Downloaded from orbit.dtu.dk on: Apr 26, 2024

DTU Library

Global Food-source Identifier (GFI)

Collaborative virtual research environment and shared data catalogue for the foodborne outbreak investigation international community

Ribeiro Duarte, Ana Sofia; Nielsen, Cecilie Liv; Candela, Leonardo; Valentin, Lars; Aarestrup, Frank Møller; Vigre, Håkan

Published in:
Food Control

Link to article, DOI:
10.1016/j.foodcont.2020.107623

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Ribeiro Duarte, A. S., Nielsen, C. L., Candela, L., Valentin, L., Aarestrup, F. M., & Vigre, H. (2021). Global Food-source Identifier (GFI): Collaborative virtual research environment and shared data catalogue for the foodborne outbreak investigation international community. *Food Control*, *121*, Article 107623. https://doi.org/10.1016/j.foodcont.2020.107623

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Food Control 121 (2021) 107623

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Global Food-source Identifier (GFI): Collaborative virtual research environment and shared data catalogue for the foodborne outbreak investigation international community

Ana Sofia Ribeiro Duarte [a,*], Cecilie Liv Nielsen [a,d], Leonardo Candela [b], Lars Valentin [c], Frank M. Aarestrup [a], Håkan Vigre [a]

[a] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark. Kemitorvet 204, DK 2800, Kongens Lyngby, Denmark
[b] Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", National Research Council of Italy, Via G. Moruzzi, 1 - 56124, Pisa, Italy
[c] Unit Food Hygiene and Technologies, Supply Chains, Food Defense, Department Biological Safety, German Federal Institute for Risk Assessment (BfR), Max-Dohrn-Strasse 8-10, 10589, Berlin, Germany
[d] Department of Animal Science – ANIS Welfare, Aarhus University. Blichers Alle 20, DK 8830, Tjele, Denmark

ARTICLE INFO

Keywords:
Foodborne outbreak investigation
Microbial typing
Food source
Data sharing
Harmonization
Virtual research environment

ABSTRACT

The source of a foodborne disease outbreak (FBO) is often difficult to identify, especially in the early phase where interventions would be most efficient. In addition, data on FBOs are mostly scattered in different formats either in national databases and reports or within pathogen-specific or regional reporting networks, both of which are often only accessible to a selected number of individuals.

Here, we developed an international, open, shared and searchable data catalogue of past FBOs – the Global Food-source Identifier (GFI). GFI was developed with two objectives: a) to create a collaborative online community of FBO investigators, encouraging the international sharing of data in a harmonized, detailed and comparable manner and b) to support foodborne outbreak investigation worldwide by providing access to detailed records of past outbreaks, which can convey valuable insight into potential 'risk foods' of a detected pathogen. GFI is hosted within a Virtual Research Environment (VRE), which offers additional features to facilitate the collaboration between the outbreak investigators. These features allow document exchange, communication and data visualization and analysis between the VRE members.

Based on scientific literature on foodborne outbreaks and discussions within a working group, we selected a total of 46 attributes characterising the outbreak records to be included in the catalogue, aggregated under the four overarching categories *causative agent*, *epidata*, *food source* and *report details*. Detailed descriptions of the attributes in the catalogue and instructions for harmonized data reporting are available on a wiki page in the VRE. At the time of writing and public launch of GFI, the data catalogue was populated with records of 102 FBOs occurred in Denmark over a period of 12 years (2005–2016) and covering the most frequent pathogens and a broad range of typing methods.

The VRE features that enable data analysis, document sharing and communication between members were applied for the graphical representation of the records available in GFI, and for the sharing of results and script files within the VRE. The descriptive analysis included the relationship between the most frequent causative agents and outbreak food sources. Such results can support a risk-based food sampling strategy in the very beginning of a foodborne outbreak investigation.

The Global Food-source Identifier is a data catalogue specifically designed to host an international collection of FBO records reported in a detailed and harmonized manner. It is implemented in a virtual research environment that offers key features to facilitate and enhance the global collaboration and data sharing among FBO investigators. Once in active use by the international food safety community, we envisage that GFI will contribute to the success of FBO investigations worldwide.

