might be the coordination among partners for the successful integration of their APIs and the design of the marketplace that caters to the needs of the diverse user base.	
Required Data & Potential Sources	
N/A: As the task revolves around design and integration, it does not specifically require data sources.	
Required Domain Expertise Support	

Agroknow and SGS, with their domain expertise in data management, analytics, and food safety, will play pivotal roles in determining the specifications of the Marketplace. Their expert guidance will be instrumental in ensuring that the marketplace meets user needs and industry standards, thereby contributing to its ultimate success.

KPIs	
KPI AGK8.1	Number of novel features identified for the Marketplace > 5

Outcome AGK9: Design of Decision Support Scenarios & Pilots	
Outcome AGK9 focuses on designing decision support scenarios for each of the project's three use-cases. This process involves the creation of concrete, realistic scenarios where the tools and methods developed by the EFRA project can be applied and evaluated. Additionally, specific piloting activities and Key Performance Indicators (KPIs) will be outlined to test the EFRA modules' effectiveness within the defined use-case scenarios.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
SGS	Design of relevant piloting activities for use-case #3
Agrivi	Design of relevant piloting activities for use-case #2
MOY	Design of relevant piloting activities for use-case #1
Related Tasks	
Task 5.2 Experimental Methodology, Use-case Plan, and Recommendations	
Related Deliverables & Content	
D1.1 (due in September 2023)	Initial listing of scenarios
D5.3 (due in December 2023)	Elaboration of scenarios and design of specific piloting activities
Current State-of-the-art (before EFRA)	
At present, the partners involved in the use-case scenarios do not have specific solutions in place for the challenges identified within the context of the three use-cases: pathogen prevention in poultry, pesticide use in agriculture, and anticipation of regulatory changes. This presents an opportunity for the EFRA project to introduce innovative, data-driven solutions to address these critical issues.	
Future State-of-the-art (after EFRA)	
Upon the completion of the EFRA project, partners will have successfully piloted specific solutions to address the identified challenges within the three use-cases. This includes pathogen prevention in poultry, pesticide use in agriculture, and anticipation of regulatory changes. The outcomes of the project have the potential to significantly enhance decision-making and risk management processes within these contexts, leading to safer, more sustainable food production practices.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Initial listing of scenarios for each use-case	July 2023
Elaboration on scenarios	September 2023
Design of piloting activities	November 2023
Envisioned Challenges & Risks	
One of the main challenges foreseen in this project is ensuring that the timing and design of the pilots align with the results of the relevant Research and Development (R&D) activities. This synchronization is crucial to effectively test and validate the solutions developed during the course of the project.	
Required Data & Potential Sources	

As this task primarily involves planning and design, it doesn't specifically require data sources. The design of scenarios and piloting activities will mainly be based on the partners' domain expertise and understanding of the project's objectives.
Required Domain Expertise Support
The partners involved in each use-case scenario – SGS, Agrivi, and MOY – will provide the necessary domain expertise. Their industry insights and experience will be essential in designing appropriate piloting activities that accurately reflect real-world conditions and challenges.

KPIs	
KPI AGK9.1	Identified use-case scenarios > 5
KPI AGK9.2	Pilots designed with at least 2 iterations = 3 (one per use-case partner)

Outcome AGK10: Report on meetings with important networks & ADRA membership	
Outcome AGK10 is centered on creating and strengthening relationships within the food industry, specifically through targeted meetings with key food networks such as the Global Food Security Institute (GFSI) and the Food Integrity Intelligence Network (fiin). Additionally, it involves the consortium members joining the Association for Data, Robotics and AI (ADRA) and actively participating in its meetings. These activities will provide opportunities for dialogue, collaboration, and knowledge exchange, enhancing the consortium's understanding of current industry needs and standards.	
Lead Partner	
Agroknow	
Contributing Partners & Roles	
ALL	Apply for ADRA membership
Related Tasks	
Task 6.3 Involvement in industry groups & networks	
Related Deliverables & Content	
D6.1 (due in December 2023)	Report of meetings with GFSI and fiin. Report on ADRA memberships established.
Current State-of-the-art (before EFRA)	
As it stands, no members of the consortium are part of ADRA. This suggests a potential lack of direct engagement and influence in the activities and discussions within this critical EU network.	
Future State-of-the-art (after EFRA)	
Post EFRA, it is envisioned that at least three consortium members will actively participate in ADRA. Their involvement will enhance the consortium's visibility and representation within the industry, promoting stronger ties and communication with key stakeholders.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Meetings with GFSI and fiin	October 2023
Consortium members apply for ADRA membership	November 2023
Envisioned Challenges & Risks	
Currently, there are no significant challenges or risks foreseen in this task. However, it is important to maintain open and ongoing communication with these networks to ensure successful engagement and collaboration.	
Required Data & Potential Sources	
No specific data sources are required for this task as it primarily involves network engagement and membership application processes.	
Required Domain Expertise Support	

There is no need for specialized domain expertise support in this task. The consortium members will use their industry knowledge and experience to effectively engage with the targeted networks and ADRA.

KPIs	
KPI AGK10.1	ADRA members in the consortium >= 3
KPI AGK10.2	Meetings with relevant networks >= 2

8.2 Outcomes lead by SU

Outcome SU1: NLP Baselines for Food Risk Prediction	
This outcome involves creating datasets and baselines for future EFRA deliverables in NLP. The baselines will be simple models for each of the tasks to allow for comparability of models developed in the project. The currently envisioned tasks are: (i) classification of food incident data, (ii) extraction of relevant entities from food news.	
Lead Partner	
SU	
Contributing Partners & Roles	
Agroknow	Provide data and help with data curation
Related Tasks	
Task 2.2 Novel explainable mining & analysis methods and tools Task 3.1 Process to design, train, and test explainable food risk prediction models Task 5.3 Experiments on performance, speed & accuracy	
Related Deliverables & Content	
D2.1 (2023-12)	Baselines for extraction of relevant entities from food news
D5.2 (2023-12)	Evaluate baselines to create benchmarks for future model development
Current State-of-the-art (before EFRA)	
There are currently no benchmarks for text-based data regarding food risk prediction from public sources.	
Future State-of-the-art (after EFRA)	
We will provide datasets and benchmarks for comparison. This will allow EFRA partners and other researchers to develop and compare novel methods for text-based food risk prediction.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Alignment call and task finalization SU and Agroknow	July 2023
Data from Agroknow	July 2023
Description of the data labelling process	July2023
SU and Agroknow finish data exploration and dataset preparation for the food incident data (i)	October 2023
Final Paper on (i)	October 2023
Agroknow annotates food news with entities	M15+
SU provides baseline models for food news (ii)	M15+
Envisioned Challenges & Risks	
This outcome and its deadlines depend on the availability of the data and the corresponding labels in time. Furthermore, a timely response on the questionnaire by participating partners is important. Low quality data may make several iterations of labelling and exploration from both SU and Agroknow necessary. The data should reflect well-defined use case applications.	

Required Data & Potential Sources	
Food incident data including food recall reports from food authorities and food news data will be provided by Agroknow.	
Required Domain Expertise Support	
We need domain expertise for the labelling of the datasets.	
KPIs	
KPI SU1.1	Benchmark performance > naïve baselines (e.g. majority class) for each task

Outcome SU2: Shared Task on Food Risk Prediction	
We will create a shared NLP task on food risk prediction. This task is concerned with the classification of food hazards and food products from texts (website titles and content). Participants are asked to develop NLP models that predict product and hazard categories and produce explanations by predicting/extracting a more fine-grained label for each of those. As an optional task, participants can submit auto generated explanations for their predictions, which will be assessed by Agroknow's domain experts. The challenge will hopefully provide a multitude of approaches to explainable food risk prediction that can be further explored within EFRA.	
Lead Partner	
SU	
Contributing Partners & Roles	
Agroknow	Provide labeled dataset (see outcome SU1), and evaluation of explanations
Related Tasks	
Task 2.2 Novel explainable mining & analysis methods and tools Task 3.1 Process to design, train, and test explainable food risk prediction models Task 5.3 Experiments on performance, speed & accuracy	
Related Deliverables & Content	
D2.1 (2023-12)	Sample approaches for a library of targeted NLU modules to extract insights from food risk texts.
D3.1 (2024-03)	Exploration of different approaches to explainable risk prediction.
D5.2 (2023-12)	Evaluate submitted methods and their capabilities in terms of performance.
Current State-of-the-art (before EFRA)	
We lack an understanding of how useful different approaches are for the task of food hazard classification based on texts.	
Future State-of-the-art (after EFRA)	
The shared task will generate several different approaches to food safety prediction and provide inspiration for future development of novel methods within EFRA. There will also be a publicly available dataset for food safety prediction, which will introduce a new NLP task to a broader audience and stimulate research around the use of NLP in this application area.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Alignment call and task finalization SU and Agroknow	August 2023
Finish data exploration and dataset preparation	October 2023
Provide baseline models	October 2023
Submit the shared task	March 2024
Envisioned Challenges & Risks	

There is a risk that the proposed shared task is rejected.
Required Data & Potential Sources
Data and baselines from SU1.
Required Domain Expertise Support
We need Agroknow's domain experts for analysing the submitted explanations.

KPIs	
KPI SU2.1	Number of participants > 20
KPI SU2.2	Number of submitted explanations > 5
KPI SU2.3	Scientific publication in a peer-reviewed venue = 1

Outcome SU3: Explainable Human-in-the-Loop System for Food Hazard Classification
This outcome involves developing a human-in-the-loop classifier using conformal prediction and attention-based explainability for food hazard prediction. This system will be designed to help human experts quickly decide whether a food-related text involves a health risk or not. For this purpose, we will reuse the dataset created in outcome SU1 for the SemEval task to create a prototype of the system. Optionally, we can also use data from the EFRA Data Hub depending on the progress in that outcome.
Lead Partner
SU

Contributing Partners & Roles	
Agroknow, MAIZE	Provide labeled dataset (see outcome SU1, or EFRA Data Hub)
Agroknow, Agrivi, SGS	Expert opinions for evaluation of system's capabilities

Related Tasks
Task 2.2 Novel explainable mining & analysis methods and tools Task 5.3 Experiments on performance, speed & accuracy

Related Deliverables & Content	
D2.2 (2023-12)	Create a targeted NLP module for food risk prediction
D5.2 (2023-12)	Conduct and document experiments for a scientific publication.

Current State-of-the-art (before EFRA)
To the best of our knowledge, XAI methods have not been used to create human-in-the-loop text classifiers for food risk prediction, while the use of conformal prediction has not been explored much in NLP.
Future State-of-the-art (after EFRA)
We will combine conformal prediction and attention-based explainability to create a human-in-the-loop system for food hazard prediction. This will allow domain experts to categorize food texts based on a set of the most probable labels, where explanations (e.g. in the form of highlighted text) are provided for each label in the prediction set.

Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Provide a prototype for an explainable user-friendly HIL system on EFRA data	December 2023
Submit at least one scientific article on the model to a peer-reviewed venue	March 2024

Envisioned Challenges & Risks
There is a risk that we do not receive expert opinions for evaluation of system's capabilities in time.

