The Food Safety Market: An SME-powered industrial data platform to boost the competitiveness of European food certification

# D2.1- Data Models & Representations

| | |
|---|---|
| **DELIVERABLE NUMBER** | D 2.1 |
| **DELIVERABLE TITLE** | Data Models & Representations |
| **RESPONSIBLE AUTHOR** | Svetla Boytcheva |

Co-funded by the Horizon 2020
Framework Programme of the European Union

| GRANT AGREEMENT N. | 871703 |
|---|---|
| PROJECT ACRONYM | TheFSM |
| PROJECT FULL NAME | The Food Safety Market: An SME-powered industrial data platform to boost the competitiveness of European food certification |
| STARTING DATE (DUR.) | 01/02/2020 (36 months) |
| ENDING DATE | 31/01/2023 |
| PROJECT WEBSITE | www.foodsafetymarket.eu |
| COORDINATOR | Nikos Manouselis |
| ADDRESS | 110 Pentelis Str., Marousi, GR15126, Greece |
| REPLY TO | nikosm@agroknow.com |
| PHONE | +30 210 6897 905 |
| EU PROJECT OFFICER | Stefano Bertolo |
| WORKPACKAGE N. \| TITLE | WP 2 \| Data |
| WORKPACKAGE LEADER | Sirma AI |
| DELIVERABLE N. \| TITLE | D 2.1 \| Data Models & Representations |
| RESPONSIBLE AUTHOR | Svetla Boytcheva |
| REPLY TO | svetla.boytcheva@ontotext.com |
| DOCUMENT URL | |
| DATE OF DELIVERY (CONTRACTUAL) | 31 January 2022 (M24) |
| DATE OF DELIVERY (SUBMITTED) | 31 January 2022 (M24) |
| VERSION \| STATUS | 2.0 \| Final |
| NATURE | Other (OT) |
| DISSEMINATION LEVEL | Public (P) |
| AUTHORS (PARTNER) | Svetla Boytcheva (SIR), Plamen Tarkalanov (SIR), Nikola Tulechki (SIR), Pavlin Gyurov (SIR), Nikola Rusinov (SIR), Vladimir Alexiev (SIR), Martin Kotov (SIR) |
| CONTRIBUTORS | |
| REVIEWER | Danai Vergeti (UBITECH) |

| VERSION | MODIFICATION(S) | DATE | AUTHOR(S) |
|---|---|---|---|
| 0.1 | Table of contents | 15 November 2020 | Svetla Boytcheva (SIR) |
| 0.2 | Initial draft version | 15 December 2020 | Plamen Tarkalanov (SIR), Pavlin Gyurov (SIR), Nikola Rusinov (SIR) |
| 0.3 | Sections contribution, corrections, and observations | 21 January 2021 | Plamen Tarkalanov (SIR), Nikola Tulechki (SIR), Pavlin Gyurov (SIR), Nikola Rusinov (SIR), Martin Kotov (SIR) |
| 0.4 | Document refinement, addition of figures and descriptions | 28 January 2021 | Svetla Boytcheva (SIR) |
| 0.5 | Minor corrections | 29 January 2021 | Branimir Rakic (PROSPEH) |
| 0.6 | Minor corrections | 29 January 2021 | Danai Vergeti (UBITECH) |
| 1.0 | Final version | 30 January 2021 | Svetla Boytcheva (SIR) |
| 1.1 | Update of Section 3 and adding more ontologies | 10 January 2022 | Martin Kotov (SIR) |
| 1.2 | Updated Section 4 Semantic Data Integration | 19 January 2022 | Martin Kotov (SIR) |
| 1.7 | Updates of the other sections | 27 January 2022 | Martin Kotov (SIR) |
| 1.8 | Proof-reading and readability improvements - version for internal review | 28 January 2022 | Zlatina Marinova (SIR) |
| 1.9 | Internal review | 31 January 2022 | Danai Vergeti (UBITECH) |
| 2.0 | Final version | 31 January 2022 | Simeon Petrov (SIR) |

| PARTNERS | | CONTACT |
|---|---|---|
| Agroknow IKE<br>(Agroknow, Greece) | | Nikos Manouselis<br>(Agroknow)<br>nikosm@agroknow.com |
| SIRMA AI EAD<br>(SAI, Bulgaria) | | Zlatina Marinova<br>(SAI)<br>zlatina.marinova@ontotext.com |
| GIOUMPITEK MELETI SCHEDIASMOS YLOPOIISI KAI POLISI ERGON PLIROFORIKIS ETAIREIA PERIORISMENIS EFTHYNIS<br>(UBITECH, Greece) | | Danai Vergeti<br>(UBITECH)<br>vergetid@ubitech.eu |
| AGRIVI DOO ZA PROIZVODNJU, TRGOVINU I USLUGE<br>(Agrivi d.o.o., Croatia) | | Filip Gerin<br>(Agrivi d.o.o.)<br>filip.gerin@agrivi.com |
| PROSPEH, POSLOVNE STORITVE IN DIGITALNE RESITVE DOO<br>(PROSPEH DOO, Slovenia) | | Martina Poberaj<br>(PROSPEH DOO)<br>martina.poberaj@tracelabs.io |
| UNIVERSITAT WIEN<br>(UNIVIE, Austria) | | Tima Anwana<br>(UNIVIE)<br>tima.anwana@univie.ac.at |
| STICHTING WAGENINGEN RESEARCH<br>(WFSR, Netherlands) | | Yamine Bouzembrak<br>(WFSR)<br>yamine.bouzembrak@wur.nl |
| TUV- AUSTRIA ELLAS MONOPROSOPI ETAIREIA PERIORISMENIS EUTHYNIS<br>(TUV AU HELLAS, Greece) | | Kostas Mavropoulos<br>(TUV AU HELLAS)<br>konstantinos.mavropoulos@tuv.at |
| TUV AUSTRIA ROMANIA SRL<br>(TUV AU ROMANIA, Romania) | | George Gheorghiou<br>(TUV AU Romania)<br>george.gheorghiu@tuv.at |
| VALORITALIA SOCIETA PER LA CERTIFICAZIONE DELLE QUALITA'E DELLE PRODUZIONI VITIVINICOLE ITALIANE SRL<br>(VALORITALIA, Italy) | | Francesca Romero<br>(Valoritalia)<br>francesca.romero@valoritalia.it |
| TUV AUSTRIA CYPRUS<br>(TUV AU CYPRUS, Cyprus) | | Sousanna Charalambidou<br>(TUV AU CYPRUS)<br>sousanna.charalambidou@tuv.at |

## ACRONYMS LIST

| | |
|---|---|
| API | Application Programming Interface |
| GTIN | Global Trade Item Number |
| JSON | JavaScript Object Notation |
| RDF | Resource Description Framework |
| RDBMS | Relational Data Base Management System |
| REST | Representational state transfer |
| SHACL | Shapes Constraint Language |
| SKOS | Simple Knowledge Organization System Primer |
| TheFSM | The Food Safety Market |
| XML | Extensible Markup Language |
| TTL | Turtle |
| URI | Uniform Resource Identifier |

# EXECUTIVE SUMMARY

This document presents the updated version of the data models developed until M24 of The Food Safety Market (TheFSM) project and their machine-readable representations. In the document we briefly present data types, data sources, and data formats. Special attention is paid to the domain-specific ontologies, standard classifications, and vocabularies. The result data models are based on the desk research and analysis, based on data inventory of publicly available datasets, as well as dataset samples provided by TheFSM consortium. The detailed analysis of data requirements of all use cases and certification schemes serves as a basis for TheFSM integrated data model. Developed machine-readable data models are in compliance with the industry-accepted standards in order to ensure interoperability of theFSM platform. This document will be further updated at M36 to describe the final data model and its extensions in terms of data interoperability and coverage.

# TABLE OF CONTENTS

# LIST OF TABLES



# LIST OF FIGURES

# 1 INTRODUCTION

## 1.1 Scope

This document describes the current of The FSM data models at the point of completion of MS2 "2nd Release of TheFSM Data Platform and Applications" (M24). It builds upon the first iteration of the data model design, presented in version 1 of D2.2 in M12, for which we analyzed data from data inventory and conducted desk research of available domain-related standards and ontologies. The data inventory contains both data from public resources, data samples provided by TheFSM project consortium and improved data models based on additional data sources, APIs and more domain-specific ontologies. A final version of this deliverable will be submitted at the end of the project, when MS8 "Final Release of TheFSM Data Platform and Applications" is completed (in M36).

## 1.2 Data

The following types of data were considered as relevant – grouped by availability (public/proprietary) and information organization in different data formats (structured/unstructured) with some examples.

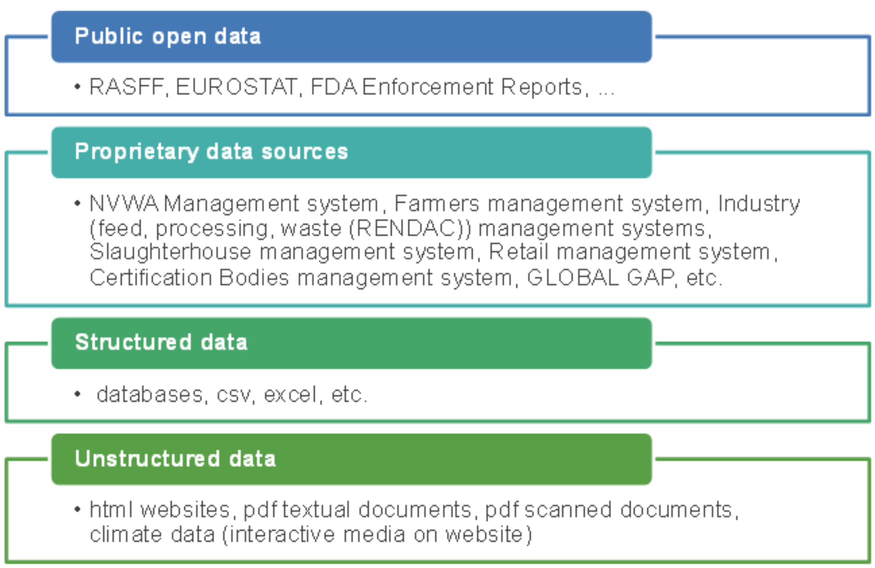


Figure 1: Types of datasets

When we use public open data we need to identify the scope and the particular subsets that are relevant to the project. For example, LangualL contains foods, beverages, ingredients, drinks which are relevant to the project, but it also contains other information for counties, how food should be stored or cooked which is not useful in the context of the project.

For proprietary data we need data samples to be shared in order to analyze the data structure and to envision the necessary data services that will be developed for such data processing.

For unstructured data we need strong justification of the particular data source and lack of any other option to gather the essential data from other data providers in a structured format.