* Corresponding author. Kemitorvet 204, DK 2800, Kongens Lyngby, Denmark.
E-mail address: asrd@food.dtu.dk (A.S. Ribeiro Duarte).

https://doi.org/10.1016/j.foodcont.2020.107623
Received 9 July 2020; Received in revised form 4 September 2020; Accepted 8 September 2020
Available online 15 September 2020
0956-7135/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In 2019 the United Nations World Food Safety Day was celebrated for the first time, with the aim to strengthen the global efforts to ensure food safety. Foodborne infectious diseases have been estimated to affect 550 million persons and cause 230,000 deaths globally in 2010, with 43% of the disease burden from contaminated food occurring among children under 5 years of age (World Health Organization (WHO), 2015). These estimates are considered conservative due to considerable underreporting, and foodborne outbreaks (FBO) are responsible for a large part of the estimated burden. In addition to these direct consequences FBO can cause major losses due to restrictions on trade and sale of specific food products.

FBO are caused by many different agents contaminating a broad range of foods (EFSA&ECDC, 2015). The task of investigating an outbreak is complicated among other factors due to the difficulty to quickly identify and recover the contaminated food. In Europe, the source of 7% of strong-evidence FBO and 73% of weak-evidence FBO was either reported as 'unknown' or 'other foods' (EFSA&ECDC, 2015). Similarly, in the US, 46% of reported outbreaks had an unidentified food vehicle (Richardson et al., 2017). Traceback studies not only allow resolving an outbreak investigation and containing disease spread, they are also important to support food safety policy evidence-based decisions (Gormley et al., 2011). Rapid, and correct identification of the source is absolutely essential, both to limit human illness and to avoid unnecessary restrictions on falsely implicated food products. Further, rapid identification of the most likely source for sampling and subsequent conventional or molecular analysis is the first and essential step in any outbreak investigation.

Characterisation of foodborne isolates differs between countries and laboratories, and may include different phenotyping or molecular typing methods. Different protocols for various typing methods exist and some typing results can be retrieved both from conventional protocols and from whole genome sequencing (Joensen et al., 2015). Nomenclature of outbreak strains can also differ between countries and regions (Peters et al., 2017). However, with an increasing evidence for the transboundary nature of foodborne pathogens there is an increased need for international collaboration and data sharing for surveillance and outbreak investigation. Indeed, one of the strategic directions of WHO's plan to lower the burden of infectious foodborne disease is to improve international collaboration and enhance communication, which are recognized major challenges when food supply chains cross multiple national borders (World Health Organization (WHO), 2015). In order to address these challenges at a truly global scale and in an all-inclusive manner, solutions that enhance collaboration, communication and data exchange need to encompass a wide range of typing methods and epidemiological data, allow reporting of the typing protocols used, and adopt standardised nomenclature, in order to ensure transparency and comparability of results.

Here we addressed the challenge of exchanging typing results and epidemiological data from FBO investigations, in a collaborative, interactive and harmonized way, internationally. We created a free and open access, shared and searchable data catalogue of foodborne outbreaks, where each record includes detailed information on the type of causative agent, the identified food source and epidemiological information. The catalogue is integrated in a virtual research environment (VRE) named Global Food-source Identifier (GFI).[1] A VRE is an open, alive, online system helping people and organisations to easily share information and results, when research and outbreak investigations are carried out in teams which span across institutions and countries. Features within a VRE usually include collaboration support (Web forums and wikis), document hosting, data analysis and data visualization

[1] Global Food-source Identifier Virtual Research Environment dashboard https://aginfra.d4science.org/web/foodborneoutbreak.

Table 1
Description of fields in the GFI data catalogue.