Required Data & Potential Sources	
Data form outcome SU1 and EFRA Data Hub	
Required Domain Expertise Support	
We will need expert opinions on the helpfulness and capabilities of the system.	
KPIs	
KPI SU3.1	Scientific publication in a peer-reviewed venue >= 1
KPI SU3.2	Native system performance > baselines from outcome SU1

Outcome SU4: Literature Review on Explainability in NLP and Time Series	
This outcome involves conducting a literature review on explainability in NLP, including time series. The literature review will describe the state-of-the-art in explainability in the context of NLP, in particular language models, both for text classification and entity extraction.	
Lead Partner	
SU	
Contributing Partners & Roles	
All	PhD students from our partners will be invited to participate in the course.
Related Tasks	
Task 3.1 Process to design, train, and test explainable food risk prediction models	
Related Deliverables & Content	
D3.1 (2024-03)	Literature review for explainability for time-series predictions and NLP.
Current State-of-the-art (before EFRA)	
There is a need for an overview on explainability methods in NLP, including time series.	
Future State-of-the-art (after EFRA)	
We provide an overview of state-of-the art approaches to explaining text and time-series data. These approaches can be recycled by the EFRA partners in order to provide transparent ML solutions in the context of food safety.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Identify relevant publications	December 2023
Write and submit paper that summarizes the findings of the literature review	March 2024
Envisioned Challenges & Risks	
The plan is for the literature review to be carried out as part of a PhD level course and there is a risk that not enough students enroll and/or do not complete the course. We plan to minimize this risk by encouraging a minimum of two PhD students to work on one topic. The literature review may also be limited in scope and could then be presented in the related work section of another publication.	
Required Data & Potential Sources	
Publications in peer-reviewed scientific journals and conferences, published by IEEE, ACM and similar.	
Required Domain Expertise Support	
No domain expertise knowledge is requested at this moment.	
KPIs	
KPI SU4.1	Scientific papers reviewed > 10
KPI SU4.2	Scientific publication in a peer-reviewed venue = 1

8.3 Outcomes lead by CNR

Outcome CNR1: enhanced AI model for pest threats prediction	
AGRIVI uses a rule-based system and data from weather forecasts, scouting activities, and past information to enhance decision-making processes for farmers, including pest alarms based on weather data. The enhancement envisioned would involve developing and integrating a supervised predictive algorithm, which uses AI to forecast pest invasions with greater accuracy and suggest optimized responses. The AI model will optimize the efficiency/effectiveness trade-off, i.e., they allow more effective predictions while being more efficient and less energy-demanding.	
Lead Partner	
CNR	
Contributing Partners & Roles	
AGRIVI	AGRIVI leads the agricultural use case of scenario AG.1: Enhanced Predictive Capabilities for Pest Alarms. It provides CNR with technical support, domain knowledge, and training data for pest threats prediction and other farmers' decision-making processes; Other partners (SU, AGROKNOW, WFSR) will be kept informed.
Related Tasks	
Task 3.4 Enhancing AI sustainability and energy-efficiency [M10-M30].	
Related Deliverables & Content	
D3.1 – Models and Components for Risk Prediction	The link to D3.1 concerns the novel efficient AI models developed.
D3.2 – Report on Deployment of Risk Prediction Modules	The link to D3.2 concerns the deployment of novel efficient AI models developed.
Current State-of-the-art (before EFRA)	
AGRIVI currently uses a rule-based system exploiting weather data to provide pest alarms and predictions. They are also collecting feedback from the farmers but are not actively using this information in the prediction model. The existing tools, although efficient, are not as sophisticated or precise as they could potentially be with further enhancements.	
Future State-of-the-art (after EFRA)	
After EFRA, AGRVI will offer a more advanced prediction algorithm, that will allow for significantly enhanced prediction of pest invasions and optimized responses based on these predictions. The improved solution will feed more effective pest management strategies exploiting historical information and a machine learned solution, and potentially higher crop yields due to more precise and timely responses to pest threats.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Study and definition of data and AI models for pest threats prediction	September 2023
Novel more effective AI model for pest threats prediction	November 2023
Improved efficiency of AI model for pest threats prediction	January 2024

Evaluation and Final delivery	March 2024

Envisioned Challenges & Risks
The primary challenges and risks revolve around the development and integration of advanced AI algorithms in the AGRIVI platform. Accurately predicting pest invasions requires in-depth data analysis and a comprehensive understanding of various complex factors. These include climate conditions, specific crop types, regional infestations, and past pest behavior. The algorithm's complexity, their data requirements combined with the vast amount and heterogeneity of data that needs to be processed and understood, presents a significant challenge. Additionally, ensuring the new capabilities smoothly integrate with the existing platform without causing disruptions is another potential challenge.
Required Data & Potential Sources
In the first iteration, CNR will work on already-available pest threats prediction data, i.e., 1) pest occurrences, 2) weather conditions, 3) crop types, 4) pesticide usage. CNR will then refine its solutions by integrating other sources of data to achieve more accurate and more efficient predictions. Examples of these data are: 1) real-time environmental data, 2) specific pest behavior data, 3) global pest outbreak data.
Required Domain Expertise Support
1. Technical support: we ask technical support from AGRIVI to integrate new AI capabilities into the existing platform. This may include updating the system with new features, providing ongoing user support to handle any issues or queries. 2. Domain knowledge support: we ask Agrivi's team of agronomists and data scientists to provide essential insights and support during the development and testing stages of the new features. 3. Pilot support: we ask Agrivi to set up appropriate piloting activities to test the new features with current Agrivi users.
KPIs

KPI CNR1.1	Prediction accuracy > 20% w.r.t. pre-EFRA AI solution
KPI CNR1.2	Energy consumption < 50% w.r.t. first non-optimized AI solution
KPI CNR1.3	Number of risk prediction AI models deployed in real-world use-cases +1 (KPI 5.1)

Outcome CNR2: energy-aware summarization of regulatory data
The use case in scenario RG.1 aims to design a novel Automated Regulatory Analysis & Summarization Module to be integrated in the SGS DIGICOMPLY platform. Such component aims at drastically reduce manual effort by automatically generate key summaries and extracts from regulatory texts to enhance user experience and efficiency. As M15 outcome, CNR will optimize the effectiveness/efficiency trade-off of this component by exploiting advanced neural network compression techniques. The activity will continue until the end of task 3.4.
Lead Partner
CNR
Contributing Partners & Roles

SGS DIGICOMPLY	SGS provides EFRA with technical requirements, domain knowledge and collections of regulatory data for the use case in scenario RG.1: Automated

MAIZE, SU	Regulatory Analysis & Summarization Module. SGS assesses the quality and effectiveness of the techniques developed within EFRA for regulatory data analysis and summarization by collecting feedback for continuous improvement. Additionally, SGS will provide NLP expertise in fine-tuning LLMs for the specific text processing tasks.

Related Tasks

Task 3.4 Enhancing AI sustainability and energy-efficiency [M10-M30]

Related Deliverables & Content

D3.1 – Models and Components for Risk Prediction	The link to D3.1 concerns the novel efficient AI models developed.
D3.2 – Report on Deployment of Risk Prediction Modules	The link to D3.2 concerns the deployment of novel efficient AI models developed.

Current State-of-the-art (before EFRA)

Currently, the SGS DIGICOMPLY platform provides comprehensive regulatory data, without any summarization and interpretation. The fruition of the relevant information included in these data require significant manual effort from the users.

Future State-of-the-art (after EFRA)

By automatically generating effective key summaries and extracts from regulatory texts the SGS DIGICOMPLY platform would drastically reduce manual effort. This could significantly enhance user experience and efficiency. The optimization of the effectiveness/efficiency trade-off of this analysis and summarization component will remarkably reduce the cost for the platform needed to produce effective multilingual summaries for the collection of regulatory data.

Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Collection of requirements, study of state-of-the-art LLMs for summarization and information extraction, collection of multilingual text collection in the food safety domain for LLM fine-tuning.	September 2023
First prototype of fine-tuned component for summarization of regulatory data.	November 2023
First prototype of energy-aware compressed model for the summarization of regulatory data; Provision of human-assessed summaries	January 2024
Evaluation of efficiency and effectiveness of the compressed model w.r.t. the original one	March 2024

Envisioned Challenges & Risks

Energy-hungry LLMs, such as GPT-3, require significant computational power to train and run. Training these models involves massive amounts of data and extensive processing, which demands substantial energy resources and can contribute to carbon emissions. Large-scale deployment of these models for text summarization may lead to increased SGS costs and limited accessibility for users with limited resources. To address these challenges and risks, CNR will explore techniques for model compression, optimization, and energy-efficient computing. Developing more

lightweight models specifically designed for text summarization of regulatory data and adopting energy-efficient hardware can also help mitigate the environmental impact of energy-hungry LLMs. Moreover, CNR will investigate novel efficient and low-energy-demanding fine-tuning strategies to derive novel models that can be used by SGS to perform summarization over regulatory data.
Required Data & Potential Sources
• Regulatory texts and amendments, compliance guidelines, product and label regulations. Multilingual text collection in the food safety domain for LLM fine-tuning. • Entities, Ontologies and multilingual vocabularies in the food safety domain. • Human assessed summaries and salient information of a significant collection of regulatory data for the evaluation of analysis and summarization solutions.
Required Domain Expertise Support
Technical support: we ask technical support from SGS for understanding the requirements and expectations of their customers. Domain-knowledge support: we ask SGS data scientists to provide essential insights and support during the development and testing stages of the original and optimized summarization module. Pilot support: we ask SGS to set up appropriate piloting activities to test the accuracy and efficiency of the developed component with their users.

KPIs	
KPI CNR2.1	Summarization accuracy for the compressed model < 10% worse that the one of the original, non-compressed, model.
KPI CNR2.2	Energy consumption for the compressed model < 50% w.r.t. the original, non-compressed model.
KPI CNR2.3	Number of risk prediction AI models deployed in real-world use-cases +1 (KPI 5.1).

Outcome CNR3: Distributed AI learning	
AI systems in the agriculture and food industries rely significantly on vast data for training and optimization. However, obtaining adequate training data can be difficult for privacy and scarcity reasons, leading to the appeal of federating different organizations to develop more resilient and accurate predictive models. This approach requires combining the capabilities of distributed learning algorithms with novel knowledge and data-sharing strategies. By exploiting topically and horizontally partitioned training data, CNR will investigate a novel distributed learning scenario focusing on a learning-to-rank task. Unlike federated learning, our instance of distributed learning aims at learning a distinct model for each organization, specializing in the peculiar characteristics of the local data. In this case, data/knowledge sharing aims to enhance the generalization power of the local models.	
Lead Partner	
CNR	
Contributing Partners & Roles	
WFSR AGROKNOW	WFSR is the main accountable partners for task 3.3. The collaboration between CNR and WFSR can be useful to exchange know-how and reach the EFRA goals Contributing with a survival analysis use case at M15+
Related Tasks	

Task 3.3 Federated Learning and Semantic Interoperability [M13-M30]	
Related Deliverables & Content	
D3.1 – Models and Components for Risk Prediction	The link to D3.1 concerns the design of the novel AI models developed.
D3.2 – Report on Deployment of Risk Prediction Modules	The link to D3.2 concerns the deployment of the novel AI models developed.
Current State-of-the-art (before EFRA)	
To the best of our knowledge the task addressed in this activity, I.e., distributed learning of topically focused learning-to-rank models, is novel. Moreover, it is relevant and adaptable to EFRA scenarios. With AGROKNOW we will investigate its application to a task of survival analysis in the food risk domain after M15.	
Future State-of-the-art (after EFRA)	
CNR aims to exploit the global knowledge about the data owned by the different stakeholders to improve accuracy and robustness of the models learned using private data only. Our approach involves implementing a specific concept of distributed learning, involving the integration of a model trained on private data with external knowledge extracted from the other models trained in federated nodes.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Collection of requirements, study of state-of-the-art, preparation of training data	November 2023
Final Prototype of EFRA component for distributed learning-to-rank	January 2024
Evaluation of efficiency and effectiveness of the models learned in distributed way w.r.t. the ones trained locally and in a centralised node	February 2024
Paper writing and polishing	March 2024
Envisioned Challenges & Risks	
Since it involves addressing unexplored questions, the main risk of this activity is the inherent uncertainty and potential challenges due to the lack of prior knowledge, established methodologies, and existing references.	
Required Data & Potential Sources	
CNR has data for properly addressing the learning-to-rank task. However, the data available does not refer to the food domain. Thus, we require data from other sources for directly assessing the solution in the context of the EFRA project.	
Required Domain Expertise Support	
The support of domain experts is required to transfer the solution in the food domain and assess its utility.	
KPIs	
KPI CNR3.1	Articles published on peer-reviewed international journals/conferences at M15+ >= 1.