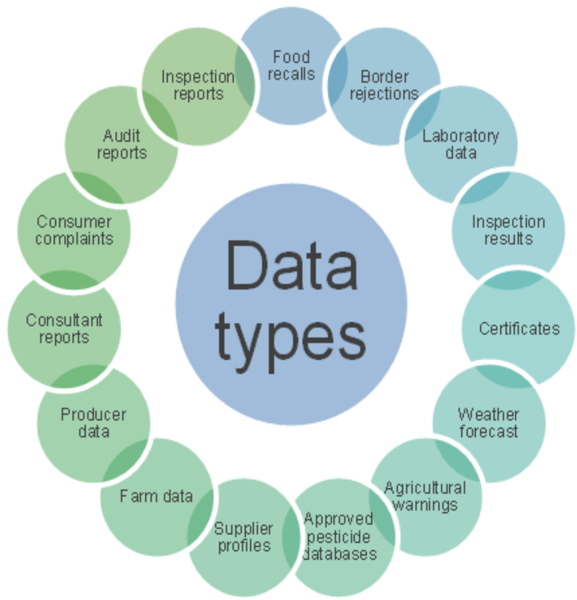


Figure 2: Data types

## 1.3 Relation to other deliverables

The data model developed is based on the detailed analysis of business, data and technical requirements of all use cases described in D1.1 “Report on Requirements for TheFSM” (delivered in M4), in compliance with D2.2 : “Data Services” (delivered in M12), D2.3 “Report on Data Population” (delivered in M12) and D3.1 “TheFSM Open Reference Architecture” (delivered in M12). The best practices, state-of-the-art and standards are taken in consideration as well.

## 1.4 Outputs of the deliverable

This deliverable provides description of the outcomes of T2.2 - the initial version of the data model. The current version of machine-readable representation of the developed data model is available at Github repository.

The Github repository contains definition of theFSM data model in various formats and the supplementary documentation. Some screenshots of this repository are provided below.

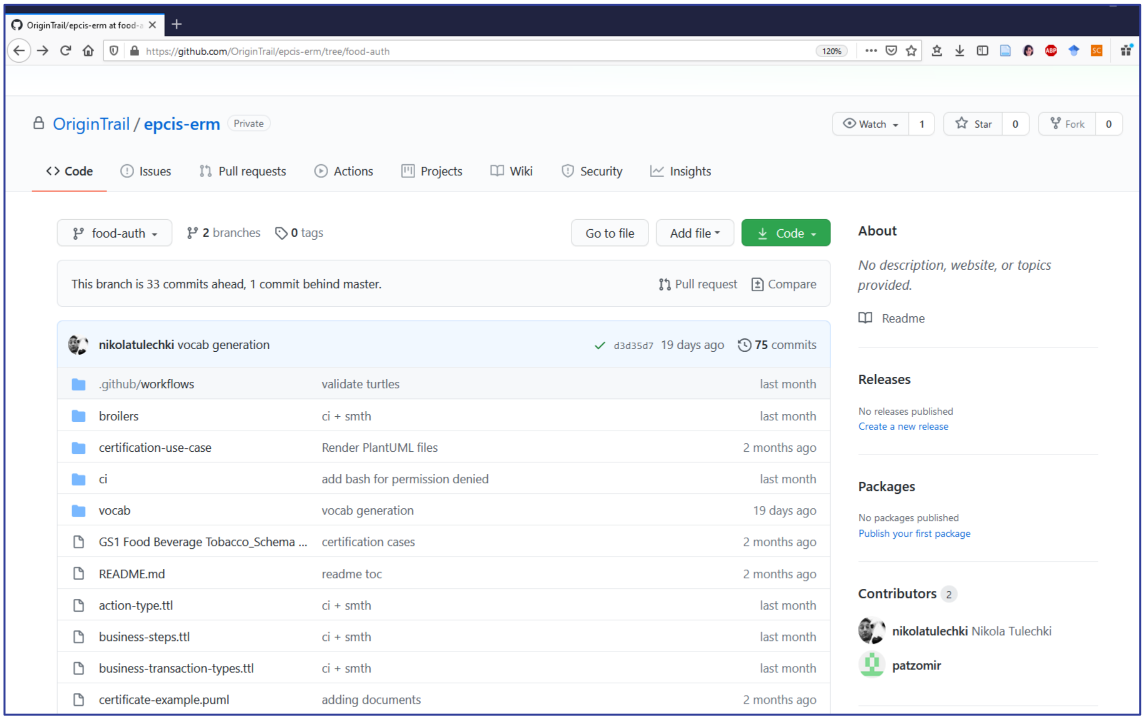


**Figure 3: Github repository – food-auth branch**

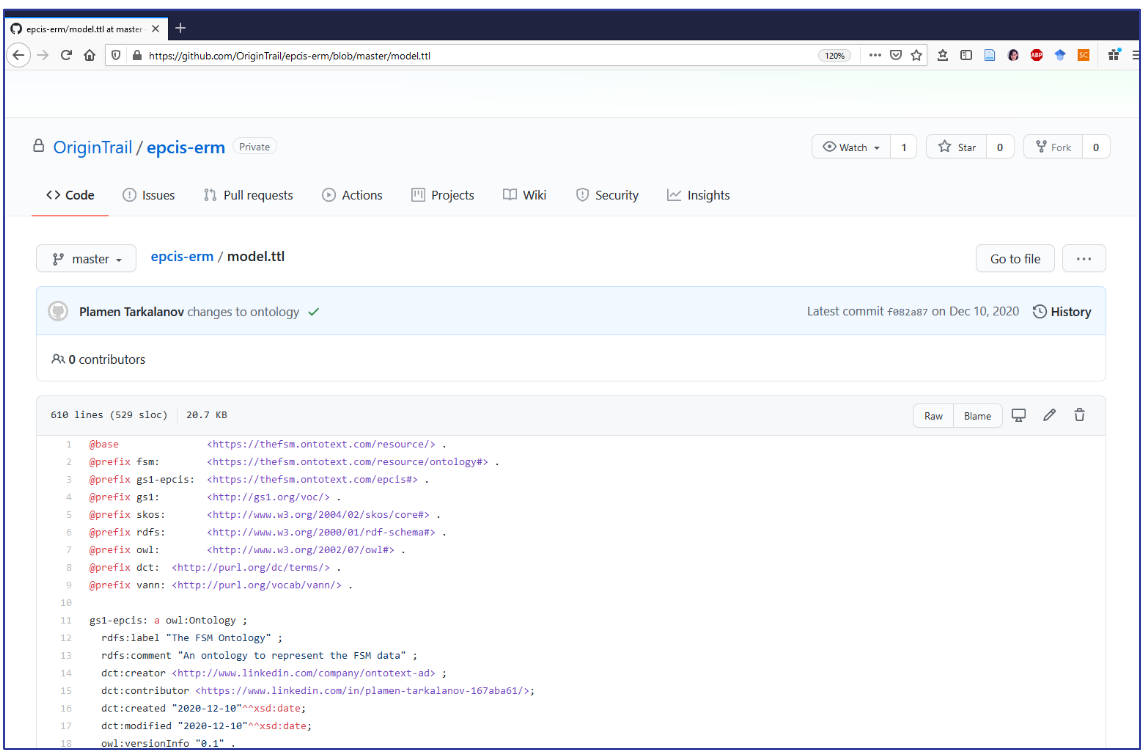


**Figure 4: Github repository – epcis-erm/model.ttl**

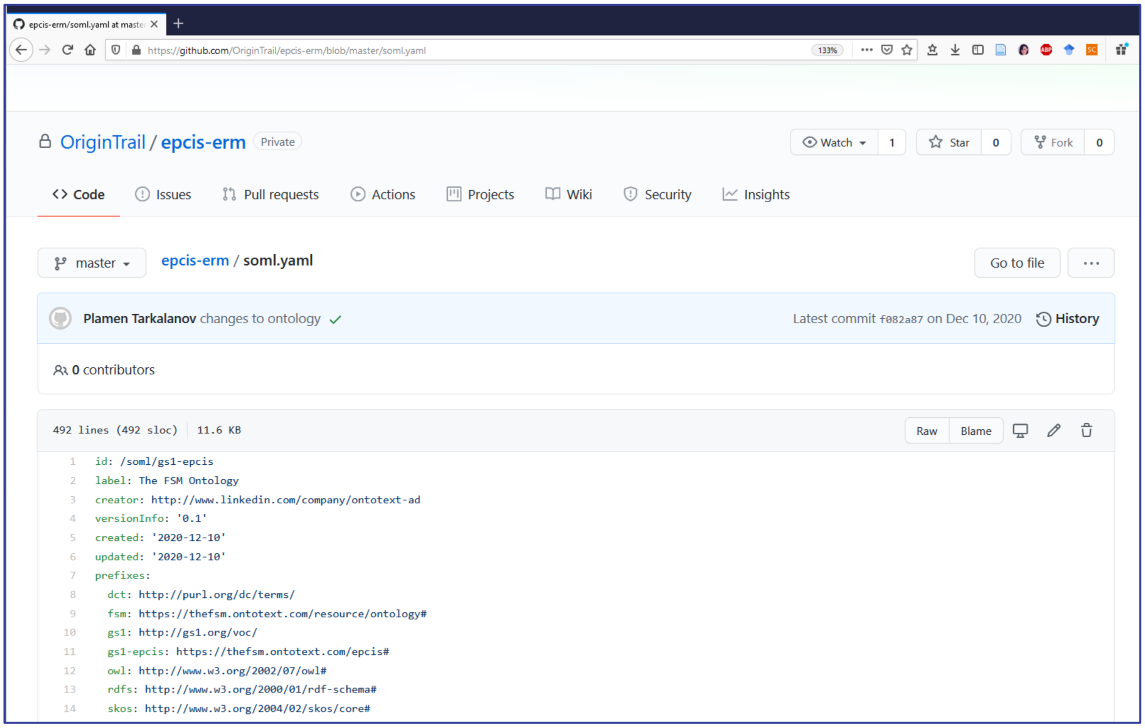


**Figure 5: Github repository – epcis-erm/soml.yaml**

# 2 PUBLIC DATASETS AND ONTOLOGIES

In this section we briefly outline the public datasets and ontologies that are relevant to the FSM data model. We start with an overview of many thesauri and ontologies, relevant to food and food certification. Then we present in more detail the ones that were selected as most appropriate for standardisation of FSM data representations.

## 2.1 Overview of domain-specific standards and ontologies

Domain-specific ontologies are a special focus in data modeling as they provide common vocabularies that allow for uniform data representation and provide machine processable meaning. Some of the popular vocabularies in the food and food certification domains are listed in the table below.