Field	Field description
Title	Title of the outbreak record (agent_country_year)
Description	Short description of the outbreak (agent, country, year, food source)
Tags	Record tags (agent, country, year, food source, …)
Data and Resources	Scientific publication, report and/or data related to outbreak
Item URL	URL of the outbreak record
Causative agent	
CausativeAgent	Causative agent of the outbreak
AMRProfile	AMR profile of the outbreak strain
AMRMethod	AMR method used
Genogroup	Norovirus genogroup of the outbreak strain
GenogroupProtocol	Norovirus genogroup protocol used
MLST	MLST type of the outbreak strain
MLSTScheme	MLST scheme used
MLVAProfile	MLVA-type of the outbreak strain
MLVAProtocol	MLVA protocol used
PFGETyped	Was the outbreak strain PFGE typed?
PFGEProtocol	PFGE protocol used
Phagetype	Phagetype of the outbreak strain
PhagetypeStandard	Phagetype standard used
Serotype	Serotype of the outbreak strain
SerotypeMethod	Serotyping method used
VirulenceProfile	Virulence profile determined for the outbreak strain
VirulenceProfileDetectionMethod	Method used to determine virulence profile
WasTheIsolateSequenced	Was the outbreak strain Whole Genome Sequenced (WGS)?
OtherTypingInformation	Additional typing information, incl. WGS accession number, if available
Epidata	
Cases	Number of patients involved in the outbreak
Countries	Country/countries involved in the outbreak
Hospitalized	Number of patients involved in the outbreak who were hospitalized
Men	Number of male patients involved in the outbreak
MultipleStrains	Was more than one outbreak strain involved in the outbreak? (Preferably, cases should be reported per outbreak strain. If cases cannot be attributed to individual strains, they should be reported in different records for each strain)
Old	Number of patients above 65 years old involved in the outbreak
OutbreakDuration	Duration of the outbreak in months (outbreaks lasting one month or less should be reported as 1 month)
ReferenceOtherStrain	Title of the record(s) referring to other strain(s) involved in the same outbreak
Setting	Specific setting for the outbreak (EFSA&ECDC, 2015)
SingleCountryOutbreak	Was there only one country involved in the outbreak?
StartMonth	Month for first registered case in outbreak
StartYear	Year for first registered case in outbreak
TravelRelated	Was the outbreak travel related? (Travel = travel abroad in a seven-day period prior to disease onset (Anon., 2017))
TravelCountry	If the outbreak was travel related, to which country have the cases travelled to?
UnknownGender	

(continued on next page)

Table 1 (*continued*)

Field	Field description
	Number of patients with unknown gender involved in the outbreak
Women	Number of female patients involved in the outbreak
Young	Number of patients below 5 years old involved in the outbreak
Food source	
MajorFoodType MajorFoodSubtype FoodSubtypeVariety A_ProcessingPreparationConsumption B_ProcessingPreparationConsumption	Food sources were categorized based on a hierarchic food categorization scheme from the Interagency Food Safety Analytics Collaboration (IFSAC) (Richardson et al., 2017; see supplementary material).
CountryOriginOfContamination	Country where food was contaminated with the disease causing agent
FoodOriginOfContamination	Production level where food was contaminated with the disease causing agent
OtherFoodsourceInfo	Specific food source causing the outbreak (e.g. raspberries), or other additional information related to the foodsource
Report details	
EvidenceForOutbreak	Evidence for outbreak (epidemiological and/or microbiological)
Reference	URL(s) to relevant scientific literature for the outbreak
Management Info	
Author	Name of user who reported the outbreak
Maintainer	Name of user who maintains the record
Version	Record version
State	State of the record (e.g. active)
Last Updated	Date of last record update
Created	Date for record creation

AMR - antimicrobial resistance.

facilities (Candela et al., 2013).

GFI is intended to host an online international community involved in outbreak investigation, who can communicate and exchange no confidential data in a harmonized, comparable manner. The main aim of GFI is to assist that international community during outbreak investigation.

To the best of our knowledge, an international networking platform which is open, allows data sharing in a transparent way and includes detailed information on a comprehensive span of pathogen typing results of such a wide range of phenotypic and genotypic typing methods, on food source and on epidemiology, is not available at the moment. A similar concept has been implemented in the past with the Pathogen-annotated Tracking Resource Network (PATRN) (Gopinath et al., 2013), developed in the US by the Center for Food Safety and Applied Nutrition(CFSAN) and Food and Drug Administration (FDA). However, to our best knowledge, this resource is no longer available. Several countries have national databases containing information on foodborne, waterborne or infectious disease outbreaks. (Anon., 2017; Gormley et al., 2011; Guzman-Herrador et al., 2016; Krause et al., 2007; Richardson et al., 2017). These databases are structured in different ways and are not interconnected. Communication between countries relies therefore many times on the individual networks of the people involved in national outbreak investigation. There are also several existing international databases that do not match the characteristics of GFI, such as the Foodborne Illness Outbreak Database (Marler Clark LLP, 2018), The European Surveillance System (TESSy) (ECDC, 2019), NoroNet (RIVM, 2018), PulseNet International (Nadon et al., 2017), The Global Microbial Identifier (Wielinga et al., 2017) or GenomeTrakr (Timme et al., 2019). Most of these data collections are targeted at a specific pathogen or typing method, and many are only fully accessible to a limited number of persons or accessible to a limited extent to the public.

Here we describe the development of GFI, including i) the data catalogue that includes multiple pathogens and a comprehensive set of typing methods and protocols, and ii) the VRE - a living environment that hosts the data catalogue, and many additional features that facilitate the collaboration between the GFI community members.