Outcome CNR4: privacy-aware green platform for distributed AI and data sharing

This outcome seeks to analyze and adopt technological solutions related to open-source container orchestration systems for the design and deployment of a cloud-based prototypal platform. This platform, dedicated to food risk safety analysis, will be enforced by federated and micro-service principles and will satisfy a variety of requirements: i) multi-tenancy; ii) scalability; iii) seamless distribution and management of workloads across different geographically distributed participants; iv) optimized resource usage and power consumption via green and hardware-aware scheduling algorithms; v) support of distributed and federated AI learning ensuring data privacy.

Lead Partner	
CNR	
Contributing Partners & Roles	
ALL	All partners contribute to the achievement of this goal by providing use cases, requirements, data to be federated, technical solutions, AI models, evaluation procedures
Related Tasks	
Task 4.1 Open cloud & edge architecture of EFRA Data & Analytics Infrastructure [M4-M21] Task 4.2 Re-allocating cloud & HPC resources for greener operations [M7-M33]	
Related Deliverables & Content	
D4.1 – EFRA Architecture & Green Operations	The link to D4.1 concerns the design of the EFRA platforms optimized for green operations.
Current State-of-the-art (before EFRA)	
Currently, stakeholders in the food safety domain typically rely on their own individual data, platforms and technological solutions. It is very likely that their solutions do not cover all the requirements that the EFRA platform intends to satisfy. Additionally, privacy concerns and strategic business considerations often act as barriers to the sharing of food safety data that could enable the training of more robust and useful food risk predictive models. Moreover, the use of energy-hungry AI in the food safety domain raises concerns about significant energy consumption and its environmental impact. The energy requirements also lead to higher operational costs, limiting accessibility to AI-driven solutions.	
Future State-of-the-art (after EFRA)	
Upon the completion of the EFRA project, there will be a platform possessing the above features and mitigating energy demands through energy-efficient algorithms and optimized hardware. At M15 we will have a complete design of the architecture and a preliminary prototype showcasing the features exploitable by the platform.	

Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Analysis and study of practical solutions and state-of-the-art literature	September 2023
Preliminary architectural design of the EFRA Platform	October 2023
Prototype implementation of some components	March 2024
Envisioned Challenges & Risks	
To the best of our current understanding, Kubernetes, with its rich ecosystem of solutions and scheduling algorithms, can represent the *de facto* standard over which to design the EFRA platform. This solution must be thoroughly evaluated and leveraged to align with the project's goals. Given the ambitious goals to be achieved, some of them naturally contrasting, the task of studying, selecting, and integrating the most suitable solutions is challenging and	

risky.
Concerning data sharing, an interesting notion to explore is that of *data space.* This concept comes from the EU document "A European strategy for data", and appears to be considered and developed in at least a very important project, Gaia-X. However, it should be noted that research and development in this area is still in a premature stage.

Required Data & Potential Sources

There are several open-source container orchestration systems available, with Kubernetes representing the widely accepted *de facto* standard. Kubernetes has a vibrant and rapidly evolving ecosystem that offers numerous tools, services, and extensions that can help to achieve the EFRA platform's goals. Several lines of active research focus on Kubernetes' container scheduling algorithms, which will highly likely be beneficial for the platform's goals.

Regarding data sharing and prevalent privacy concerns among stakeholders in the food safety sector, it may be interesting to analyze solutions that are being developed in EU projects such as GAIA-X. Here, it is of particular interest the notion of "data space", a layer in a platform that facilitates sharing of specific data types, enhancing data availability across the economy and society while ensuring that control remains with data producers.

Finally, results from other EU projects such as MobiDataLab, DEMETER, TheFSM, and BigDataGrapes, might be considered for extension and re-usage.

Required Domain Expertise Support

Pilot support: domain experts should support the piloting activities to test the efficiency and usability of the developed components and of the whole platform.

KPIs

KPI CNR4.1	Energy consumption with the green EFRA scheduler and compressed AI models < 50% w.r.t. the use of a traditional scheduler with non-compressed AI models.

Outcome CNR5: Explainable AI for food risk prediction

This outcome aims to analyze the applicability of methods that belong to the so-called field of eXplainable Artificial Intelligence (XAI) with focus on food risk prediction. The idea is to evaluate the applicability of well-established AI explainability approaches for EFRA risk prediction scenarios. CNR expertise regards XAI solutions tabular data using interpretable ensembles of decision trees.
That said, CNR will investigate new AI models and techniques that are explainable to the end user but also very efficient. In fact, more explainable models typically have fewer parameters and fewer interactions between features with respect to the black-box ones, which can be exploited to combine explainability and efficiency. To this regard, CNR will also explore the trade-offs between explainability and accuracy.

Lead Partner

CNR

Contributing Partners & Roles

SU	SU is one of the main accountable partners for task 2.2 and task 3.1. The collaboration between the two partners can be useful to exchange know-how and reach the EFRA goals.

Related Tasks

Task 2.2 Novel explainable mining & analysis methods and tools [M4-M36] Task 3.1 Process to design, train & test explainable food risk prediction models [M7-M30]	
Related Deliverables & Content	
D3.1 – Models and Components for Risk Prediction	The link to D3.1 concerns the novel efficient AI models developed.
D3.2 – Report on Deployment of Risk Prediction Modules	The link to D3.2 concerns the novel XAI models deployed.
Current State-of-the-art (before EFRA)	
To increase trust in AI-enabled proactive approaches, the AI model cannot only deliver actionable insights and recommendations but must also explain the process with which the insights and suggestions are reached, and in a very specific sense: the human expert must be shown explanations that assure him/her that if s/he had all the data available to the AI, and infinite time, s/he would reach similar insights and suggestions. However, not all AI models are explainable, and many powerful deep neural network approaches are hard to explain, creating a tension between explainability and performance.	
Future State-of-the-art (after EFRA)	
After EFRA, new dashboards based on novel explainable AI algorithms will be provided to support expert human decision makers in real-life scenarios. Specifically, novel approaches based on intrinsically interpretable tree-based boosting algorithms will be tested and evaluated for food risk prediction problems.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Literature review of XAI models specifically tailoring food risk prediction models	September 2023
Novel XAI efficient solution exploiting the trade-off in between accuracy, efficiency and explainability.	January 2024
Evaluation of the proposed solution	March 2024
Envisioned Challenges & Risks	
The risk involved with created a novel more efficient and explainable model is not actually succeeding in improving the state of the art in a specific way or that the data is not representative of the prediction outcome so the model might not achieve the desired performance.	
Required Data & Potential Sources	
The initial analysis will be done with benchmark dataset for the evaluation of effective and explainable AI models for tabular data for different tasks. The second part of the analysis will be done on data for a food-risk application provided by interested partners, in this case AGRIVI, in trying a new explainable approach for interpretable boosting.	
Required Domain Expertise Support	
Pilot support: we ask interested partners (AGRIVI) to provide the dataset and/or set up appropriate piloting activities to test the new models and the plausibility of the AI model explanations.	
KPIs	
KPI CNR5.1 **KPI CNR5.2**	Offering explainable predictions with a loss in prediction effectiveness < 10% w.r.t. non-explainable models trained on the same dataset.

	Number of articles submitted in peer-reviewed international journal/conference >= 1 for M15+.

Outcome CNR6: Compression of complex AI models for green AI inference	
By leveraging AI technologies, farmers and agricultural stakeholders can make data-driven decisions, optimize resource allocation, and mitigate environmental impact. Artificial intelligence models are becoming increasingly complex, requiring substantial computational resources for deployment and execution. The enhancement envisioned would involve reducing the computational burden of DNNs. First, this permits execution of AI models on resource-constrained devices. Second, the compression of AI models will contribute to reducing the carbon footprint associated with AI deployments in agriculture.	
Lead Partner	
CNR	
Contributing Partners & Roles	
AGROKNOW	Provide domain expertise for the evaluation of the compression techniques proposed to the food sector. AGROKNOW will also help with the investigation of the impact of data locality and caching strategies of sparse neural network derived by the application of compression methods.
Related Tasks	
Task 3.4 Enhancing AI sustainability and energy-efficiency [M10-M30]	
Related Deliverables & Content	
D3.1 – Models and Components for Risk Prediction D3.2 – Report on Deployment of Risk Prediction Modules	The link to D3.1 concerns the novel efficient AI models developed. The link to D3.2 concerns the deployment of novel efficient AI models developed.
Current State-of-the-art (before EFRA)	
In this task, CNR aims to define a novel compression technique for Deep Neural Networks that combines two well-known research lines in literature, i.e., I) low-bit quantization (binarization) with ii) pruning. We aim to jointly maximize the accuracy achieved by the network while minimizing its memory impact by identifying an optimal partition of the network parameters among these two sets. CNR intends to work at the definition of the novel approach by explicitly tailoring efficiency by explicitly exploiting data locality and caching strategies on modern CPUs.	
Future State-of-the-art (after EFRA)	
A novel, efficient, and application-generic model compression technique that, by combining low-bit quantization and pruning, can enable a transformative shift in AI-powered systems for agricultural applications.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Design of a novel deep neural network compression technique that blends pruning and quantization in the same optimization framework.	September 2023
Implementation of the novel deep neural network compression technique along with state-of-the-art baselines and competitors.	December 2023

Evaluation of the novel compression technique and writing of a research article to be submitted to top-notch international journals or conferences	March 2023

Envisioned Challenges & Risks

Model compression allows for reducing the computational requirements of AI-based systems. However, it is important to identify the optimal trade-off between resource demands and model accuracy to ensure the usability of the tool is not compromised. The optimal trade-off point depends on the model, on the requirements, and on the specific features of the application. For this purpose, there exist various model compression techniques, such as quantization, pruning, and knowledge distillation, that allow reduction of the computational requirements of AI-powered models by targeting a specific dimension of the problem. As an example, pruning allows reduction of the memory footprint of pre-trained neural networks, while quantization permits for efficient inference on almost any computational platform, e.g., CPU, GPU, FPGA.

Required Data & Potential Sources

CNR will work on public datasets and benchmarks that are already available in literature.

Required Domain Expertise Support

The activity envisioned so far is academic and we need the support of domain experts (AGROKNOW) to address the food domain.

KPIs

KPI CNR6.1	Accuracy for the compressed model < 10% worse that the one of the original, non-compressed models.
KPI CNR6.2	Inference time for the compressed model < 30% w.r.t. the original, non-compressed model.
KPI CNR6.3	Energy consumption for the compressed model < 30% w.r.t. the original, non-compressed model.
KPI CNR6.4	Number of articles published on peer-reviewed international journals/conferences >= 1.

Outcome CNR7: Report on the current EFRA AI-enabled technologies

This outcome involves preparing an extensive report on the current AI-enabled technologies used by each use case partner, focusing primarily on i) how data is gathered; ii) how computational and storage resources are employed; iii) which kind of AI models are used for the predictive tasks already in place. The objective is to establish a more comprehensive understanding of opportunities for cloud-based or edge-based computations or other approaches for energy savings and appropriate pooling-up of resources between partners along the vision of an integrated platform for food risk analytics and federated AI training.

Lead Partner

CNR

Contributing Partners & Roles

MOY, AGRIVI, SGS, Agroknow	AGRIVI, SGS, MOY, and Agroknow need to provide information on their respective use cases to outline the technologies currently used by the consortium partner before the EFRA project.