**Table 1: Domain-specific Standards and Ontologies**

| | |
|---|---|
| ***AgroVoc***[1] | A controlled vocabulary for describing food, nutrition, agricultural, marine, forestry, environmental information. It is also part of the GACS initiative, which aims to map the core concepts of three major thesauri AgroVoc, CAB, and NAL. |
| ***CAB Thesaurus***[2] | CAB Thesaurus is the essential search tool for all users of CAB Abstracts and Global Health databases and related products. It provides a controlled vocabulary resource of over 2.9 million descriptive terms, for the applied life sciences including: agriculture, forestry, horticulture, soil science, entomology, mycology, parasitology, veterinary medicine, nutrition, rural studies. |
| ***NAL Thesaurus***[3] | The National Agricultural Library's Agricultural Thesaurus is now available as Linked Open Data. The Thesaurus and Glossary are online vocabulary tools of agricultural terms in English and Spanish and are cooperatively produced by the National Agricultural Library[4], USDA[5], and the Inter-American Institute for Cooperation on Agriculture, as well as other Latin American agricultural institutions belonging to the Agriculture Information and Documentation Service of the Americas[6] (SIDALC). NAL |

1 http://aims.fao.org/vest-registry/vocabularies/agrovoc
2 https://www.cabi.org/cabthesaurus/
3 https://agclass.nal.usda.gov/
4 http://www.nal.usda.gov/
5 http://www.usda.gov/
6 https://www.iica.int/en

| | |
|---|---|
| | Thesaurus is used to select controlled vocabulary terms for subject indexing of AGRICOLA, PubAg and other databases. |
| ***ICASA*** | Data format for documenting experiments and modelling crop growth and development, facilitating exchange of information and software.<br>• The ICASA V2.0 can be found in this paper[7]<br>• The current update version (ICASA V3.0) can be requested from the DSSAT development team[8]: |
| ***Food Safety Knowledge Markup Language[9] (FSK-ML)*** | In development with the participation of BfR, aiming to describe data and models relevant to risk assessment tasks. With the Food Safety Knowledge Markup Language (FSK-ML) we now extend the PMF-ML format to enable the exchange of knowledge / information that is embedded in specific script-based programming languages (e.g., "R", Matlab, Python). The FSK-ML guidance document primarily aims at harmonizing the exchange of food safety knowledge (e.g. predictive models) including the associated metadata where this knowledge is only available in a software dependent format.<br>• In the provided link, there are tutorials and a developer guide pdf document[10]<br>• One can also read about the FSK-Lab - an open source food safety model integration tool[11] - |
| ***FoodEX2[12]*** | The European Food Safety Authority (EFSA) has developed a standardized food classification and description system called FoodEx2. EFSA has established a collaboration for the use of FoodeEx2 with several institutions. Among these, the Food and Agriculture Organisation (FAO) of the United Nations and the Friedman School of Nutrition Science and Policy at Tufts University.<br>• In the provided link, there are YouTube tutorials for the software. |

[7] https://dssat.net/wp-content/uploads/2014/02/White2013ICASA_V2_standards.pdf
[8]https://dssat.net/data/standards_v2/#:~:text=IBSNAT%20Data%20Standards&text=The%20ICASA%20standards%20have%20evolved,original%20plain%20text%20file%20format
[9] https://foodrisklabs.bfr.bund.de/fsk-ml-food-safety-knowledge-markup-language/
[10] https://foodrisklabs.bfr.bund.de/wp-content/uploads/fsk/FSK_guidance_document_V3_1.pdf.
[11] https://foodrisklabs.bfr.bund.de/fsk-lab/.
[12] https://www.efsa.europa.eu/en/data/data-standardisation

| | |
|---|---|
| ***GeoNames***[13] | Should be used as an international standard for country codes. |
| ***eCl@ss Classification and Product Description***[14] | An ISO/IEC-compliant industry reference-data standard for the classification and unambiguous description of products and services. The classification establishes a uniform semantic standard, enabling Internet of Things and product master data to be exchanged digitally across all borders – across sectors, countries, languages and organizations. |
| ***XACML OASIS Standard*** | In TheFSM, access to data assets is regulated through Attribute-Based Access Control (ABAC) policies, based on the XACML OASIS standard, that allows the data providers to protect and share their data assets, even when they do not have any prior knowledge of the potential individual data consumers in the food certification data value chain. XACML promotes common terminology and interoperability between access control implementations by multiple vendors. |
| ***GS1***[15] | GS1 standards provide a common language to identify, capture and share supply chain data– ensuring important information is accessible, accurate and easy to understand. GS1 Global Data model[16] plays important role in interoperability.<br>GS1 vocabulary has many props related to food[17] and perishable goods. Another benefit of using GS1 is the Global Location Number[18] (GLN) is used to identify locations and legal entities. This unique identifier is comprised of a GS1 Company Prefix, Location Reference, and Check Digit. Another specification provided by GS1 is Global Trade Item Number[19] (GTIN) – it can be used by a company to uniquely identify all of its trade items. |

13 https://www.geonames.org/
14 https://www.eclass.eu/fileadmin/downloads/eCl%40ss_company_brochure.pdf
15 https://www.gs1.org/
16 https://www.gs1.org/standards/gs1-global-data-model
17 https://www.gs1.org/industries/retail/fresh-foods
18 https://www.gs1us.org/DesktopModules/Bring2mind/DMX/Download.aspx?Command=Core_Download&EntryId=158&language=en-US&PortalId=0&TabId=134#:~:text=The%20Global%20Location%20Number%20(GLN)%20is%20used%20to%20identify%20locations,is%20a%20GLN%20used%20for%3F
19 https://www.gs1.org/standards/id-keys/gtin

| | |
|---|---|
| ***EPCIS***[20] | Electronic Product Code Information Services (EPCIS) is a global GS1 Standard for creating and sharing visibility event data. The goal of EPCIS is to enable disparate applications to leverage Electronic Product Code (EPC) data via EPC-related data sharing, both within and across enterprises. Ultimately, this sharing is aimed at enabling participants in the EPC[21] global Network to gain a shared view of the disposition of EPC-bearing objects within a relevant business context. |

## 2.2 Metadata

In addition to the data collection activities, TheFSM will also generate its own valuable data assets in terms of metadata that will improve the description, interlinking, normalization, unification, and quality assessment of the collected datasets. The use of W3C standards such as PROV-O for provenance and DCAT for data catalog description will be encouraged.

- PROV-O[22]: The PROV Ontology -The PROV Ontology is expected to be both directly usable in applications as well as serve as a *reference model* for creating domain-specific provenance ontologies and thereby facilitates interoperable provenance modeling.
- Data Catalog Vocabulary (DCAT) - Version 2[23] - DCAT is a vocabulary for publishing data catalogs on the Web, which was originally developed in the context of government data catalogs such as data.gov and data.gov.uk, but it is also applicable and has been used in other contexts.

## 2.3 SKOS

Simple Knowledge Organization System[24] (SKOS) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary and will be used to represent relevant entities that are not covered by other schemas such as GS1 and EPCIS. The ontology generation pipeline described in D2.2 also covers creating small and medium Certificate and certificate families

## 2.4 W3C Cube ontology

In order to express statistical data such as laboratory measurements, meteorological data, or any other dataset resulting from a systematized observation of project-relevant entities, we require a proper representation of data. Whenever necessary, data that is commonly expressed (and exchanged) as spreadsheets will be transformed into a homogeneous semantic model and

[20] https://www.gs1.org/epcis/epcis/1-1
[21] https://www.gs1.org/sites/default/files/docs/epc/EPCIS-Conformance_Requirements.pdf
[22] https://www.w3.org/TR/prov-o/#introduction
[23] https://www.w3.org/TR/vocab-dcat/#introduction
[24] https://www.w3.org/2004/02/skos/

described using the **W3C Cube** ontology[25]. This process will ensure high interoperability of observation data as well as easier maintenance and backward compatibility of novel datasets.

The W3C CUBE ontology (QB) captures multidimensional observations (data cubes) using the following terminology (in bold). We map roughly these QB terms to the data items discussed in previous sections and expressed using tGS1 and EPCIS. The first 3 are called "components".

- **Dimension**: entity, quality, method
- **Attribute**: unit, context
- **Measure**: value
- **Observation** = Dimensions + Attributes + Measures

QB defines what components are expected in a specific ***qb:DataSet*** by using a ***qb:DataStructureDefinition***. QB provides some flexibility that affords data efficiencies, and avoiding combinatorial explosion:

- QB allows using several dimensions per observation, without tying them up together. E.g., you can use 3 dimensions **entity="plant", quality="height", method="whole height"**
- We could also use several measures per observation (e.g. as taken by a combination instrument), although this is less commonly used.
- One could split a dataset into Slices (or other kinds of ObservationGroups) by fixing some of the dimensions, so one doesn't need to repeat them for every observation.

[25] https://www.w3.org/TR/vocab-data-cube/

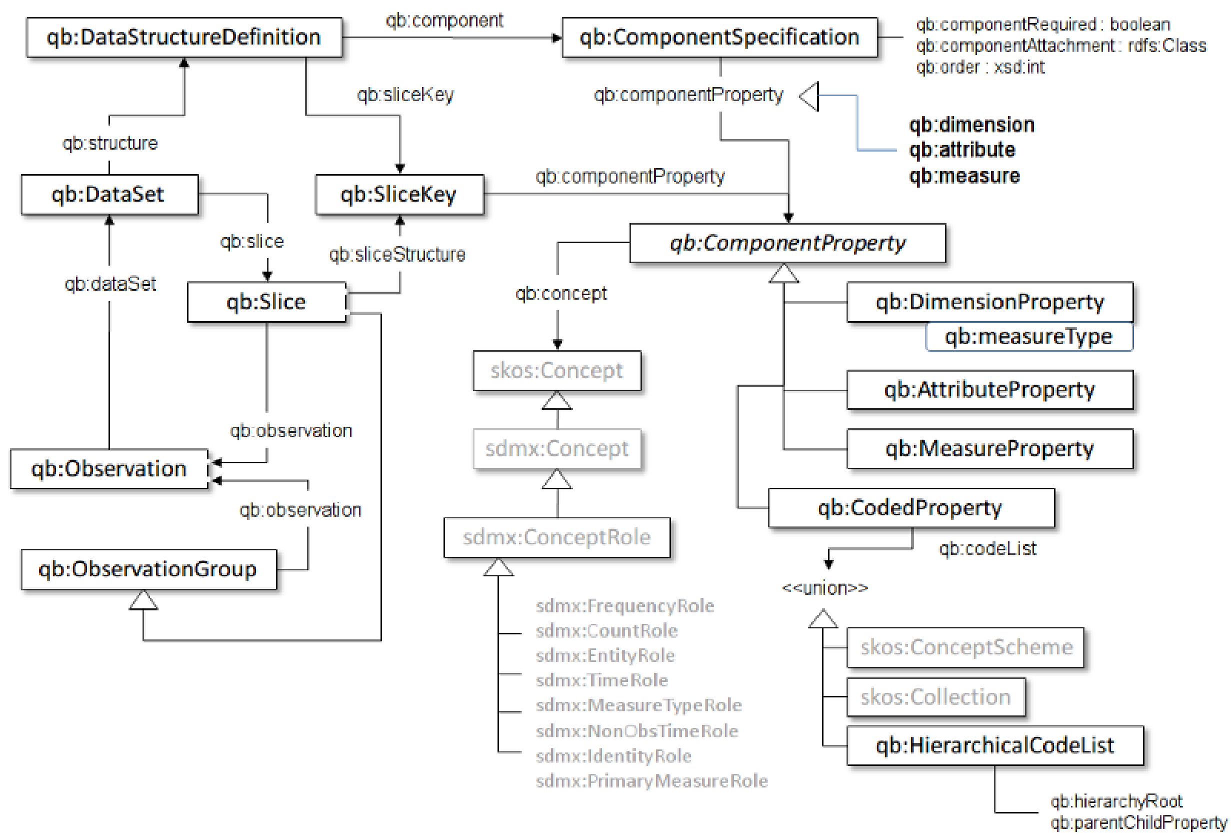


**Figure 6: W3C CUBE Ontology**

Adapting the model to the project-specific entities will be done by defining a large number of classes. In order to both streamline this process and ensure that it can be done collaboratively by domain experts and in a fully decentralized fashion, we employ a technique using shared Google sheets and automatic ontology generation by converting the data from the sheets into the final model. This setup is described in D2.2.