2. Methods

2.1. Data catalogue of foodborne outbreaks

2.1.1. Selection of variables

The outbreak record format supported by the catalogue has been developed as parsimoniously as possible, including only variables considered relevant for foodborne outbreak investigation and those generally present in reports of FBO (Table 1). The selection was based on a review of scientific literature, advice from two experts in food- and waterborne outbreak investigation from the Danish Statens Serum Institut (SSI) and discussions within a working group, which included three experts on epidemiological investigation of foodborne disease from the National Food Institute (Technical University of Denmark), a food microbiologist from SSI, and a computer scientist expert on virtual research environments development and catalogue exploitation from the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (National Research Council of Italy).

2.1.2. Population of the catalogue

The catalogue was initially populated with outbreak data described in the Annual Report on Zoonoses in Denmark in the sections of 'Food- and Waterborne outbreaks' from 2013 to 2016, 'Outbreaks of special interest' from 2007 to 2012 and 'Outbreaks of zoonotic gastrointestinal infections' from 2005 to 2006 (Anon., 2006, 2012, 2013, 2017). Outbreaks considered neither of food- nor of waterborne origin were excluded. The intention was to populate the data catalogue with a minimum of 100 outbreaks covering a period of approximately 10 years, the most frequent foodborne pathogens and a broad range of typing methods. This helped to complete as much as possible the list of typing methods and protocols used for different pathogens. Besides the information retrieved from the Danish zoonoses reports, additional data were extracted from the national database of foodborne outbreaks (Fødevareudbrudsdatabase (FUD)) and whenever needed, from scientific literature describing the outbreaks in more detail. For outbreaks reported both in the zoonoses report and in FUD, the data from FUD was preferred if discrepancy between the two sources existed. Scientific literature used included references mentioned in the zoonoses reports and a selection of additional references from a list provided by one of the outbreak investigation experts (Steen Ethelberg, *personal communication*).

2.2. Virtual research environment for outbreak investigation

2.2.1. GFI features

The GFI VRE was created to support the establishment of an international collaboration environment for foodborne outbreak investigation, including a data sharing opportunity as well as features that allow document exchange, communication and data visualization and analysis.

The VRE is hosted by the AGINFRA Gateway (Assante et al., 2019a), a science gateway developed by relying on the D4Science infrastructure (Assante et al., 2019c) and the gCube technology (Assante et al., 2019b). This VRE realizes a web-based working environment (Fig. 1a) providing its users with: (a) a *shared workspace*, for storing, organising and sharing any version of an artefact, including dataset and model implementation

Fig. 1. Features of the GFI virtual research environment.
Snapshots of the virtual research environment: *a* - welcome page; *b* – data catalogue; *c* –shared workspace; *d* – data analytics options and R studio; *e* – social networking.

(Fig. 1c); (b) a *social networking area* enabling collaborative and open discussions on any topic and disseminating information of interest for the community, e.g. the availability of a new outbreak record (Fig. 1e); (c) the *catalogue* recording the outbreak records worth being shared thus to make it possible for others to be informed and make use of these assets (Fig. 1b); (d) an RStudio-based development environment for R enabling to perform statistical computing tasks in the cloud. The environment provides its users with a powerful integrated development environment

Table 2
Number of outbreaks, outbreak strains and cases per year available in the catalogue at time of public launch.

Year	Outbreaks	Strains	Cases
2005	10	10	1263
2006	5	6	533
2007	4	23	429
2008	7	13	1825
2009	6	6	786
2010	25	54	1091
2011	5	5	4024
2012	9	10	546
2013	8	14	729
2014	7	7	525
2015	6	6	296
2016	10	12	1419
Total	102	166	13,466

Extracted data can be found within the VRE environment.[3] An additional output file is the KNIME workflow in its executed status, which the user can open in KNIME (version 3.7 and above) to check individual nodes, if wished. Note that sharing the executed workflow file with other users will imply sharing the personal user token, which is not advisable.

Visualization plots and descriptive analysis of the present catalogue data were performed using R by RStudio (R version 3.4.4) (R Core Team, 2018; RStudio Team, 2016) within GFI, and the results were both stored in GFI's workspace and shared in the social networking area, to illustrate GFI's potential to support collaborative outbreak investigation. The R script is provided in a R notebook available to users in the shared workspace.