Related Tasks

Task 1.3 Energy-efficient Cloud/Edge HPC architecture & integration requirements Task 4.2 Re-allocating cloud & HPC resources for greener operations	
Related Deliverables & Content	
D1.1 – EFRA Requirements Roadmap (M9)	Initial report presenting the findings of the study, based on the current technologies used by the consortium partner and on available state-of-the-art solutions aimed at enhancing the efficiency of AI-enabled food risk prevention through the EFRA Tools.
D4.1 – EFRA Architecture & Green Operations (M10)	The link to D4.1 concerns the design of the EFRA platforms optimized for green operations.
Current State-of-the-art (before EFRA)	
Prior to EFRA, partners Agroknow, SGS, MOY, and AGRIVI were already active in incorporating new data sources and records into their respective platforms and employing AI models on the gathered data for AI predictive tasks. However, all these tools were not shared and integrated in a unified solution, thus being inefficient, not energy-aware, redundant, and sub-optimal.	
Future State-of-the-art (after EFRA)	
Upon the successful realization of the EFRA initiative, there will be an integrated distributed solution balancing the load between cloud and edge-based computations and delivering advances in green AI training and deployment of predictive AI models.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Report on the use-case partners technologies	June 2023
Review of technological solutions aimed at exploiting opportunities for improving energy efficiency and pooling-up of resources between partners	July 2023
Roadmap of concrete outcomes concerning the integrated EFRA platform	September 2023
Envisioned Challenges & Risks	
Potential challenges include the difficulties in collecting the relevant information from use case partners and the expected heterogeneity of the data, methodologies and tools used. Point-to-point bilateral meetings will be planned to mitigate the first risk.	
Required Data & Potential Sources	
Outcome CNR7 will rely on the information provided by the use-case partners related to how data are collected and from which sources, which AI predictive models are used, which resources are employed for training and optimizing such models. Once this information is gathered, we can think about a unified platform for sharing resources and data in such a way to lower the power consumption and create more efficient AI procedures.	
Required Domain Expertise Support	
Support and active participation are expected from all the use-case partners.	
KPIs	
KPI CNR7.1	Surveys compiled by use-case partner = 4
KPI CNR7.2	Energy-efficient technological solutions analyzed >= 2

8.4 Outcomes lead by MAIZE

Outcome MAIZE1: Report on Data Sources	
Structured Report of provided data sources (news websites and blogs, institutional portals, video channels, scientific publications aggregators) along with the given dimensions: availability, ownership, quality, reliability, and format.	
Lead Partner	
MAIZE	
Contributing Partners & Roles	
AGROKNOW, SGS	Partners might point out additional data sources to be assessed
Related Tasks	
T1.2 Heterogeneous data mining requirements	
Related Deliverables & Content	
D1.2	Data Sources will be assessed
Current State-of-the-art (before EFRA)	
Not applicable	
Future State-of-the-art (after EFRA)	
Not applicable	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
~~Source Assessments template document~~	~~May 2023~~
Draft Document	August 2023
Envisioned Challenges & Risks	
Not applicable	
Required Data & Potential Sources	
Not applicable	
Required Domain Expertise Support	
Reviewing the Report document	
KPIs	
KPI MAIZE1.1	Number of assessed web sources > 20
KPI MAIZE1.2	Number of assessed video sources > 5
KPI MAIZE1.3	Number of assessed scientific publications aggregators > 5

Outcome MAIZE2: Video Crawler
Development of a video crawler capable of periodically monitoring a set of video channels (configured in the Data Source Registry), downloading any new content in a staging area, then processing it through a NLP pipeline (video segmentation, transcription and relevance classification using AI models) and finally saving the collected relevant data in the EFRA Data Hub to be further processed and enriched. Transcribed contents will maintain a reference to

the original video fragments. The pipeline will be implemented following a micro service approach to enable experimenting/testing/usage of different solutions (e.g., different relevance classifiers)	
Lead Partner	
MAIZE	
Contributing Partners & Roles	
SU	Evaluating and integrating (in future iterations) the same classifier used in the Web Crawler component (AGK1)
AgroKnow	Evaluating domain relevance of the video fragments transcriptions
Related Tasks	
T2.1 EFRA Data Hub population T1.2 Heterogeneous data mining requirements	
Related Deliverables & Content	
D2.1 (due in December 2023)	The model's design along with a first implementation will be prepared.
D1.2 (due in March 2024)	Video sources will be assessed
Current State-of-the-art (before EFRA)	
At present, Agroknow and similar food intelligence companies do not automatically integrate video contents in their processes of data collection, ingestion and analysis.	
Future State-of-the-art (after EFRA)	
Post-EFRA, contents from video sources will be integrated in the Source Registry and in the Data Hub	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Collect the relevant datasets	June 2023
Design the NLP Classifier	July 2023
Train the NLP Classifier & onboard in a crawler	September 2023
Conduct experiments & measure performance	December 2023
Envisioned Challenges & Risks	
The main technological challenges involve performances of the video segmentation and classification of the source data.	
Required Data & Potential Sources	
The video channels currently analysed and assessed are: EFSA, USDA and CDC channels on YouTube	
Required Domain Expertise Support	
Required domain expertise for evaluating collected fragments transcriptions' relevance	
KPIs	
KPI MAIZE2.1	Number of integrated video channels > 3
KPI MAIZE2.2	Precision > 80% of correctly classified transcriptions by the NLP Classifier
KPI MAIZE2.3	Relevant scientific publication in a peer-review venue = 1

Outcome MAIZE3: Semantic Backbone

Defining a common conceptual model for EFRA by exploring state-of-the-art ontologies for Food Risk Analysis and selecting relevant standards to ensure data interoperability both internally and between other EU Data Hubs. The selected standard(s) will be extended, if needed, with missing domain concepts (e.g., food risk/safety concepts). Designing and implementing an automated process for semantically aligning historical annotated data from partners and processing such data and mapping them to the common EFRA model, taking into account the following concepts: • Food safety incidents [Agroknow] • Data relevant to the use and expected behaviour of pesticides [Agrivi] • Regulatory information, especially concerning MRLs [SGS] • Pathogen sampling and other relevant data (e.g., WGS) [MOY]
Lead Partner
MAIZE
Contributing Partners & Roles

Agroknow, Agrivi, SGS	Providing historical weather data, annotated historical data about Food Safety incidents and regulations, and providing support for mapping them to EFRA conceptual model.

Related Tasks
T2.2 Semantic backbone, data annotation & interoperability
Related Deliverables & Content

D2.2 (due in December 2023)	The first instance of EFRA conceptual model, to be further enhanced / expanded if needs arise from the use cases

Current State-of-the-art (before EFRA)
Despite the interest from organisations and stakeholders, currently a standard for data interoperability in the food safety domain does not exist.
Future State-of-the-art (after EFRA)
Post-EFRA, a common framework for describing food related data and food safety/risk prediction will be available.
Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Report on the state-of-the-art ontologies and dictionaries.	July 2023
Draft version of EFRA Conceptual Model	September 2023
Semantic Alignment of historical data	October 2023
Common EFRA Conceptual Model	December 2023

Envisioned Challenges & Risks
The main challenge with respect to the semantic annotations remains and originates from the diversity of data types considered.
Required Data & Potential Sources
Relevant research initiatives for harmonising food data and providing a common semantic framework for food safety and traceability include the FoodOn and ISO-FOOD ontologies and the multilingual thesaurus Agrovoc of FAO
Required Domain Expertise Support
Required domain expertise for evaluating the coverage of the conceptual model
KPIs

KPI MAIZE3.1	Number of mapped/covered concepts from historical data > 85%

Outcome MAIZE4: Data Annotation Tool

Configuration of a Data Annotation Tool for the manual annotation of data (historical data as weel as harvested by EFRA crawlers). Whenever it is possible (e.g., entities detection, relevance classification) we aim at integrating AI classifiers (trained on historical data and public datasets) in order to pre-annotate the collected data and to involve domain experts only for the validation of contents and for the more complex annotations (e.g., relations between entities, risk detection).
Lead Partner
MAIZE
Contributing Partners & Roles

SU	Define the procedure for training models on the annotated data (e.g., either downloading annotated data or integrating the NLU models in label studio)

Related Tasks
T2.2 Semantic backbone, data annotation & interoperability
Related Deliverables & Content

D2.2 (due in December 2023)	The annotation tools.

Current State-of-the-art (before EFRA)
Not applicable
Future State-of-the-art (after EFRA)
Not applicable
Subtasks & Timing up to next Deliverable

Subtask	**Delivery Month**
Set up and test the environment	July 2023
Importing pre-annotated historical data	October 2023
Hands-on Annotation Workshop on harmonized historical data (online events for EFRA domain experts)	November 2023

Envisioned Challenges & Risks
The main technological challenges involve the mapping of historical data and the granularity of the labels
Required Data & Potential Sources
An open-source component named label studio (https://labelstud.io/)
Required Domain Expertise Support
Required domain experts fort testing the environment.
KPIs

KPI MAIZE4.1	Annotation Platform up and running.
KPI MAIZE4.2	Organize and host a first Annotation workshop, involving all domain expert partners

Outcome MAIZE5: EFRA Data and Analytics Architecture
Design of the EFRA platform architecture, including the Data Hub (storage for configurations, models and raw/enriched data) and the Analytics Powerhouse (AI model execution/scheduling over the Data Hub). Based on open, distributed/cloud/edge, and micro-service-oriented principles. EFRA architecture must enable the overall project's goals of federated learning (privacy by design) and Green AI (optimizing resources allocation and consumption).
Lead Partner
MAIZE

Contributing Partners & Roles	
CNR	Feedback on Architecture and technological solutions with respect to the interactions with HPC architecture.
WFSR	Feedback on Architecture and technological solutions with respect to the Federated Learning framework.
Related Tasks	
Task 4.1 Open cloud & edge architecture of EFRA Data & Analytics Infrastructure	
Related Deliverables & Content	
D4.1 (due in October 2023)	The design of the architecture
Current State-of-the-art (before EFRA)	
Currently Federated Learning is not supported by EFRA partners and Green AI principle and strategies are not taken into account in their architectures.	
Future State-of-the-art (after EFRA)	
After EFRA Federated Learning will be enabled and strategies for optimizing resources and energy consumption will be implemented.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Architecture Design	September 2023
Envisioned Challenges & Risks	
The main technological challenges involve the design of the Federated Learning components.	
Required Data & Potential Sources	
The architecture will be implemented over a Kubernetes cluster to take full advantage of configurable Kubernetes resources management and its horizontal scalability, as well as the virtualization of software modules (e.g., crawlers, NLU modules, AI risk prediction, etc) through Docker container	
Required Domain Expertise Support	
No domain expertise knowledge is requested at this moment.	
KPIs	
KPI MAIZ5.1	Document with the first release of EFRA Data and Analytics Architecture

Outcome MAIZE6: Data Hub
Design and development of EFRA Data Hub, a data centric, cloud-based storage solution that allows for the storage, annotation, enrichment and retrieval of EFRA data points and supports the different phases of data processing (e.g., weather data, raw data stored by crawlers, fed into annotation platform, enriched by multilingual NLU models, grouped in clusters by means of intelligent linking components and acting as input for risk prediction models). The Data Hub will be accessible by means of APIs that will be described and documented in D4.2.
Lead Partner
MAIZE
Contributing Partners & Roles
All EFRA partners will be contributing to this outcome.
Related Tasks
T4.3 Deployment of EFRA Data Hub and Analytics Powerhouse (T2.1 EFRA Data Hub population)
Related Deliverables & Content

D4.2 (due in December 2023)	First implementation of the Data Hub, including the documented APIs for saving, updating and accessing food safety data.
Current State-of-the-art (before EFRA)	
Not applicable	
Future State-of-the-art (after EFRA)	
Not applicable	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Kubernetes cluster setup	June 2023
Data Hub, APIs design and documentation	October 2023
Data Hub, First Implementation	December 2023
Envisioned Challenges & Risks	
To be initially populated it requires data collected by crawlers (video, web). If not yet available, historical data can be used for the initial testing phase.	
Required Data & Potential Sources	
Elasticsearch 8 (documents relations and dense vector types)	
Required Domain Expertise Support	
No domain expertise knowledge is requested at this moment.	
KPI	
KPI MAIZE 6.1	APIs documented
KPI MAIZE 6.2	Data Hub is online