## 2.5 Standardisation and interoperability of TheFSM data

Interoperability of food safety data in FSM is ensured by mapping main data types to ontologies that are compliant with a wide set of standards. In the following subsections we will briefly present the ontologies to which data is mapped at the moment (M24). Further datatypes will be processed in the final year of the project and this section will be updated accordingly.

### 2.5.1 LanguaL

As a basis for the food data interoperability, we map data to LanguaL[26]. LanguaL™ is a framework for description of foods and related terms, including food ingredients, cooking methods, geographic places and regions and other facets for classification. LanguaL™ facilitates links to

[26] https://www.langual.org/

many different food data banks and thus contributes to coherent data exchange. LangualL data contains links to the following standards or datasets: EC, EFSA FOODEX2[27], EUROCODE2[28], EFG, CCPR, GS1 GPC[29], USDA SR[30]. In addition, LanguaL is a multilingual thesaurus where each descriptor is identified by a unique code pointing to equivalent terms in different languages (e.g. Czech, Danish, English, French, German, Italian, Portuguese, Spanish and Hungarian). This support for multilingual food terms was also an important argument for mapping TheFSM data to LanguaL.

We used the Food Product Indexer tool for browsing the LanguaL thesaurus. Next, a screenshot of the tools is provided.

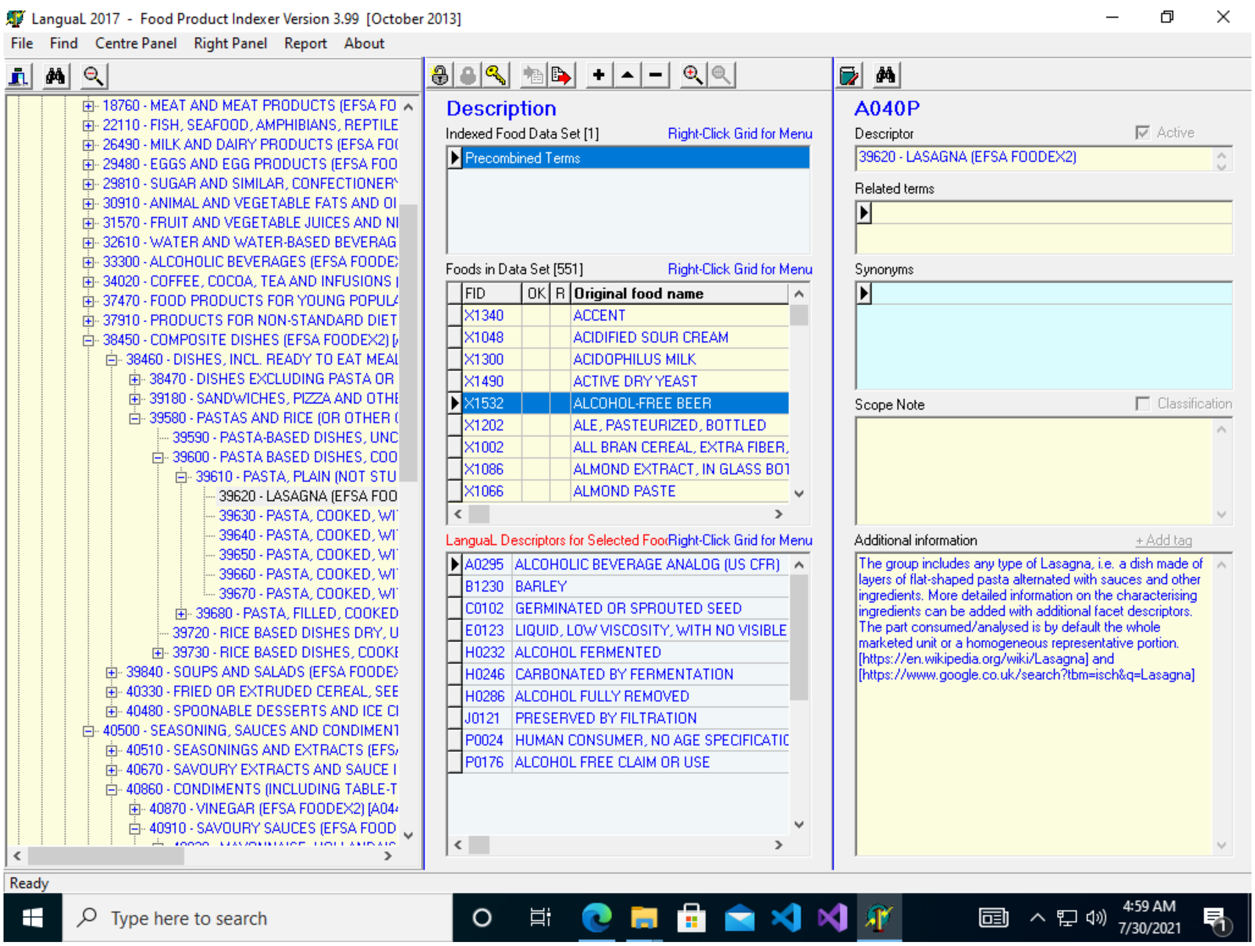


**Figure 7: LanguaL food product classification**

[27] https://www.efsa.europa.eu/en/data/data-standardisation
[28] http://www.danfood.info/eurocode/
[29] https://www.gs1.org/standards/gpc
[30] https://fdc.nal.usda.gov/

In order to use LanguaL as a vocabulary to map TheFSM data to, we first extracted the data with the help of the **mdbtool**[31]. Next we have removed all unnecessary data like how to cook your food or how to store it. After that we've extracted the food labels, codes and standards from descriptions, so we can transform it into RDF format and then use it in our reconciliation service.

### 2.5.2 USDA

We have also considered using USDA Open Data[32] as an extension of the LanguaL taxonomy. Such an extension is easily enabled using a mapping file provided in the LanguaL website. For version 2, LanguaL has provided sufficient coverage of food related data values. In the final year, based on the actual data processed by the FSM platform, we will decide whether the USDA extension shall be added to the taxonomy.

### 2.5.3 Wikidata

Wikidata[33] is a free and open knowledge base, containing 96,764,844 data items represented in a way that is understandable both by humans and by machines. Wikidata contains structured data representation of the information in Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

In the context of TheFSM we are using Wikidata data provided in RDF format[34], the so called *truthy* dataset, to semantically enrich data entires of companies and locations. For the purpose of semantic enrichment we are constructing TTLs for each domain and creating ElasticSearch[35] connectors which are used by the reconciliation service (described in D2.2). In the index we include links to other public data sets like Geonames for locations and the Virtual International Authority File[36] (VIAF) identifiers for companies.

### 2.5.4 PubMed and MESH

PubMed[37] is a free digital repository that archives open access full-text scholarly articles published in biomedical and life sciences journals. As one of the major research databases developed by the National Center for Biotechnology Information, PubMed Central is more than a document repository. The Medical Subject Headings[38] (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH includes the subject headings appearing in MEDLINE/PubMed, the NLM Catalog, and other NLM databases. In

31 https://github.com/mdbtools/mdbtools
32 https://www.usda.gov/content/usda-open-data-catalog
33 https://www.wikidata.org/wiki/Wikidata:Main_Page
34 https://www.wikidata.org/wiki/Wikidata:Database_download/en
35 https://www.elastic.co/
36 http://viaf.org/
37 https://pubmed.ncbi.nlm.nih.gov/
38 https://www.ncbi.nlm.nih.gov/mesh/

the context of TheFSM, we will use MESH as a thesaurus to which unwanted substances, activities and additives that appear as data values will be mapped.

#### 2.5.5 GS1 GPC

GS1 standards provide a common language to identify, capture and share supply chain data– ensuring important information is accessible, accurate and easy to understand. Thus, the GS1 Global Data model plays an important role in data interoperability in the FSM. GS1 vocabulary has many props related to food and perishable goods. Another benefit of using GS1 is the Global Location Number (GLN) is used to identify locations and legal entities. This unique identifier is comprised of a GS1 Company Prefix, Location Reference, and Check Digit. Another specification provided by GS1 is Global Trade Item Number (GTIN) – it can be used by a company to uniquely identify all of its trade items.

We are using the standard as a basis for the food taxonomy service, which allows users of the FSM Platform to tag datasets they upload with information about what type of food-related data these datasets contain. We have made use of the GS1 GPC hierarchical structure and we have converted it to RDF by using the SKOS schema.

#### 2.5.6 FSCC 22000

The Food Safety System Certification 22000 (FSSC 22000) FSSC is a global non-profit and independent scheme owner that provides trust and delivers impact to the consumer goods industry. FSSC 22000 has significant impact on global food safety for over 12 years through supporting the consumer goods industry in implementing ISO-based management systems, reaching its objectives and the Sustainable Development Goals.

We use the FSCC 22000 scheme to standardise FSM food certification data in the semantic model (see section 4.3.1 for more details).

# 3 SEMANTIC DATA INTEGRATION

Semantic Data Integration is one of the best ways to integrate diverse data across organizations and enterprises, and to exploit available datasets both as commercial data and in the LOD cloud. Initially adopted by Life Science and Biology researchers, semantic web techniques have by now found a wide following, notably in the Agro sector, which in many cases leverages ontologies developed in the Bio community.

## 3.1 Semantic integration process

Semantic Data Integration aims to harmonize data from heterogeneous sources often from various providers and convert it to a semantic form. Semantic Data Integration employs a data-centric architecture built upon a standardized model for data publishing and interchange, namely the Resource Description Framework (RDF).

In the RDF framework heterogeneous data - structured, semi-structured and unstructured, is expressed, stored and accessed in the same manner. This is made possible because the data structure is expressed through the links within the data itself instead of being constrained to a structure imposed by the database. Alleviating this constraint has an additional benefit - removing the costs and complications that come with replacing a database solution that has become obsolete due to the evolution of data. As changes in the data structure occur they are reflected in the database through changes in the links within the data.

Furthermore, RDF not only provides a simple and standard way to express all data, but it is also the backbone of Semantic Technologies. This enables the inference of new facts from the data, as well as the enrichment of the available knowledge base by accessing Linked Open Data (LOD) resources.

A substantial part of the semantic data integration is to match instances about the same entity coming from different databases (aka coreferencing). The final result is an integrated Knowledge Graph of data in a specific domain. The following steps are carried out as part of the process of semantic integration:

- Sampling tabular and/or RDF data from project partners
- Sample data analysis
- Listing data requirements and competence questions
- State of the art research. listing  existing ontologies and other related ontologies
- Selection, combination and extension of ontologies (Ontology engineering)
- Develop semantic models that show precisely how RDF data comes together. Long experience shows that while ontologies describe the terms (classes and properties) that can be used to describe a domain of interest, they do not describe how these terms come together to represent a particular data scenario or situation.
- Data cleaning and normalization
- Data enrichment

- Creation of semantic conversion scripts.

## 3.2 Semantic Integration within WP2 of TheFSM

In this section we summarise the specific semantic integration work carried out in WP2 by M24 of the FSM project, inline with the generic steps above.

We have collected and analysed the following data types for which samples were provided by the partners:

- Food taxonomy – a CSV file that contains product codes, labels and codes of product category
- Hazard taxonomy – a CSV file that contains hazard codes, labels and codes of hazard category
- Suppliers – a list of company names from FOODAKAI. Includes many duplicates and similar values. This is one of the major challenges to data integration since the risk events (incidents) processed may use different names for the same company. There is no identifier that we can rely on to link internal data (incidents) or to link to other datasets, so mapping to a public dataset with company names can help solve this issue.
- Country taxonomy - tabular data that contains the identifier of a country, name and codes for the continents.
- Incidents - tabular data with information about products and product categories, who has reported the problem, what hazards have been found in the food, who is the supplier(company), country of origin and from where the alert was received.

First, we have started with the foods provided by AGROKNOW in tabular format. We have searched for a proper dataset which has the right food standards and covers the data provided by the partner. Based on that we have choosen LanguaL (see section 3.5.1 for more details) which is compliant with several standards and covers the necessary information. We have then cleaned up the dataset and converted it into RDF by using OntoRefine. And finally, we have created an ElasticSearch index which is used by the reconciliation service for the semantic data integration.

Once we have finished with the foods we have continued with the countries. For them we have decided to use Wikidata (see section 3.5.3) due to its large georgraphic coverage, links to other datasets and ISO standard codes. Here we have downloaded the RDF version of Wikidata, loaded it in GraphDB and created an ElasticSeach index over the countries with the necessary properties.

Next, we moved on to the food incidents which includes data about foods, hazards and countries. We analysed the samples provided and reached the conclusion that the data is very consistent and there is no need for cleaning up or any other action. For this data we use the previously created reconciliation services.

For the food taxonomy service used in the FSM platform we have used GS1 GPC (see section 3.5.5) and its hierarchy. We have downloaded only the foods and beverages part of GS1 GPC, converted it to RDF by using OntoRefine and appling SKOS ontology.

Then we continued by processing the Food Safety Certificates data(FSSC), which contains information about companies with their address and information if they are certified to sell or manufacture food. First, we have created a semantic model by mapping the tabular data to SKOS ontology, EPCIS 2.0 standard and SCHEMA (model can be seen in the next section). We are planning this to be used in the next steps in the semantic integration service.

Finally, we have created enrichment services for other types of data samples, such as hazards and suppliers(companies). For hazards we currently use data from MeSH and PubMED (see section 3.5.4), but based on the outcome and partner feedback, other public datasets might be selected. Meanwhile for the suppliers we will use Wikidata and will process the sameway just like for the countries.

The semantic model created as basis for the integration of food safety data is presented in the next section. It combines the existing ontologies and datasets in the food domain, as presented in section 3. This model is represented in RDF and is easily extensible, shall some of the data processed in the platform require this. When instance data is created, based on the model, identifiers (URIs) should be unique and for this purpose we are combining the data label either with the code or row index.

In the final phase of the project, we will continue with the integration of more food safety data types, validation of semantic data (through SHACL or SHAPEs) for conformance to the model, instance matching through services developed in WP2 (as presented in D2.2) and creation of sample queries for the query explorer developed in WP2.

## 3.3 The FSM semantic model

The integrated semantic model created for the needs of the FSM data builds upon and extends the domain specific ontologies presented in section 3. The classes which are specific to the FSM platform use cases are defined as subclasses of different classes from the extended ontologies. By using the subclass property we enable all unique FSM objects and data to be read through the ontology classes thus enriching the objects and improving interoperability with other systems.

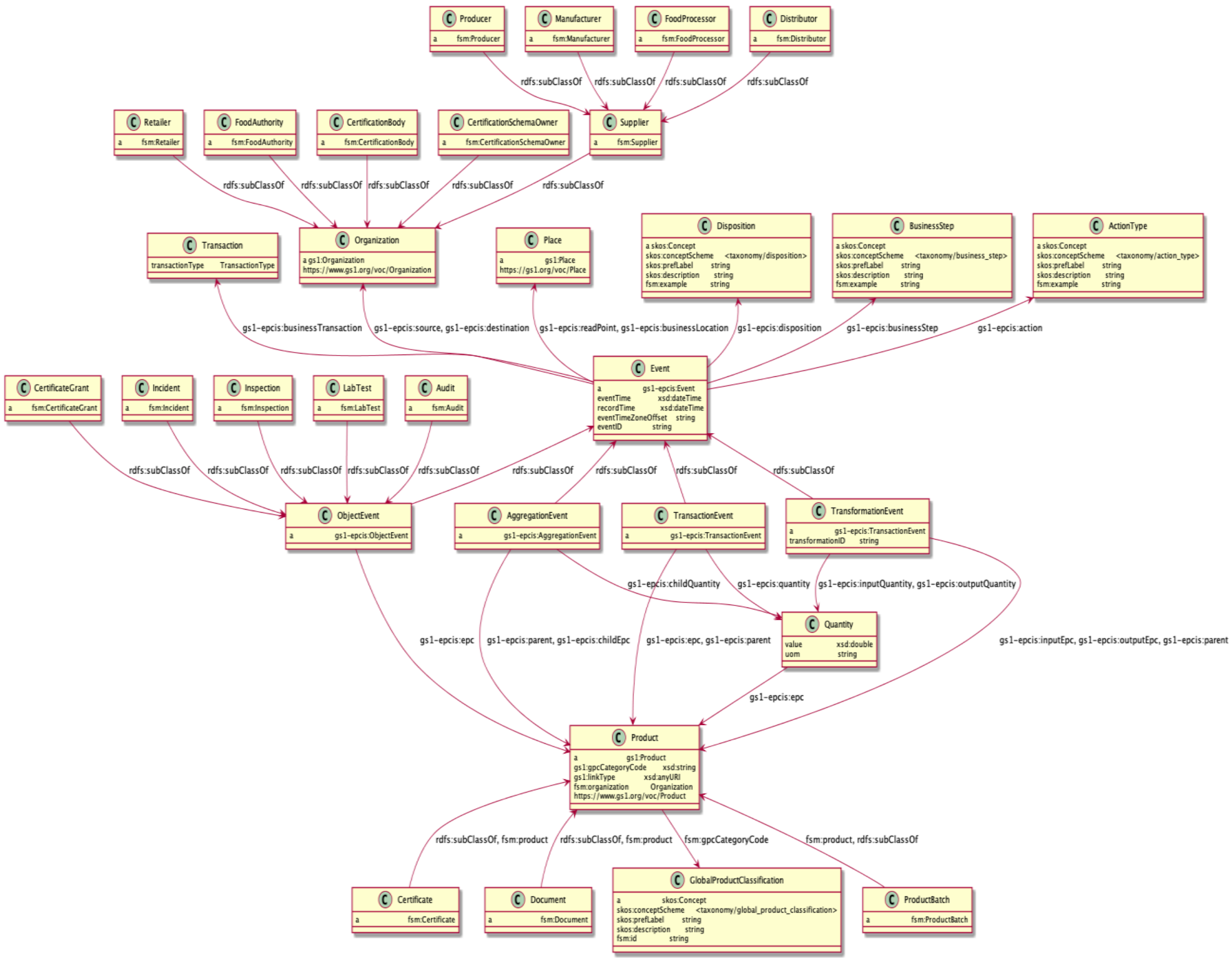


**Figure 8: FSM integrated data model**

The model defines the following list of subclasses specific to the FSM:

1. The CertificateGrant, Incident, Inspection, LabTest and Audit are made subclasses of the EPCIS ObjectEvent
2. Producer, Manufacturer, FoodProcessor, Distributor are subclasses of Supplier which together with Retailer, CertificationBody, CertificationSchemaOwner, FoodSafetyAuthority is a subclass of GS1 Organization.
3. The Certificate, Document and ProductBatch classes are subclasses of GS1 Product.
4. Products and ingredients classes are based on LanguaL and USDA representations of foods, ingredients and drinks.
5. Food Safety hazards classes enable linking to MESH and to PubMed articles that provides more information or lab test data.

### 3.3.1 Certifications

In order to express certification of producers, food processors and other parties involved in the supply chain the model needs to handle two generic cases – organization level certification information and product level certification information.

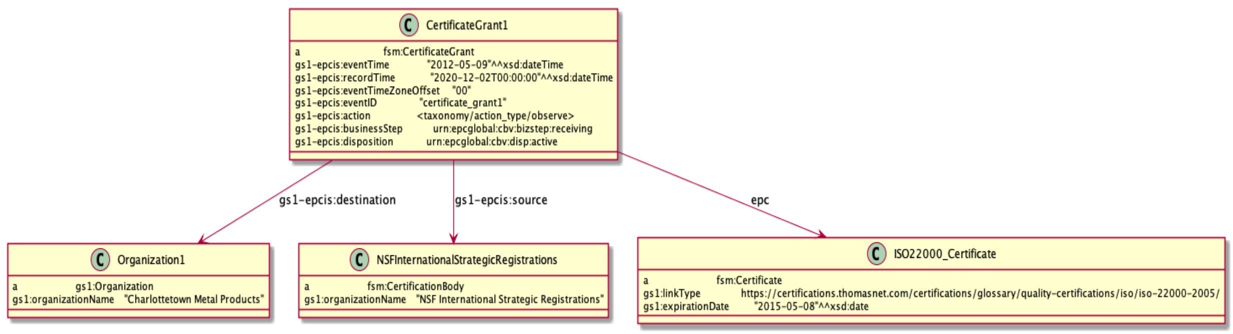


**Figure 9: Organization level certification sample**

As a start, during the first cycle of semantic model development (presented in version 1 of this deliverable), we enabled modelling of organization certification through GS1 and EPCIS. In the above example an ISO 22000:2005[39] certificate is presented. We treat the certificate as a product of the Certification Body. The certification body gives the certificate to the Organization.