Outcome MAIZE7: Analytics Powerhouse	
Design and implementation of the Analytics Powerhouse, a component capable of scheduling and monitoring AI model execution. The component will expose standard REST APIs to load and validate an AI model (defined in a docker image), assigning an ID to the model and storing it in the Data Hub (along with timestamp and an ID) and allow for its on-demand, scheduled or event-based execution. The powerhouse will enable evaluating analytics components with respect to benchmark datasets (in the Data Hub) enriched with annual annotations on the expected analysis outcome (i.e., a gold standard for each task offered by the analytics components); a dedicated API of the powerhouse will provide performance metrics from the automated tests (e.g., trends, comparison).	
Lead Partner	
MAIZE	
Contributing Partners & Roles	
CNR	Feedback on Architecture and technological solutions with respect to the interactions with HPC architecture.
Related Tasks	
T4.3 Deployment of EFRA Data Hub and Analytics Powerhouse	
Related Deliverables & Content	
D4.2 (due in December 2023)	Definition of the Analytics model deployment and validation strategy. First implementation of the Analytics Powerhouse, allowing loading a model into the Data Hub and then starting, stopping and monitoring its execution.
Current State-of-the-art (before EFRA)	
Not Applicable	

Future State-of-the-art (after EFRA)	
Not Applicable	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Kubernetes cluster setup	June 2023
Powerhouse, APIs design and documentation	October 2023
Powerhouse, First Implementation	December 2023
Envisioned Challenges & Risks	
None	
Required Data & Potential Sources	
None	
Required Domain Expertise Support	
None	
KPIs	
KPI MAIZE 7.1	APIs documented
KPI MAIZE 7.2	APIs online

8.5 Outcomes led by WFSR

Outcome WFSR1: Baseline Early Warning System for Unknown Risks	
We need to identify long-term, systemic, unknown risks in the food supply chain and predict their insurgence. The emergence of unknown systemic risks often causes long-term shocks on the food supply chain which can be mitigated with short term solutions (i.e., early warning). Consequently, the development of predictive models becomes crucial for a risk prevention approach. This task will focus on developing a library of trained, tested, and calibrated prediction models for long-term, systemic, unknown risks, with significant performance, speed, and accuracy. These models will use state-of-the-art machine/deep learning approaches and supervised and unsupervised techniques over heterogeneous/unstructured data through a direct link with T2.2.	
Lead Partner	
WFSR	
Contributing Partners & Roles	
MOY	The data that will be used to train and test the system can be provided by MOY.
SU	T2.2 will turn the noisy textual data into high-quality food risk signals for WP3 AI models
Related Tasks	
Task 1.1 Scientific requirements on short- and long-term food risk prediction Task 3.2 Methods & tools for prediction of long-term unknown risks	
Related Deliverables & Content	
D3.1 (due in March 2024)	Models and components for risk prediction
Current State-of-the-art (before EFRA)	
Current systems mainly focus on predicting known risks. A few systems that are developed for predicting unknown risks are mainly based on thresholds of manually identified parameter values.	
Future State-of-the-art (after EFRA)	
The output of this system will provide experts with an automated system that requires much less manual intervention and will yield precise and complete unknown risk predictions.	

Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Alignment call add to this document. MOY and Agroknow Requirements analysis	September 2023
Training and test data preparation	November 2023
Baseline system	March 2024
Envisioned Challenges & Risks	
Description of the concept of an 'unknown risk' and collection of a dataset that can be used for training and evaluating a system for predicting unknown risks is a challenge. When does a risk become unknown? Moreover, the concept of 'early' should be identified as well. The following lines of research will be investigated to identify the optimal solution for our target system: outlier detection, one-class classification, and domain generalization. Another type of experiment that will be conducted is a leave-one-risk-out training and evaluation setting on a known risks dataset.	
Required Data & Potential Sources	
This baseline system will be developed on available datasets that should be identified during the literature review for this task.	
Required Domain Expertise Support	
Definition of unknown risks in the scope of EFRA should be refined in collaboration with domain experts.	
KPIs	
KPI WR1.1	Relevant scientific publication in a peer-reviewed venue = 1
KPI WR1.2	A system that predicts unknown risks with recall = 1 & precision > .20

Outcome WFSR2: Baseline Federated Learning System	
Data sharing is not always possible across stake-holders. But utilization of all possible data for a challenge like food safety is mandatory as use of big and diverse data in training machine learning systems yields optimal performance. In addition to accuracy, these systems should be privacy-preserving and explainable.	
Lead Partner	
WFSR	
Contributing Partners & Roles	
AGROKNOW CNR Moy Park	A baseline model that predicts food risk will be the base of WFSR's efforts. CNR will guide the setup of the computational infrastructure for this use case. MOY Park will provide data and assess quality of the model.
Related Tasks	
Task 1.4 Public & private data for AI training and data sharing requirements Task 2.3 Semantic backbone, data annotation & interoperability Task 3.3 Federated Learning and Semantic Interoperability Task 5.2 Experimental Methodology, Use-case Plan, and Recommendations Task 5.3 Experiments on performance, speed & accuracy	
Related Deliverables & Content	
D3.1 (due in March 2024)	Models and components for risk prediction
Current State-of-the-art (before EFRA)	

For the food safety domain such infrastructure and related AI models have been tested in the laboratory of WFSR and showed promising results.
Future State-of-the-art (after EFRA)
Since new developments in this domain are in flux, new AI models will be built for food safety and food fraud applications. These models could be equally applicable in other domains and shall be contributed to the open source community and made available on version control and code sharing repositories (e.g. git). The relevant developments will be demonstrated in the corresponding EFRA use-cases.
Subtasks & Timing up to next Deliverable

Subtask	Delivery Month
Review	September 2023
Semantic backbone	November 2023
FAIR data station setup	January 2023
Baseline model	March 2023

Envisioned Challenges & Risks
Measuring the performance improvement of a model trained in a federated learning setting is a challenge. The combination of the performance scores at each station should be carefully performed in order to ensure a proper evaluation. Label errors and interoperability issues pose a risk in this line.
Required Data & Potential Sources
Training and evaluation data will be utilized without being transferred to WFSR.
Required Domain Expertise Support
Domain experts are needed from each partner that allow utilization of their data in order to ensure semantic interoperability of the data.
KPIs

KPI WR2.1	Relevant scientific publication in a peer-reviewed venue = 1
KPI WR2.2	FAIR data stations > 2
KPI WR2.3	Performance of the model created using Federated learning > Performance of all models created using data only from one station

Outcome WFSR3: Scientific Review of Explainability and Privacy-preservation in Federated Learning
Federated learning ensures the input needed for modelling is used at data source without any data sharing. This setting is privacy-preserving and enables building trust in modelling using data from multiple stakeholders. The predictions of the model created should be explainable too so that trust and usability of the predictions could be increased. However, there may be a trade-off between privacy-preserving characteristics and explainability of a model. A detailed analysis is required to reveal this relationship in a federated learning setting.
Lead Partner
WFSR
Contributing Partners & Roles

SU CNR	Collaboration on expertise on explainability of machine learning models

Related Tasks
Task 2.2 Novel explainable mining & analysis tools
Related Deliverables & Content

D2.1 (June 2024)	EFRA Data Registry, Discovery & Mining Stack

D3.1 (March 2024)	Models and Components for Risk Prediction
Current State-of-the-art (before EFRA)	
Current state-of-the-art at the intersection of explainability and privacy preserving machine learning has not been investigated	
Future State-of-the-art (after EFRA)	
The trade-off between explainability and privacy preservation in food domain will be investigated and reported in a federated learning scenario in detail.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Keynote at EFRA virtual summit	June 2023 (done)
Literature review	December 2023
Experimental setting ready	February 2024
Detailed report	March 2024
Envisioned Challenges & Risks	
None for the next subtask.	
Required Data & Potential Sources	
No additional data requested at this moment.	
Required Domain Expertise Support	
No domain expertise knowledge requested at this moment.	
KPIs	
KPI WR3.1	Relevant scientific publication in a peer-reviewed venue = 1

Outcome WFSR5: XAI to identify sensitive information from federated learning	
Federated learning is designed to be privacy-conserving, i.e., to keep sensitive information being accessed by unauthorized parties. While models such as deep neural networks become more and more complex, it is possible that sensitive information might somehow be stored in these models. In this task, we will identify whether we can identify sensitive information from federated learning platforms using eXplainable Artificial Intelligence (XAI).	
Lead Partner	
WFSR	
Contributing Partners & Roles	
CNR, SU	We ask partners for input and collaboration. CNR and SU expressed interest.
Related Tasks	
2.2, 3.1	
Related Deliverables & Content	
D3.1 (March 2024)	This is a new proposed topic so no clear deliverables have been set yet. M24, second version of this deliverable.
Current State-of-the-art (before EFRA)	
Evaluation of privacy conservation for federated learning in food safety with XAI has not yet been performed.	
Future State-of-the-art (after EFRA)	
Ability of XAI to extract sensitive information from federated learning is assessed	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
Implement federated AI learning systems with different complexity	TBD

Use XAI to probe the federated learning systems to identify private information	> M18
Disseminate obtained knowledge	> M18
Envisioned Challenges & Risks	
None at the moment	
Required Data & Potential Sources	
None at the moment	
Required Domain Expertise Support	
No domain expertise knowledge is requested at this moment.	
KPIs	
KPI WR5.1	Scientific paper in peer reviewed journal = 1

8.6 Outcomes lead by RAINNO

Outcome RAIN1: Maximize the Project's Impact in line with DEC plan	
This outcome involves all the Dissemination, Communication and Exploitation activities, we have engaged to implement the first 18 months of the project through the DEC plan. These activities refer to Rainno's, as well as, to partners' obligations, according to the allocation of the relevant KPIs, considering the PMs and budget dedicated to WP6.	
Lead Partner	
RAIN	
Contributing Partners & Roles	
ALL	All partners have been assigned with Communication and Dissemination KPIs for contributing to the maximisation of the project's impact. The KPIs and targets can be found in the GA and in the project's shared folder. The allocation of the KPIs among partners and reporting periods of the project is in the shared folder, as well.
Related Tasks	
Task 6.1 Dissemination, exploitation and community engagement Task 6.2 Building and engaging a Public-Private Network on risk prediction Task 6.3 Involvement in industry groups and networks	
Related Deliverables & Content	
D6.1 v2 (due in December 2023)	The second updated version of the DEC plan with the results of the outreach activities.
Current State-of-the-art (before EFRA)	
• Not applicable for the DEC activities	
Future State-of-the-art (after EFRA)	
• Not applicable for the DEC activities	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
D&C Partners Reporting • Enrich and annotate DEC Plan highlighting important things for each partner specifically. • Send link with DEC Plan Working document.	June 2023-May 2024

• Reminders will be sent each month for reporting activities to partner's assigned person.	
Envisioned Challenges & Risks	
Not applicable for the DEC activities	
Required Data & Potential Sources	
All partners to report on their activities and project results	
Required Domain Expertise Support	
Expert support for Dissemination, Communication and Exploitation of EFRA'S activities and results will be provided by RAINNO. WFSR and CNR can facilitate the access to public sector stakeholders.	
KPIs	
KPI RAIN 1.1	D&C KPIs as written in the GA

Outcome RAIN2: Business Models and IPR Management	
This outcome involves an initial IP partners' collection that will be conducted as well as Business Models state-of-play, both compiled into the first version of D6.2.	
Lead Partner	
RAIN	
Contributing Partners & Roles	
ALL	• All partners to provide an update on their exploitation plans and any IPR that may arises • Use cases to provide information about their existing business models
Related Tasks	
Task 6.4 EFRA Sustainability Plan and Business Models	
Related Deliverables & Content	
D6.2 v1 (due in December 2023)	The toolset consists of an IPR plan, a Business Models Playbook and a Sustainability plan.
Current State-of-the-art (before EFRA)	
• Prior to EFRA, no tailored made Business Model exists addressing specific food risk cases.	
Future State-of-the-art (after EFRA)	
• Tailor-made Business Models for the 3 use cases aiming to provide incentives for adopting the EFRA proposed data and analytics solution.	
Subtasks & Timing up to next Deliverable	
Subtask	**Delivery Month**
IP Partner's Collection	October 2023
IP 1st Webinar	December 2023
Envisioned Challenges & Risks	
• Conflicts over ownership of the exploitable results • Reluctance on providing accurate data on the existing business models for the use cases	
Required Data & Potential Sources	
We need to analyse the current state of play on the business models in order to design tailor-made business models for each use-case incorporating the trends in the market coupled with feedback obtained through the two cycles of each use-case.	
Required Domain Expertise Support	

• Domain experts will be needed to provide us with relevant information about the existing business models in the use cases.	
KPIs	
KPI RAIN 1.1	An analysis with the existing business models and the market
KPI RAIN 1.2	1 IP EFRA Webinar

9 Conclusions

The work conducted in the scope of this deliverable serves the purpose of identifying and analyzing data and system development options available for designing systems for emerging food safety risk prediction. The capacities of the EFRA consortium partners are documented in line with the goals of the EFRA project. This practice yielded a roadmap, that consists of actions for tackling one or more challenges described as tasks and goals in the EFRA Grant Agreement. The information reported in this document such as drivers of emerging food safety risk and federated learning design is what EFRA consortium need for implementing the use cases. The data and computational resources available and needed are described comprehensively.

The collaboration with industrial partners of EFRA consortium has served as a formalization of their problems and preparation for the implementation of the respective uses cases. The preparation is at both sides as industrial partners are formatting and sharing data in a way that can be used for machine learning and the other partners identified solutions and requirements that suit use case owners.

The next steps will be the detailed design and implementation of the machine learning systems that will mainly predict emerging food risk events in a real-time big data setting. The progress will be tracked using the roadmap provided in Chapter 8. The roadmap will be updated in Deliverable 1.2, which is due to M15 of the EFRA project, according to developments in the implementation.

Annex I: Individual data source assessment notes

Abu Dhabi Agriculture and Food Safety Authority

The website provides access to about 20 PDF documents dated 2013 to 2015.

URL

- https://www.adafsa.gov.ae/English/Foodcontrol/Pages/default.aspx

T&C

- There are no database rights in UAE, but data is subject to copyright by Abu Dhabi Agriculture and Food Safety Authority.

Abu Dhabi Food Control Authority

The website is a portal through which you can request various services. It does not seem to offer information, so we did not assess this data source any further.

URL

- https://www.adafsa.gov.ae/english/Pages/default.aspx

ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail)

Data from July 1996 onwards is available, about 4 documents each month.

URL

- https://www.anses.fr/fr/content/avis-et-rapports-de-lanses-sur-saisine

T&C

- Use of data requires attribution.
- Full T&C at https://www.anses.fr/en/content/legal-information

  The information and data on the ANSES website www.anses.fr are public; they can be freely reused, without charge, in accordance with Articles L. 321-1 et seq. of the French Code of Relations between the Public and the Administration. They are protected by the Berne Convention for the Protection of Literary and Artistic Works, by other international conventions and by national laws on copyright and associated rights.

  Information and data from the ANSES website can be reused for personal or public purposes, provided that ANSES (and where applicable, the partners associated with the data or information) is mentioned as the source of the information.

  Any reproduction, translation or reuse of the information contained on the ANSES website is subject to the condition that the public information is not altered, its meaning is not distorted and the date of the latest update is indicated.

  Therefore, any data or information from the ANSES website must be repeated in full, without any modification or addition and without the addition of advertising; it must be available for downloading free of charge.

  At no time shall the reuse of data, regardless of the media or form, give the impression that ANSES participates in or endorses the action of the user.

  The ANSES logo is a registered, protected design whose use is authorised in limited cases. Any use requires a request to be submitted to ANSES via the website's contact form; it can only be granted on formal written authorisation of the Director General of ANSES.

Person responsible for access to administrative documents and questions related to the reuse of public information: Bérénice Renard, Legal Affairs Director. This adress mail can be used: demande.acces@anses.fr.

Australian Department of Agriculture Imported Food Reports

Data from January 2017 onwards is available, 77 monthly PDF reports with data laid out in tables.

URL

- https://www.agriculture.gov.au/biosecurity-trade/import/goods/food/inspection-testing/failing-food-reports

T&C

- Use of data requires attribution. Data is copyright Australia, Department of Agriculture, Fisheries and Forestry.
- There are no database rights in Australia.
- Full T&C at https://www.agriculture.gov.au/about/copyright
  Unless otherwise specified, you can use our material under a Creative Commons Attribution (CC-BY) 4.0 International licence.

Austrian Food Safety Authority

About 300 articles and reports dated 2021 and onwards.

URL

- https://www.ages.at/en/

T&C

- Data is copyright Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH, subject to EU sui generis database rights.
- No specific T&C are available.

BarfBlog

About 12,500 blog posts dated between 2007 and 2021, last updated July 2021.

URL

- https://www.barfblog.com/

T&C

- Data is copyright Doug Powell and Ben Chapman, based in Australia.
- There are no database rights in Australia.
- No specific T&C are available.

Brazilian Health Regulatory Agency

About 10,000 blog posts dated 2015 onwards

URL

- http://antigo.anvisa.gov.br/alertas

T&C

- Data is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported License.
- There are no database rights in Brazil.

BVL German Federal Office of Consumer Protection and Food Safety

Available historical data consists of 493 PDF press reports, 944 announcements, mainly in German (30 are in English), and 250 warnings in HTML tables

URL

- Reports: https://www.bvl.bund.de/EN/Service/MediaCenter/03_reports/infothek_berichte_node.html
- Warnings: https://www.lebensmittelwarnung.de/bvl-lmw-de/liste/lebensmittel/deutschlandweit/10/24

T&C

- Data is copyright Federal Office of Consumer Protection and Food Safety, subject to EU sui generis database rights
- Commercial use is not permitted, attribution is required, BVL consent is required.
- Full T&C https://www.bvl.bund.de/DE/Meta/Impressum/impressum_node.html
  All content from this website may only be published if the source is stated (FEDERAL OFFICE FOR CONSUMER PROTECTION AND FOOD SAFETY, [date]: [document title], [URL], status: [date]) may be published or passed on to third parties. The same applies to any form of duplication, translation, storage and processing in electronic systems.
  The publication of complete pages with unchanged content from this website and their integration into another website is only permitted with the written consent of theBVLpermitted. The duplication of texts, parts of text and images requires the prior consent of theBVL.

Canadian Food Inspection Agency

Food alerts starting from 2006. 3775 html pages, about 200 per year in recent years.

URL

- https://recalls-rappels.canada.ca/en/search/site?search_api_fulltext=&archived=1&f%5B0%5D=issue%3A20

T&C

- Commercial reproduction requires written permission. For non-commercial use, a link to the original resource is required.
- Full T&C https://www.canada.ca/en/transparency/terms.html
  Non-commercial reproduction
  Unless otherwise specified you may reproduce the materials in whole or in part for non-commercial purposes, and in any format, without charge or further permission, provided you do the following:
  exercise due diligence in ensuring the accuracy of the materials reproduced
  indicate both the complete title of the materials reproduced, as well as the author (where available)
  indicate that the reproduction is a copy of the version available at [URL where original document is available]
  Commercial reproduction
  Unless otherwise specified, you may not reproduce materials on this site, in whole or in part, for the purposes of commercial redistribution without prior written permission from the copyright administrator. To obtain permission to reproduce any content owned by the Government of Canada available on this site for commercial purposes, please contact the institution responsible for that content by referring to the institutions list available on the Government of Canada contacts page.
  Some of the content on this site may be subject to the copyright of another party. Where information has been produced or copyright is not held by the Government of Canada, the materials are protected under

the Copyright Act, and international agreements. Details concerning copyright ownership are indicated on the relevant page(s).

CDC (Centers for Disease Control and Prevention) YouTube channel

Videos, lectures, webinars, podcasts, about 6 per week, with 6100 videos already published since 2007.

URL

- https://www.youtube.com/@CDC

T&C

- We assume data is subject to https://www.cdc.gov/other/agencymaterials.html

  Most of the information on the CDC and ATSDR websites is not subject to copyright, is in the public domain, and may be freely used or reproduced without obtaining copyright permission.

  There are, however, a few exceptions. A federal government website may have a mix of public domain and copyright-protected materials. First, some resources, as well as images, on the CDC and ATSDR websites are restricted in their use because they were developed by government contractors or grantees, or have been licensed by a third party. Second, the U.S. government work designation does not apply to works of state and local governments; works of state and local governments may be protected by copyright. Third, copyright laws also differ internationally. While U.S. federal copyright laws may not protect U.S. government works outside the United States, the work may still be protected under the copyright laws of other countries when used in these jurisdictions. Copyright-protected materials featured on the CDC and ATSDR websites should include a copyright statement. However, if in doubt, please write to the contact point for that site.

  The following requirements must be followed to utilize CDC's public domain content:

  1) Attribution to the agency that developed the material must be provided in your use of the materials. Such attribution should clearly state the materials were developed by CDC ATSDR and/or HHS (e.g., "Source: CDC"; "Materials developed by CDC");

  2) You must utilize a disclaimer which clearly indicates that your use of the material, including any links to the materials on the CDC, ATSDR or HHS websites, does not imply endorsement by CDC, ATSDR, HHS or the United States Government of you, your company, product, facility, service or enterprise. All such disclaimers must be prominently and unambiguously displayed (e.g., "Reference to specific commercial products, manufacturers, companies, or trademarks does not constitute its endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention;

  3) You may not change the substantive content of the materials; and

  4) You must state that the material is otherwise available on the agency website for no charge.

  Note: Many of CDC's on-line publications are continually updated as the agency learns more about a specific disease or condition. Occasionally, sites that copy and re-post CDC materials fail to check for updates, which may result in out-of-date information being offered to users. For that reason, we urge you to link directly to our resource documents rather than re-posting. If you do re-post, please check back periodically to see if there are revisions.

  Linking to CDC, ATSDR or HHS content should open up a new browser window to our site/page. CDC content should not appear within the original window, framed by the existing site.

EFSA (European Food Safety Authority)

URL

- EFSA https://www.efsa.europa.eu/en/publications/food-risk-assess-europe
- OpenEFSA https://open.efsa.europa.eu/
  The single public interface for all information related to EFSA's scientific work. Follow the risk assessment process from receipt of the dossier to adoption of the opinion: status of assessments, dossier and studies (non-confidential versions), meetings agenda and minutes, info on experts, etc.
- EFSA Journal https://www.efsa.europa.eu/en/publications
  The scientific output of the European Food Safety Authority is published in the EFSA Journal, an open-access, online scientific journal. This concerns risk assessment in relation to food and feed and includes nutrition, animal health and welfare, plant health and plant protection.

T&C

- Requires attribution.
- Full T&C https://openscaie-dev.portal.azure-api.net/terms-and-conditions
  Unless otherwise indicated, the data and any related materials on the EFSA API Portal is in the public domain and made available with a Creative Commons Attribution 4.0 International (CC BY 4.0) license. You can copy, modify, distribute, reproduce and reuse the data, even for commercial purposes, without asking permission, provided that: EFSA is properly acknowledged as source and you provide a link to the license. You do not distort the original meaning of the data provided through the API Portal and you indicate when any modification is made.

Level of service

- i) Starter: Subscribers will be able to run 10 calls/minute up to a maximum of 200 calls/week; without access to the Questions REST (dismissed) API
- ii) Unlimited: Subscribers will be able to run 10 calls/minute up to a maximum of 500 calls/week. Administrator approval is required.