[39] https://www.iso.org/iso-22000-food-safety-management.html

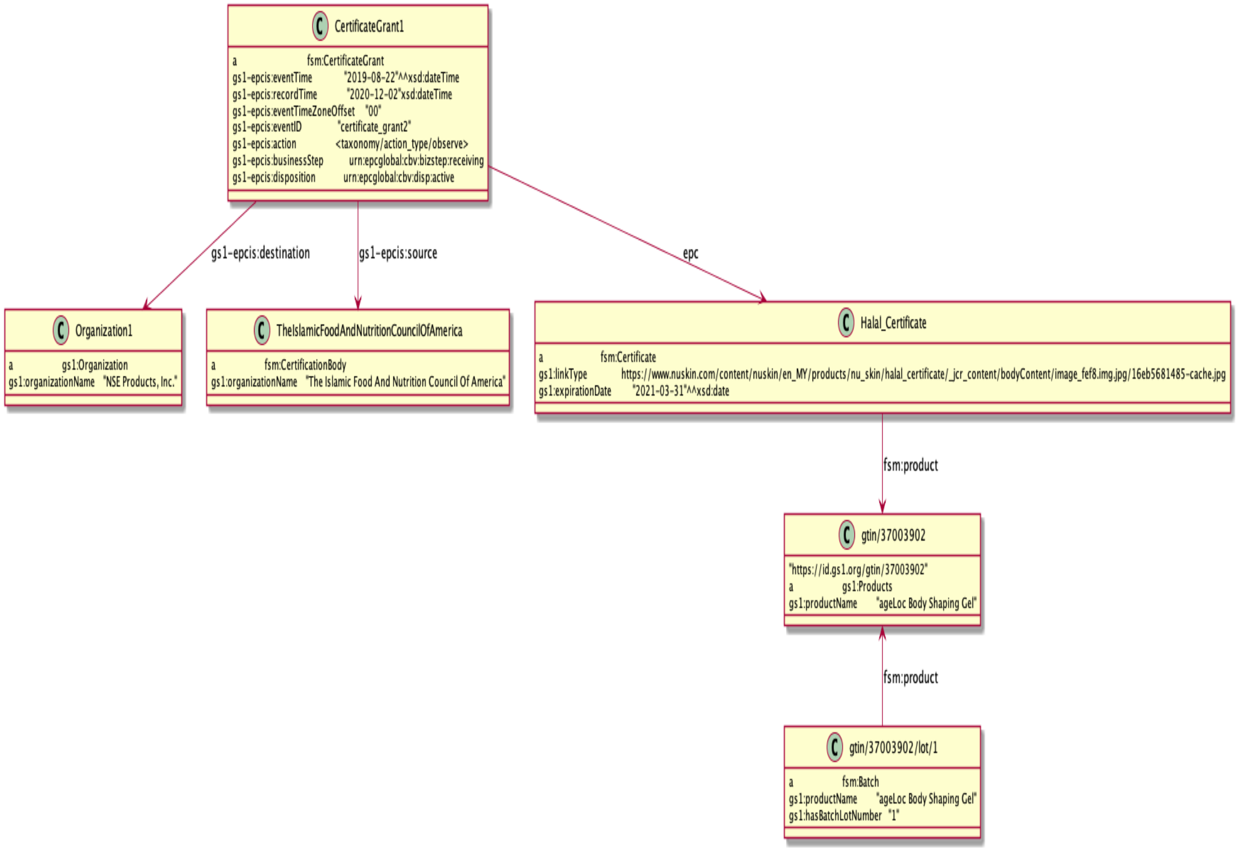


**Figure 10: Product level certification sample**

Another aspect of information covered by the model is related to the product subject of certification. Above, we provide an example of how a Halal certificate is represented. These certificates are provided only for some products provided by a company. We treat the certification grant in the same manner as in the previous example.

The extension compared to the previous example is that the certificate node links to a product. This product is a GTIN prefix for a company product. All batches of this product will be linked to it with the fsm:product link.

During the second iteration of semantic model development, carried out between M13 and M24 of the project, we have extended the certification model and mapped the FSCC to EPCIS 2.0. This allowed us to cover more details of the certification information. Below, we provide a snapshot of the model with mappings to FSCC.

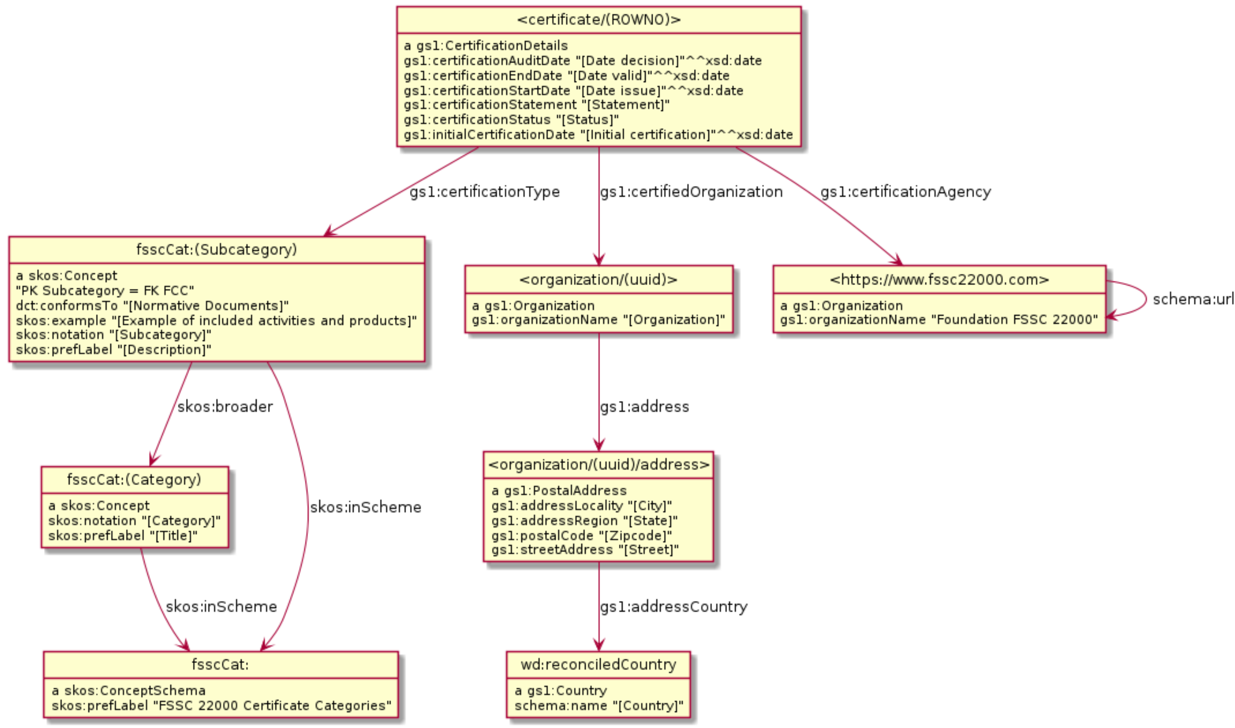


**Figure 11: FSCC categorization extension of the certification sample**

### 3.3.2 Documents

The documents are represented in a similar manner as the certifications. We treat as documents the following classes:

- Incidents
- Audits
- Inspection
- Lab tests

They can be represented using the same logic in the model Below we take the following sample from the FOODAKAI[40] API and illustrate how it is represented in the semantic model.

[40] https://www.foodakai.com/

```
"id":"FDK_18341544",
'                 "title":"Other hazard in ready to eat - cook meals by «  1001
France » from France",
'                 "description":"Date : 29 NOVEMBRE 2020\r\nINFORMATION
CONSOMMATEURS\r\nRAPPEL DE PRODUIT\r\nLa Société «  1001 FRANCE » procède
aujourd'hui au retrait de la vente de sa \r\nrecette « Mon mijoté de courges
butternut et bœuf »  suite à la mise en évidence de \r\npossibles morceaux de
graines de courges dans le produit.\r\nIl s'agit des lots portant les
caractéristiques suivantes : \r\nNature du Produit   :  Repas complet pour
bébé\r\nMarque : HAPPYLAL BABY\r\nGENCOD : 3770007731081\r\nFORMAT   :   220
grammes\r\nDLC   :  15/03/2021 (LOT : MEL46V11) et 09/05/2021(LOT :
MEL02B06)\r\nCode emballeur/Estampille Sanitaire   (ou numéro de lot) :  FR
29.174.020 CE\r\nL'ensemble des lots sont retirés de la
commercialisation.\r\nCertains de ces produits ont cependant été commercialisés
avant la mesure de \r\nretrait.\r\nIl est donc recommandé aux personnes qui
détiendraient des produits appartenant aux \r\nlots  décrits  ci-dessus  de ne pas
les  consommer et  de les  détruire  ou  de les \r\nrapporter au point de vente.
\r\nLa société « 1001 FRANCE » se tient à la disposition des consommateurs pour
\r\nrépondre à leurs questions au numéro de téléphone : « 06 62 26 38 63
».\r\nAfficher jusqu'au 29 décembre 2020",
'                 "entityType":"incident",
'                 "createdOn":"2020-11-29T00:00:00",
'                 "updatedOn":"2020-11-30T14:05:06.916673",
'                 "dataSource":"FOODAKAI",
'                 "tags":[
'                    "france",
'                    "ready to eat - cook meals",
'                    "other hazard",
'                    "prepared dishes and snacks",
'                    "europe"
'                 ],
'                 "published":true,
'                 "privateData":0,
'                 "linkedEntities":[
'                    {
'                       "id":"FDK_18341542",
'                       "title":"«  1001 France »",
'                       "description":"",
'                       "entityType":"supplier",
'                       "createdOn":"2020-11-30T14:05:06.916704",
'                       "updatedOn":"2020-11-30T14:05:06.916707",
'                       "dataSource":"FOODAKAI",
'                       "tags":[

'                       ],
'                       "published":true,
```

```
'                   "privateData":0,
'                   "linkedEntities":[

'                   ],
'                   "internalId":null
'                 }
'              ],
'              "internalId":"18341544"
'           },
```

**Figure 12: Example for an incident from FOODAKAI**

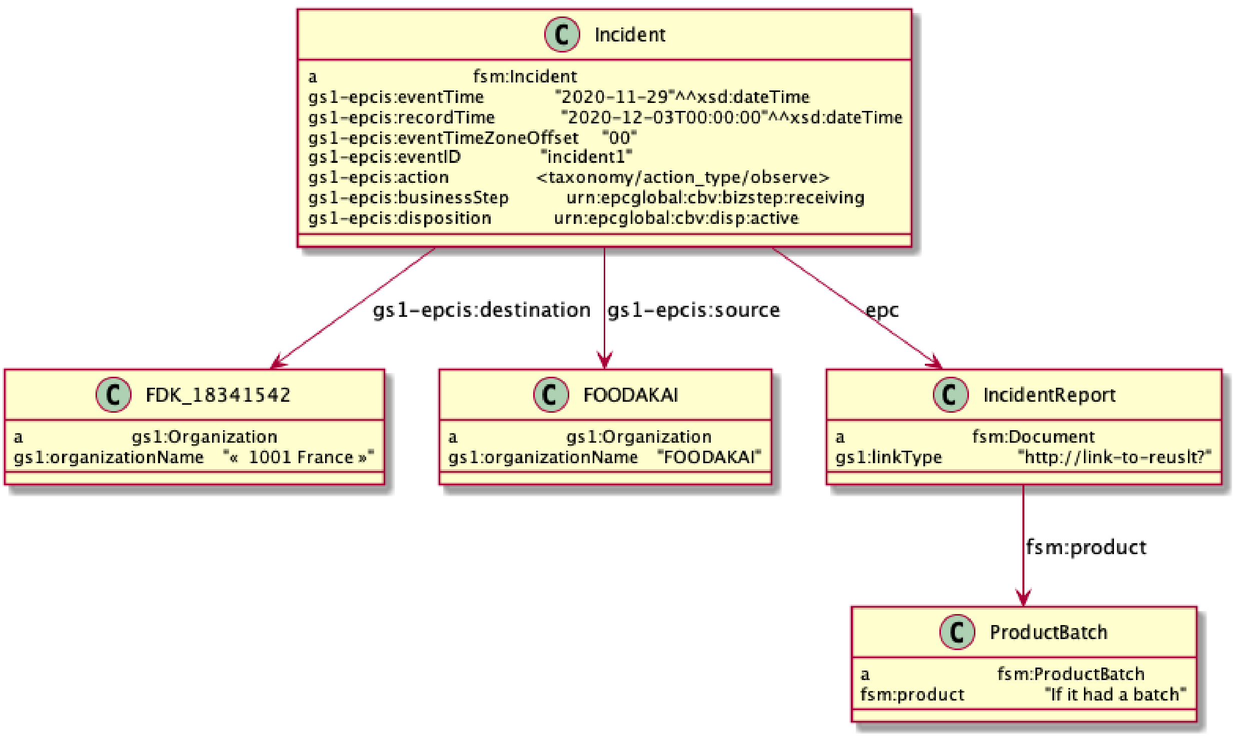


**Figure 13: Incident document example**

We treat the FOODAKAI as the source organization of the event as such information is not provided by FOODAKAI and no one outside Agroknow is able to validate it. The example incident has structured information only for the organization which supplied the food product that has led to the incident. In TheFSM Platform, we would like to be able to identify a product or a batch of products. In order to accommodate for such granularity, we use the fsm:product relation to a product.

# 4 INTEROPERABILITY PER USE CASE SCENARIO

In this section we present some implementations for TheFSM use cases scenarios: food supply chain, retailer scenario, certification body scenario and food safety authority scenario. In order to ensure interoperability for the FSM platform, the data is integrated through the semantic model and is thus compliant with GS1, EPCIS, W3C and other domain-specific ontologies.

## 4.1 Food supply chain example

We took the use case regarding the life cycle of broilers and their supply chain. In the diagram we can see the different organizations and their interaction in the process of making broiler meat for consumers.

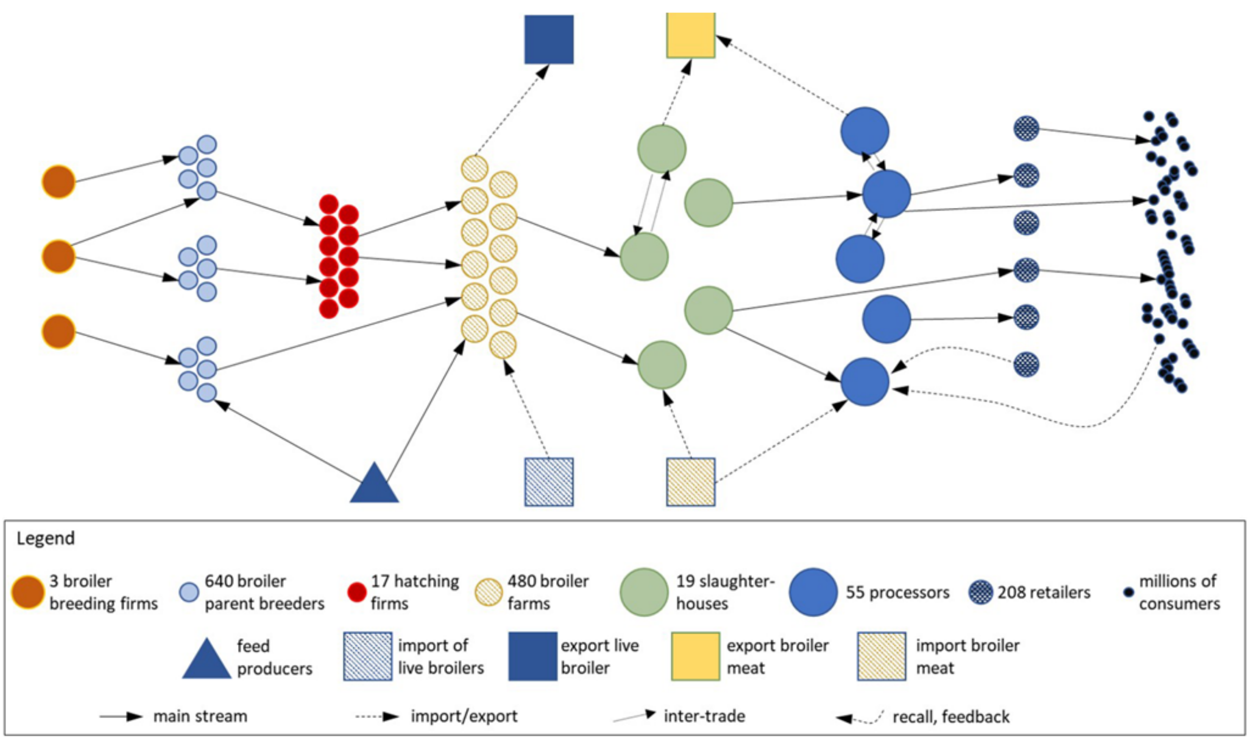


Figure 14: Food supply chain scenario

First, we apply the GS1 EPCIS standard to the use case. We represent any movements of products and consumption of such in the case of feeding as an event.

- We have 6 Transportation events that reflect the mainstream from the presentation. These events are of type observation. We assume that with a single event we can capture the transition of the product from one organization to another. We could think of the event as the place in time which the product is received in the second organization. We omit and in fact infer the event that a product has been sent by the first organization.

- We add additional events for laying an egg, hatching an egg, slaughtering a broiler, and packaging the meat. These events are of type Transformation. We are assuming that one or more products are consumed to produce another product. In the case of laying an egg, the transformation gets as input the broiler and outputs the same broiler and the egg. For the other cases, the logic is similar.

We assume that every broiler, packaged food, broiler egg, broiler meat and broiler packaged meat is uniquely identified. The model currently is not complicated with the logic of handling aggregated packages of broilers and etc.

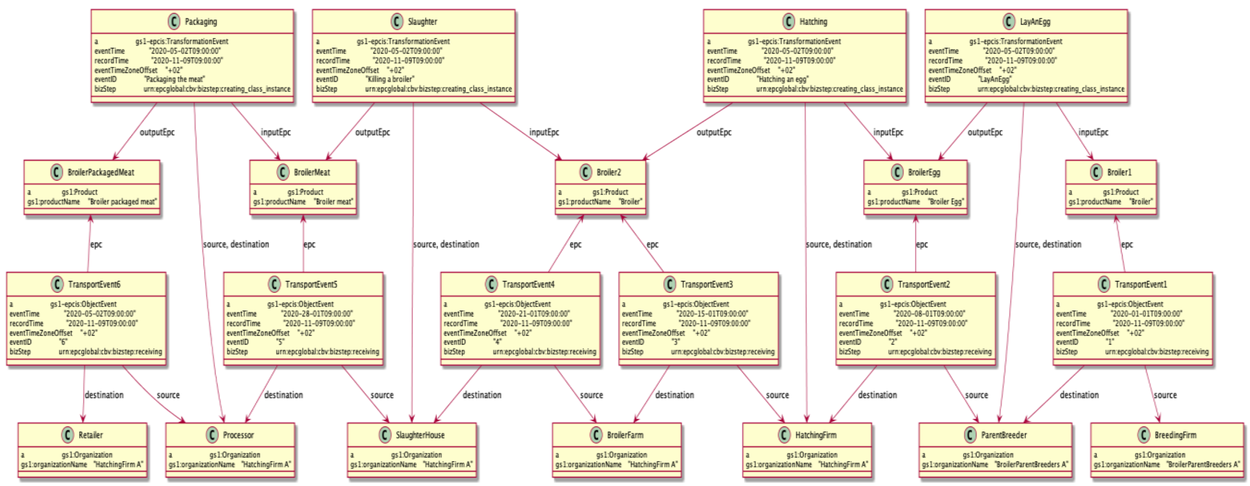


**Figure 15: Scenario implementation with the FSM model**

The above diagram shows the model of the main flow of the products. We will add additional diagrams displaying the secondary relations between the organizations.

The feeding of the broilers is important in order to keep track of how they have been raised and the quality of the meat. The feeding is relevant to Broiler Parent Breeders and Broiler Farms. The process of feeding is captured in two events:

- Buying of food - the point in time when a batch of food is being received in a farm or breeder organization. The food has its unique identifier signifying exactly which food package we refer to.

- The feeding itself - the point in time when a broiler is fed with a particular package of food.

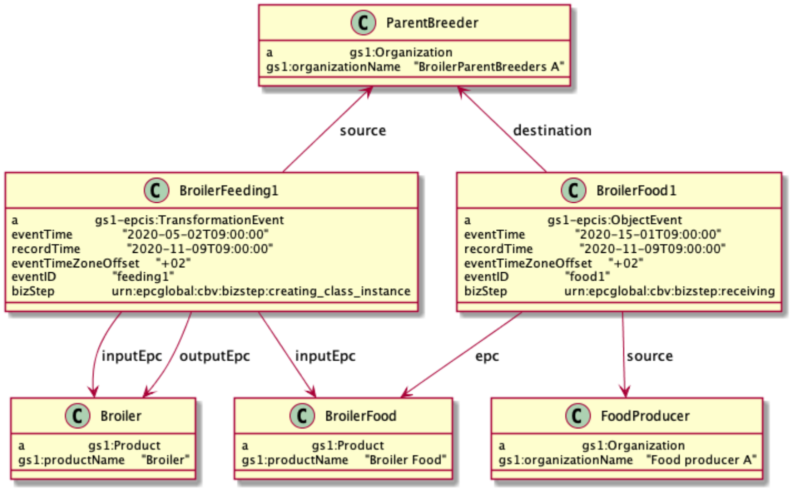


Figure 16: Scenario feeding implementation

The import and export of broilers are represented in the diagram below. An import or export is a single event. The event is the point in time when the product reaches the destination organization.