Format

- API: https://openapi-portal.efsa.europa.eu/docs/services/
- Catalogues REST API (there is also a SOAP version)
  This API allows retrieval of the catalogues published on Data Collection Framework.
  Catalogues (Harmonized controlled terminology) are a key element in the process of data validation and reporting. A harmonised terminology is used to collect and analyse data in a coherent way with the aim to support scientific research.
  CatalogueGroupList
  This operation allows downloading the list of catalogue groups defined in the system. The operation does not have any input parameters.
  The web service replies with a message containing the list of Groups in XML format.
  CatalogueList
  This operation allows downloading the list of catalogues, possibly restricted to a certain catalogue group or data collection.
  The web service replies with a message containing the list of catalogues in XML format.
  CatalogueFile

Export functionalities are supported by the method ExportCatalogueFile which differentiates the operation through the exportType parameter:

- Export catalogue
- Export the release note for a catalogue
- Export a hierarchy (deprecated)
- Export a group

- DataCollections REST API (there is also a SOAP version)
  This API allows retrieval of the configurations of data collections.
  Data collection is an important task of EFSA and a fundamental component of many of its risk assessment activities.
  DataCollectionList
  This method is used to retrieve the list of data collections defined in the EFSA Data Collection Framework. The operation does not have any input parameters. The web service replies with a message containing a list of data collections as a string in XML format.
  ResourceList
  This method is used to retrieve the list of resources related to a data collection. The parameter expected in this method is the Data Collection code.
  ResourceFile
  This method is used to retrieve a file resource stored in DCF and identified by a specific resourceId. With this method you can export the following resources:
  XML table definition (TABLE_METADATA prefix 01)
  Business rules file (BRS prefix 02)
  Transformation file (STX prefix 03)
  Validation schema (XSD prefix 04)
  XML file of the data collection configuration (DATA_COLLECTION prefix 05)
  Ack details (DATAILED_ACK_RES_ID prefix 06)
- Datasets API
  The dataset API gives access to all the data published by EFSA.
  The API returns the metadata in DCAT-AP standard format with the DOI link to download the dataset.
  A set of filters are available to retrieve a list of the datasets of interest (for instance the newly published), or the full list could be retrieved at once.
  search
  This API allows the retrieval of metadata of EFSA published Datasets. Metadata are expressed in DCAT_AP.
  In order to retrieve metadata it is possible to execute a query on the main fields like title (dcterms:title), abstract (dcterms:description), digital object identifier (dcterms:identifier), keywords (dcat:keyword), publication date (dcterms:issued), contact email (vcard:hasEmail).
- Deposits API
  The deposits API allows to retrieve all deposits (documents and data) published by EFSA.
  The API users can access from a unique standard interface all the EFSA depositions stored in different repositories.

The API returns the metadata either in Dublin Core standard (dc) or in EFSA specific format (Business Information Entity - BIE).

search (GET / POST)

This API allows to retrieve metadata of EFSA published documents.

Metadata can be expressed in Dublin Core (dc) or EFSA Business Information Entity (bie).

In order to retrieve metadata it is possible to execute a query on the following list of fields:

- BIE: title, type, abstract, content_doi, keywords, affiliation, publication_author, published_date, correspondence, language, publisher, publication_issn, journal_number, panel_members, question_number, adoption_date, acknowledgment, disclaimer_text, volume_number, issue_number
- Dublin Core: title, type, description, identifier, subject, rightsHolder, creator, issued, language, contributor, rights, relation, publisher, isRequiredBy, format

EFSA (European Food Safety Authority) YouTube channel

795 videos, lectures, webinars, podcasts, since 2012, about 5 or 6 per month

URL

- https://www.youtube.com/@EFSA_EU

T&C

- Requires attribution
- Full T&C https://www.efsa.europa.eu/en/legalnotice

EU Knowledge centre for food fraud and quality

A collection of News (2014 onwards, 86 articles) and publications (2005 onwards, 156 items).

News data is in various formats, sometimes pdf, sometimes pictures, sometimes it is a link to other institutional websites, video on VIMEO. Publications are pdf, with quite regular structure.

URL

- NEWS https://knowledge4policy.ec.europa.eu/search_en?f%5B0%5D=content_type%3Anews&f%5B1%5D=knowledge_service%3AFood%20Fraud%20and%20Quality
- PUBLICATION https://knowledge4policy.ec.europa.eu/search_en?f%5B0%5D=content_type%3Apublication&f%5B1%5D=knowledge_service%3AFood%20Fraud%20and%20Quality

T&C

- Requires attribution
- Data is © European Union, 1995-2023 licensed under CC BY 4.0.
- Full T&C https://commission.europa.eu/legal-notice_en#copyright-notice

FDA Enforcement Reports

Reports of product recalls. Setting product type “food” returns 24524 documents. Data is accessible via CSV or via API (https://www.accessdata.fda.gov/scripts/ires/apidocs/)

URL

- https://www.accessdata.fda.gov/scripts/ires/index.cfm#tabNav_advancedSearch

T&C

- All FDA websites publish their content with no copyright restrictions.
- Full T&C https://www.fda.gov/about-fda/about-website/website-policies#web
- Copyright owners are based in USA, so this licence covers database rights too.
  Unless otherwise noted, the contents of the FDA website (www.fda.gov) — both text and graphics — are not copyrighted. They are in the public domain and may be republished, reprinted and otherwise used freely by anyone without the need to obtain permission from FDA. Credit to the U.S. Food and Drug Administration as the source is appreciated but not required.
  People are also free to link to any URL on FDA's site. FDA's preference is that people link to the material on the FDA site (rather than copying it to their personal websites) because the agency continuously updates the information on the website as better information becomes available. A person copying documents to another website, instead of linking to them, would then have to monitor the original documents to know when these documents were updated by FDA or else risk giving bad or incorrect advice to visitors to their website. Providing consumers or health professionals with advice that is not fully up to date can lead to serious public health consequences. Providing industry advice that is not fully up to date can lead to companies being out of compliance with regulatory requirements.
  If a person, nonetheless, decides to copy content or images, FDA strongly recommends that the copied item lists the date that the material was copied and provides a link back to its source on the FDA website. Users can then see for themselves if the copied material has been updated or changed.
  FDA appreciates being informed about the use of website materials. Please email us at webmail@oc.fda.gov.

Schema

- Recalling Firm: The firm that initiates a recall.
- Classification:
- Class I: A situation with a reasonable probability of serious adverse health consequences or death from a violative product's use or exposure.
- Class II: A situation where use or exposure to a violative product may cause temporary or medically reversible adverse health consequences or the probability of serious adverse health consequences is remote.
- Class III: A situation where use or exposure to a violative product is not likely to cause adverse health consequences.
- On-Going: A recall currently in progress.
- Completed: A recall where the firm has retrieved and impounded all outstanding product or completed all product corrections.
- Terminated: A recall terminated when all reasonable efforts have been made to remove or correct the product according to the recall strategy.
- Distribution Pattern: General area of initial distribution (states, countries, territories). Subsequent distribution may not be included.
- Product Description: A brief description of the product.
- Code Information: List of lot/serial numbers, expiration dates, etc., on the product or its labelling.

- Reason for Recall: Information describing how the product is defective.
- Product Quantity: The amount of product subject to recall.
- Voluntary/Mandated: Indicates if the recall was initiated voluntarily by the firm or mandated by FDA.
- Recall Initiation Date: The date when the firm first notified the public or their consignees of the recall.
- Initial Firm Notification of Consignee or Public: The method(s) used by the firm for the initial recall notification.
- Recall Number: Alphanumeric designation assigned by FDA for tracking purposes.
- Event ID: Numerical designation assigned by FDA for tracking purposes.
- Center Classification Date: The date when FDA classified the recalled products as Class I, II, or III.
- Date Terminated: The date when FDA terminated the recall.
- Press Release URL(s): Link(s) to press release(s) published by FDA for the recall. Multiple press releases may be listed if applicable.

FDA Import Alerts

A list of recent import alerts, probably in the low thousands.

URL

- https://www.fda.gov/industry/actions-enforcement/import-alerts#list

T&C

- All FDA websites publish their content with no copyright restrictions (see FDA Enforcement Reports)

Schema

- Import Alert #: This is the number issued by the FDA. The first 2 numbers are the industry code of the product. For example, any import alert that starts with a 16 will be related to seafood.
- Published Date: This is the last date that there was an update to the alert. This is not the original date the alert was published.
- Type: This describes whether the alert is DWPE or DWPE with surveillance. Import Alerts that are DWPE with surveillance include additional guidance for the field. Such as, IA 20-05 states: Surveillance of heavy metal levels in fruit juices and fruit juice concentrates from all countries is warranted.
- Import Alert Name: This is the name of the alert; it is a brief description of what the alert applies to.
- Reason for Alert: This section describes why the alert was issued.
- Guidance: This section describes what actions the FDA may take and may provide guidance on how to be removed from the alert. This section can vary based on the type of alert.
- Product Description: This section describes what products are subject to DWPE.
- Charge: This section describes the FDA's laws and regulations applicable to the import alert.
- Countries: This section is included for country- or area-wide import alerts and includes the countries/areas subject to DWPE.
- List of firms and their products subject to Detention without Physical Examination (DWPE) under this Import Alert (a.k.a. Red List): This section lists the firms and/or products that are on the red list of the import alert. If a firm/product is on the red list of an import alert, it means they are subject to DWPE.
- List of firms and their products that have met the criteria for exclusion from Detention without Physical Examination (DWPE) under this Import Alert (a.k.a. Green List): This section lists the firms and/or

products that are on the green list of the import alert. If a firm/product are on the green list of an import alert it means they are not subject to DWPE.

FDA Import Refusals

About 440 thousands reports, downloadable by CSV

URL

- INFO https://www.fda.gov/industry/fda-import-process/import-refusals#listrefusals
- DATA https://www.accessdata.fda.gov/scripts/ImportRefusals/index.cfm

T&C

- All FDA websites publish their content with no copyright restrictions (see FDA Enforcement Reports)

Schema

- MANUFACTURER FEI - An identifier assigned internally by FDA for each firm/location.
- MANUFACTURER NAME - Identifies the name of the establishment declared as being responsible for the product refused.
- MANUFACTURER/ADDRESS/CITY/PROVINCE-STATE/COUNTRY - Identifies the manufacturer's street address, city, province or state, and country/area.
- PRODUCT CODE - A unique identifier assigned to products regulated by FDA.
- FDA PRODUCT DESCRIPTION - The FDA's description of the product offered for entry.
- REFUSAL DATE - Identifies the date when the action was taken.
- FDA DISTRICT - Identifies FDA District Offices that have jurisdiction over the refused product.
- ENTRY NO. - A unique identifier assigned to each entry.
- DOCUMENT/LINE/SUFFIX - A unique identifier for the product within an entry. An entry may have one or more of these number/letter identifiers.
- FDA SAMPLE ANALYSIS - Yes or No flag indicating whether or not a FDA sample analysis was conducted.
- FDA RECORD OF PRIVATE LAB SAMPLE ANALYSIS - Yes or No flag indicating whether or not FDA records show receipt of private laboratory analysis results package.
- CHARGES - Identifies the reason for the agency actions. The specific reason for the refusal can be accessed by clicking the reason given in the IRR or by searching under the file titled "Violation Code Translations".
- Partial Refusal - If this is present on a listing, it means that there was a reconditioning action which resulted in a portion of the shipment being refused.

FDA Inspections Citations

About 280000 entries from 2012 to 2023, expecting 10000 new entries per year. Download an XLSX file from the web interface (see “Download Dataset” button) or via API.