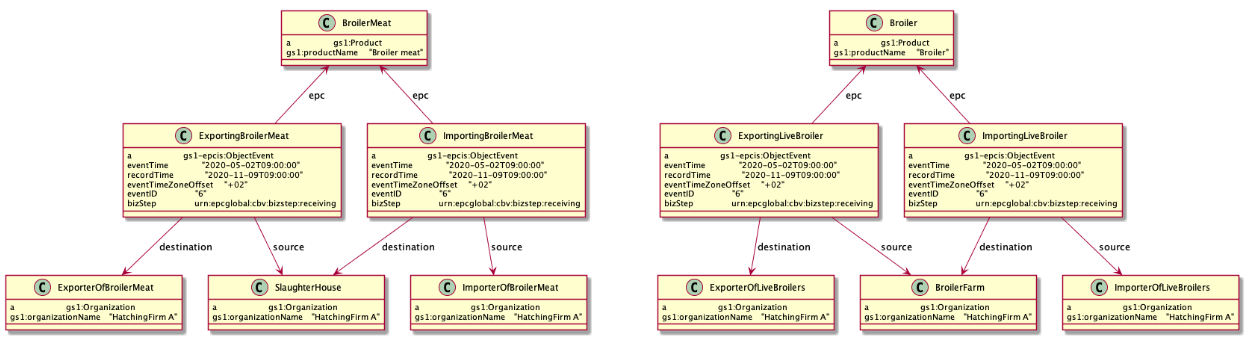


Figure 17: Scenario import/export implementation

## 4.2 Retailer Scenario

We use these requirements:

1. Browse information
2. list of the Certification Bodies and Certification Schema owners that the supplier (producer, processor) is working with.
3. able to perform a remote supplier verification using critical information like incidents, inspection results, certificates, and lab tests
4. able to perform a supplier risk assessment in order to prioritise the audits and lab tests
5. Be able to access fully traced information
6. Be able to access information regarding findings of the inspection of suppliers in the food chain

7. Be able to access the current status of food supply actors, as far as audit results of certified organizations are concerned
8. Have access to innovative tools

How we cover the above requirements:

1. Based on the FSM model we can browse organizations, products, product batches. These organizations can be Retailers, Suppliers, Certification bodies, etc. We can query the database for Events regarding these objects. The events can be transactions between organizations, transformations of a product to a new product, certification grants, lab tests etc.
2. Certification bodies can be listed by querying for objects of type fsm:CertificationBody.
3. We can query directly inspection results, certificates and lab tests relating to an organization. This can be done using the event objects fsm:Inspection, fsm:CertificateGrant, fsm:LabTest, fsm:Audit respectively. The incidents are usually harder to trace back. Based on the presumption that we have the full supply chain data we can query the graph for all the events and parties involved with this product. The entry point for such an analysis should be the fsm:Incident object.
4. This functionality should be extracted from the raw data within the FSM model. It should be part of down-stream processing.
5. As per 3. we can fully trace a product batch through the supply chain as long as we have the full data for it.
6. This is part of 3. browsing inspection results should give you access to view the information within the system.
7. The current state should be logically aggregated based on historical event data. All the data will be within the system and will be queryable. A question like: How long results of Inspection/Audit are relevant after their release?
8. This should be part of another layer above the database.

## 4.3 Food Processor Scenario

Based on an overview of the responsibilities of 2 EU food authorities (Bulgarian Food Safety Agency, Netherlands Food and Consumer Product Safety Authority) we have extracted their obligations:

- Animal Health - Keep animals healthy. Animal welfare. Animal products processing secure
- Plant health - Plant diseases and prevent pests
- Food safety - Supervising the production, preparation, transport and sale of food products and sale of tobacco products.
- Product safety - Safe personal care products, indoor and outdoor games, consumer products in and around the house.
- Border inspections

The data we deem necessary for the FSM Platform are the:

- Reports of inspections - they could include information about safety risks.[41]
- News of new regulations[42] - In order to make this type of data queryable it would require to add the concepts of geo locations. As such regulations are issued for specific countries.

Currently we decided to use only the reports of inspections. For example, inspections are part of the Documents heading explained above.

## 4.4 Certification Body Scenario and Certification Body for Organic Wine Production Scenario

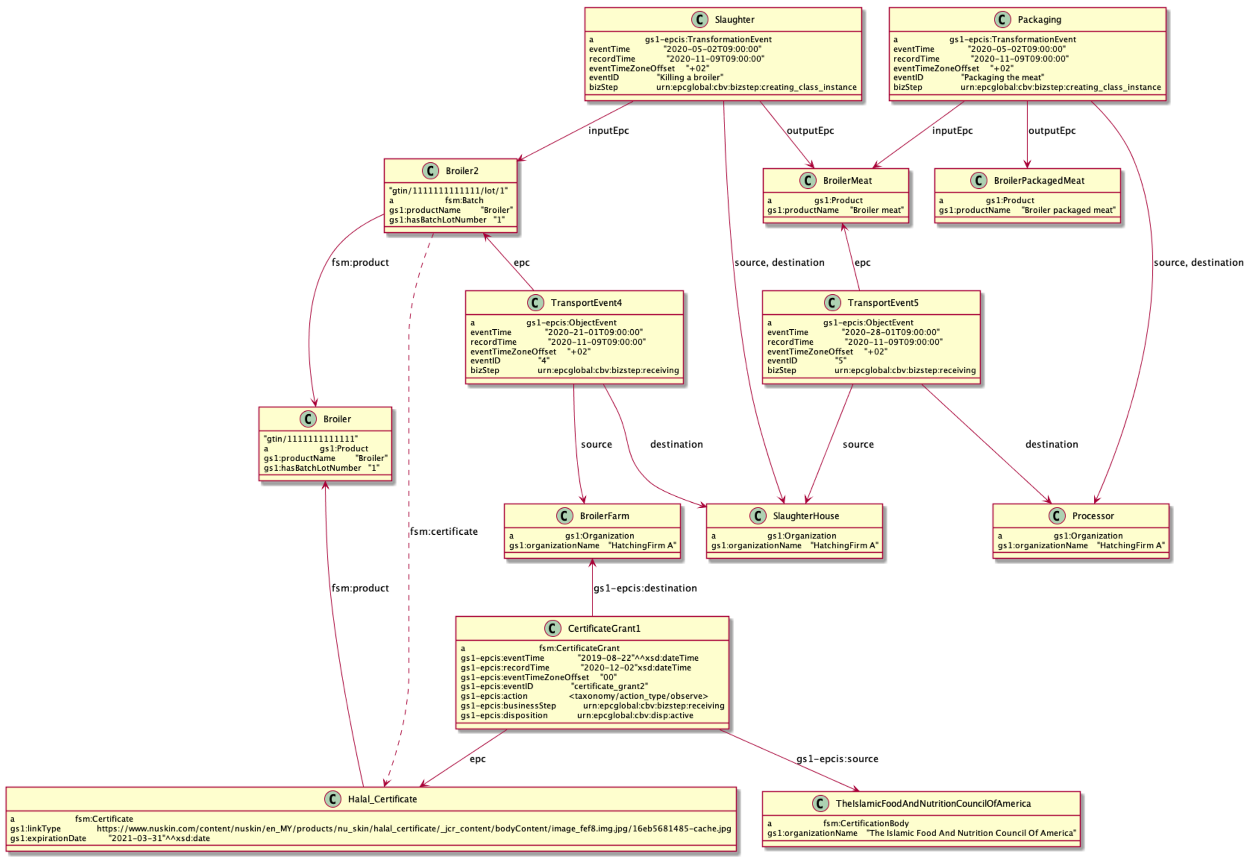


**Figure 18: An example of the use cases involving certifications**

The organization BroilerFarm has Halal certificate for broilers. The Broiler2 is a product batch for broilers. Based on the hatching of the broiler and the time of its slaughter we can deduce whether the certificate is relevant to this exact broiler batch. For the example we assume that the hatching of the broiler is after the Halal certificate grant. We can see that the transportation of the broiler

[41] https://english.nvwa.nl/news/news/2017/08/03/nvwa-website-publishes-codes-of-fipronil-contaminated-eggs-from-investigated-farms
[42] https://english.nvwa.nl/news/news/2020/11/30/bird-flu-avian-influenza-now-as-well-in-belgium-and-poland-mandatory-additional-cleansing--disinfection-transport-vehicles-in-the-netherlands

to the slaughterhouse is prior to the expiration of the Halal certificate. Thus we infer that the broiler batch is under halal certification.

## 4.5 Food Safety Authority Scenario

The food safety authority can be identified using the fsm:FoodSafetyAuthority. Events can be associated with the authority. The currently supported events are of type Incident. They state that an incident has been reported for a certain product batch or a company product. These events fall in line with the Document scenarios explained in the FSM integrated model section Documents.

The information for incidents and the reporter is searchable and can be traced back through the supply chain to identify the companies involved in producing this batch of food.

# 5 CONCLUSION AND NEXT STEPS

In this document we have presented how semantic enrichment is applied over partners data by using public datasets and automation processes, how we have mapped tabular data to different ontologies and standards during the conversion to RDF. The process of semantically enriching food safety data is performed in an iterative manner.

The first iteration was performed by M12 of the project and involved creation of data inventory with samples and datasets provided by partners, studying applications APIs that serve data, analysis of public thesauri and ontologies in the food safety domain, as well as defining the initial semantic model.

During the second year of the FSM project we proceeded with another iteration by processing more data and transforming it into the semantic model. More standards were covered as well to enhance interoperability through extensions of the semantic model.

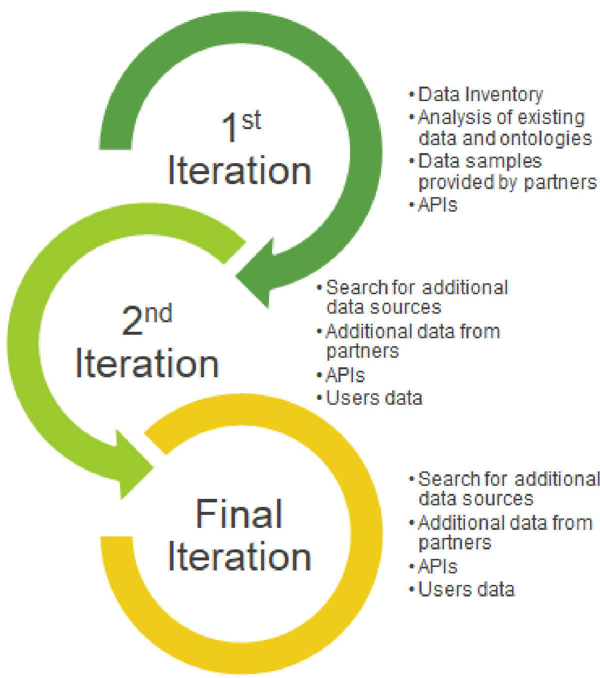


**Figure 19: The agile process of data modelling**

The final iteration will take place in year 3 of the project when we will transform the remaining data types to the semantic model and enable their mapping to standard vocabularies.