URL

- https://datadashboard.fda.gov/ora/cd/inspections.htm

T&C

- All FDA websites publish their content with no copyright restrictions (see FDA Enforcement Reports)

Schema

- FEI Number: FEI stands for Facility Establishment Identifier. It is a unique identifier assigned by the U.S. Food and Drug Administration (FDA) to track and monitor establishments involved in the manufacturing, processing, packaging, or holding of FDA-regulated products.
- Legal Name: The legal name refers to the official or formal name of a business or organization as recognized by the government or relevant authorities.
- City: The city refers to the specific urban or metropolitan area where the facility or establishment is located.
- State: The state refers to the specific region or subdivision within a country where the facility or establishment is situated. In the United States, it corresponds to one of the 50 states.
- Zip: Zip is short for ZIP Code, which is a numerical code used in the United States to identify specific geographic regions for efficient mail delivery.
- Country/Area: It represents the country or geographic area where the facility or establishment is located.
- Fiscal Year: Fiscal year refers to a 12-month financial reporting period used by businesses and organizations to calculate and report their financial performance. It may or may not align with the calendar year.
- Inspection ID: Inspection ID is a unique identifier assigned to an inspection conducted by regulatory agencies or authorities to assess the compliance of a facility with relevant rules, regulations, or standards.
- Posted Citations: Posted citations refer to the documented violations or non-compliance issues discovered during an inspection and made available for public view.
- Inspection End Date: The inspection end date indicates the date on which the inspection process was concluded.
- Classification: Classification refers to the categorization or grouping of a facility based on various factors, such as its purpose, industry, or regulatory requirements.
- Project Area: The project area refers to a specific section or division within the facility or establishment where a particular project or operation is carried out.
- Product Type: Product type refers to the category or classification of the goods or products manufactured, processed, or handled by the facility.
- Additional Details: Additional details may include any relevant information or specifics about the facility, its operations, certifications, or other notable attributes.

FDA Recalls

From 2017 to 2023 there are 662 entries with product type Food&Beverages. Excel export available.

URL

- https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts

T&C

- All FDA websites publish their content with no copyright restrictions (see FDA Enforcement Reports)

Schema

- Date: The specific date associated with the event or information mentioned.
- Brand-Names: The names or trademarks under which a particular product is marketed or sold by a company.

- Product-Description: A brief description or summary of the product involved in the mentioned context.
- Product-Types: The category or classification of the product based on its nature, purpose, or characteristics.
- Recall-Reason-Description: A description or explanation of the reason behind the product recall, highlighting the issue or concern that led to the recall action.
- Company-Name: The name of the company or manufacturer responsible for the production, distribution, or sale of the recalled product.
- Terminated Recall: Indicates whether the product recall has been terminated or completed. This may suggest that the recall process has reached its conclusion, and the necessary actions have been taken to address the issue.

Food Safety Authority of Ireland

490 alerts, 270 publications

URL

- Alerts https://www.fsai.ie/news-alerts
- Publications https://www.fsai.ie/publications

T&C

1. Use of data requires attribution.
2. Full T&C https://www.fsai.ie/getmedia/137e7e70-16e8-49ba-853a-9b099d93ebb5/Public-Sector-Information-Licence.pdf
3. Contains Irish Public Sector Information licensed by a Creative Commons Attribution 4.0 International (CC BY 4.0) license

Food Safety Dot Com

About 2570 documents dated 2000 onwards.

URL

- https://www.food-safety.com/topics/296-news

T&C

- Not available.

Food Safety News

About 8000 documents dated 2009 onwards.

URL

- Outbreaks: https://www.foodsafetynews.com/foodborne-illness-outbreaks/
- Recalls: https://www.foodsafetynews.com/food-recalls/

T&C

- Personal use only
- Copyright © 2023, Marler Clark, Inc., PS. All Rights Reserved.
  You are hereby granted a nonexclusive, nontransferable, limited license to view and use information retrieved from this website. The information is provided solely for your personal, informational, and non-commercial purposes on the condition that you do not remove or obscure the copyright notice or other notices. Except as expressly provided above, no part of this website, including but not limited to materials retrieved there from and the underlying code, may be reproduced, republished, copied, transmitted, or distributed in any form or by any means. In no event shall materials from this website be stored in any

information storage and retrieval system without prior written permission from Marler Clark, L.L.P., P.S.. Use, duplication, or disclosure by or for the United States Government is subject to the restrictions set forth in DFARS 252.227-7013 (c)1)(ii) and FAR 52.227-19.

Food Safety Tech

About 4000 articles and opinions.

URL

- https://foodsafetytech.com/

T&C

- Copyright Innovative Publishing Co., Inc. USA
  Copyright and trademarks:[SEP] Any written text, images, graphics, artwork, animations, videos, sounds and any other content of this website, including the arrangement thereof, are protected by copyright and other protective laws. Without our prior approval, no duplication, modification or usage of the content named above in other electronic or printed publications is permitted. Unless otherwise indicated, all trademarks are protected by trademark laws, including, but not limited to Innovative Publishing Co. LLC trade names, logos, emblems and name plates.The patents and trade names presented in this website are the intellectual property of the companies listed within.

FSIS USDA

1400 recall notices, with about 50 new every year.

URL

- https://www.fsis.usda.gov/recalls

T&C

- No restrictions.
- Full T&C https://www.usda.gov/policies-and-links
  Most information presented on the USDA Web site is considered public domain information. Public domain information may be freely distributed or copied, but use of appropriate byline/photo/image credits is requested. Attribution may be cited as follows: "U.S. Department of Agriculture."
  Some materials on the USDA Web site are protected by copyright, trademark, or patent, and/or are provided for personal use only. Such materials are used by USDA with permission, and USDA has made every attempt to identify and clearly label them. You may need to obtain permission from the copyright, trademark, or patent holder to acquire, use, reproduce, or distribute these materials.

Schema

- FSIS Announcement: recall details
- Product image: image of the product involved
- Company information: details of the company involved (name, contact, address)

LNV - Dutch Ministry of Agriculture, Nature and Food Quality

27 news in the last 20 years, mostly about animal disease.

URL

- https://english.nvwa.nl/news/news

T&C

- No restrictions, CC0
-

New Food Magazine

About 2600 news articles

URL

- https://www.newfoodmagazine.com/topic/food-safety/

T&C

- Non commercial use only.
- Copyright Russell Publishing Ltd, UK

  You may print off one copy, and may download extracts, of any page(s) from our site for your personal non-commercial use and you may draw the attention of others within your organisation to content posted on our site.

  You must not modify the paper or digital copies of any materials you have printed off or downloaded in any way, and you must not use any illustrations, photographs, video or audio sequences or any graphics separately from any accompanying text.

  Our status (and that of any identified licensor contributors) as the authors of content on our site must always be acknowledged.

  You must not use any part of the content on our site for commercial purposes without obtaining permission or a licence to do so from us. If you wish to do so, please contact us at admin@newfoodmagazine.com.

  If you print off, copy or download any part of our site in breach of these terms of use, your right to use our site will cease immediately and you must, at our option, return or destroy any copies of the materials you have made.

PubMed

Citations aggregator. A query for food hazard OR food safety retrieves 22k records.

URL

- https://pubmed.ncbi.nlm.nih.gov/

T&C

Full T&C https://www.nlm.nih.gov/web_policies.html#copyright

The aggregator displays content that may or may not be subject to copyright, no automatic way to tell. User Responsibility: It is your responsibility to determine and satisfy copyright or other use restrictions when using materials that are not in the public domain. NLM cannot guarantee the copyright status for any item.

Users who republish or redistribute the data (services, products or raw data) agree to: i)

maintain the most current version of all distributed data, or ii) make known in a clear and conspicuous manner that the products/services/applications do not reflect the most current/accurate data available from NLM.

RASFF Window

There are 15.242 records publicly available.

URL

- https://webgate.ec.europa.eu/rasff-window/screen/search

T&C

- Requires attribution.
- Full T&C https://data.europa.eu/en/copyright-notice

Science Direct

URL

- https://www.sciencedirect.com/

T&C

- https://www.elsevier.com/legal/elsevier-website-terms-and-conditions

  Copyright © 2023 Elsevier B.V. or its licensors or contributors. ScienceDirect® is a registered trademark of Elsevier B.V.

  - You may not copy, display, distribute, modify, publish, reproduce, store, transmit, post, translate or create other derivative works (including resulting from the use of artificial intelligence tools) from, or sell, rent or license all or any part of the Content, or products or services obtained from the Services, in any medium to anyone, except as otherwise expressly permitted under these Terms and Conditions, or any relevant license or subscription agreement or authorization by us. You may not use Content from the Services in combination with an artificial intelligence tool, (including to train an algorithm, test, process, analyse, generate output and/or develop any form of artificial intelligence tool).

  - You may not use any robots, spiders, crawlers or other automated downloading programs, algorithms or devices, or any similar or equivalent manual process, to: (i) continuously and automatically search, scrape, extract, deep link or index any Content; (ii) harvest personal information from the Services for purposes of sending unsolicited or unauthorized material; or (iii) cause disruption to the working of the Services or any other person's use of the Services. If the Services contain robot exclusion files or robot exclusion headers, you agree to honor them and not use any device, software or routine to bypass them. You may not attempt to gain unauthorized access to any portion or feature of the Services, any other systems or networks connected to the Services or to any Elsevier server, or any of the products or services provided on, accessed from or distributed through the Services. You may not probe, scan or test the vulnerability of the Services or any network connected to the Services or breach or attempt to breach the security or authentication measures on the Services or any network connected to the Services.

Scopus

URL

- https://www.scopus.com/

T&C

- https://beta.elsevier.com/legal/elsevier-website-terms-and-conditions?trial=true

  Commercial Users (Researchers in Private Sector & Commercial Institutions): APIs are available (for commercial use), with an API license and subscription, please contact us here to discuss your request

  - Non-Commercial Users (Researchers in Academic, Public Sector & Not-for-Profit Institutions): Most APIs (except SciVal and Embase APIs) are available for no charge, for non-commercial use, subject to Elsevier's policies and limits on usage

  - Copyright © 2023 Elsevier: The specific terms of use, copyright, and licensing agreements for accessing and using Scopus data would be outlined by Elsevier, the company that owns and operates Scopus. The copyright for the individual articles and papers indexed in Scopus typically belongs to the respective authors or publishers who hold the rights to those works.

TWEET-FID

A dataset consisting of selected tweets.

T&C

- https://aclanthology.org/2022.lrec-1.668.pdf

  The person in request ("the user") may receive and use TWEET-FID ("the dataset") only after accepting and agreeing to both the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy and the following terms and conditions: Commercial and academic use
  The dataset is made available for non-commercial purposes only. Any commercial use of this data is forbidden.

  Redistribution
  The user is not allowed to copy and distribute the dataset or parts of it to a third party without first obtaining permission from the creators.

  Publications
  The use of data for illustrative purposes in publications is allowed. Publications include both scientific papers and presentations for scientific/educational purposes.

  Citation
  All publications reporting on research using this dataset have to acknowledge this by citing the following article:
  Ruofan Hu, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng, Elke Rundensteiner, "TWEET-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks", in Submission at the 13th Conference on Language Resources and Evaluation (LREC 2022).

  For specific software output that is shared as part of this data, the user agrees to respect the individual software licenses and use the appropriate citations as mentioned in the documentation of the data.

  TWEET-FID changes
  The creators of this dataset are allowed to change these terms of use at any time. In this case, users will have to accept and agree to be bound by new terms to keep using the dataset

X (Twitter)

The service is provided at different tiers, and the only one that would be useful to EFRA is the Enterprise API tier

> *[…] which enables continued access to v1.1, v2 and additional Enterprise APIs. Pricing starts at $42,000 / Month based on usage and needs.*

UK Foods Standard Agency

From 2015 onwards, 192 allergy alerts and 152 food alerts

URL

- https://www.food.gov.uk/search?keywords=&filter_type%5BFood%20alert%5D=Food%20alert

T&C

- Requires attribution.
- Full T&C https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

You are free to: copy, publish, distribute and transmit the Information; adapt the Information; exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.

USDA YouTube channel

156 videos

URL

- https://www.youtube.com/user/usdafoodsafety

T&C

- Standard copyright, same terms and conditions of FSIS USDA