3. Results

The catalogue records are characterised by 46 attributes divided under four overarching categories: (i) *causative agent*, (ii) *epidata*, (iii) *food source* and (iv) *report details* (Table 1). Additionally, each record is automatically associated to tags for outbreak year, causative agent, country and food source (type, subtype and variety), and a unique record URL is generated. Management info is also added by default, including author, maintainer, version, state, creation date and date of last update. Users are encouraged to attach complementary data and/or resources to the records, including e.g. scientific publications where the outbreak was described, or a related national/international outbreak report. Record visibility may be set as 'restricted' (only VRE members will see the full record) or 'public'. GFI members have access to all catalogue records, while external users only have access to public records and the title of restricted records.

A detailed description of the catalogue record' attributes with predefined choice- or syntax-options is presented in Appendix A. For the category *causative agent* the following conditions applied: *(i)* fields were included in order to allow a detailed report of typing results; *(ii)* for each typing method, a list of the most common protocols is provided in a

Table 3
Number of records by causative agent and typing method available in the catalogue at time of public launch.

Causative agent		Serotype	PFGE type	Phagetype	MLVA profile	MLST	AMR profile	Virulence profile	Genogroup	WGS
Campylobacter	3									1
Campylobacter jenuni	6					2	1			3
Clostridium perfringens	2									
Cryptosporidium hominis	1									1
Enteropathogenic *Escherichia coli* (EPEC)	3	3						1		
Enterotoxigenic *Escherichia coli* (ETEC)	22	21								
Shiga toxin/verocytotoxin-producing *Escherichia coli* (VTEC/STEC)	8	6	1			3		5		3
Other pathogenic *Escherichia coli*	2	0					1	1		
Entamoeba histolytica	1									
Giardia intestinalis/lamblia/duodenalis	1									
Hepatitis A virus	2									
Listeria monocytogenes	4	1	2			2				2
Norovirus	42								27	22
Other parasitic agents	3									
Rotavirus	2									
Salmonella enterica subsp. enterica	54	54	7	25	37	1	30			15
Sapovirus	1									
Shigella sonnei	4		1				1			
Staphylococcus aureus	1									
Streptococcus	1		1							
Yersinia enterocolitica	3	1								
Total	166	86	12	25	37	8	33	7	27	47

(IDE) including a console, a source code editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management (Fig. 1d); (e) a Jupyter-based notebook environment for documenting and recording analytics processes (Fig. 1d). Every notebook is a rich document that contains live code, equations, visualizations and narrative text aiming at capturing a research activity.

2.2.2. Collaborative data analysis

The full content of the data catalogue can be extracted in the format of a.xlxs or the proprietary KNIME table file via the execution of a KNIME workflow integrated into the VRE.[2] The workflow was developed with the software KNIME (version 3.8) (Berthold et al., 2007) and is offered by the *as-a-Service* delivery model. When extracting the data, the user's personal token is automatically provided in the workflow. After execution of the workflow, the user can proceed with analysis of the extracted data within the VRE, using RStudio or Jupyter notebook.

[2] KNIME workflow available at: https://aginfra.d4science.org/group/foodborneoutbreak/data-miner.

[3] Data extracted from the catalogue and the R notebook can be accessed and downloaded at: https://aginfra.d4science.org/group/aginfra-gateway/workspace?itemid=8b1cf941-9571-4c12-80e7-42c525472f03.

Table 4
Number of records by food category available in the catalogue at time of public launch.

MajorFoodType		MajorFoodSubtype		FoodSubtypeVariety		A_ProcessingPreparationConsumption		B_ProcessingPreparationConsumption	
Fish	1								
Dairy	1	Fluid milk	1			Unpasteurized	1		
Meat-Poultry	29	Meat	25	Beef	8	Ready-to-eat	6	Acidified/Fermented	1
				Pork	16			Salt-cured	1
								Fully-cooked	4
						Raw/other processed	1		
						Non-intact raw	3	Ground beef	2
		Poultry	3	Chicken	2				
				Other poultry	1				
Eggs	6	Shell Eggs	3						
Produce	38	Fruits	9	Small	9	Frozen	9		
		Vegetables	29	Sprouts	2				
				Seeded	3	Fresh	25		
				Row crops	22				
				Herbs	1				
Composite meal	3								
Buffet meal	10								
Beverages	5								
Unknown	9								
Total	102		70		64		45		8

The table presents only category levels for which there are presently records available in the catalogue. For a complete list of all levels available for reporting see Appendix 1.16.

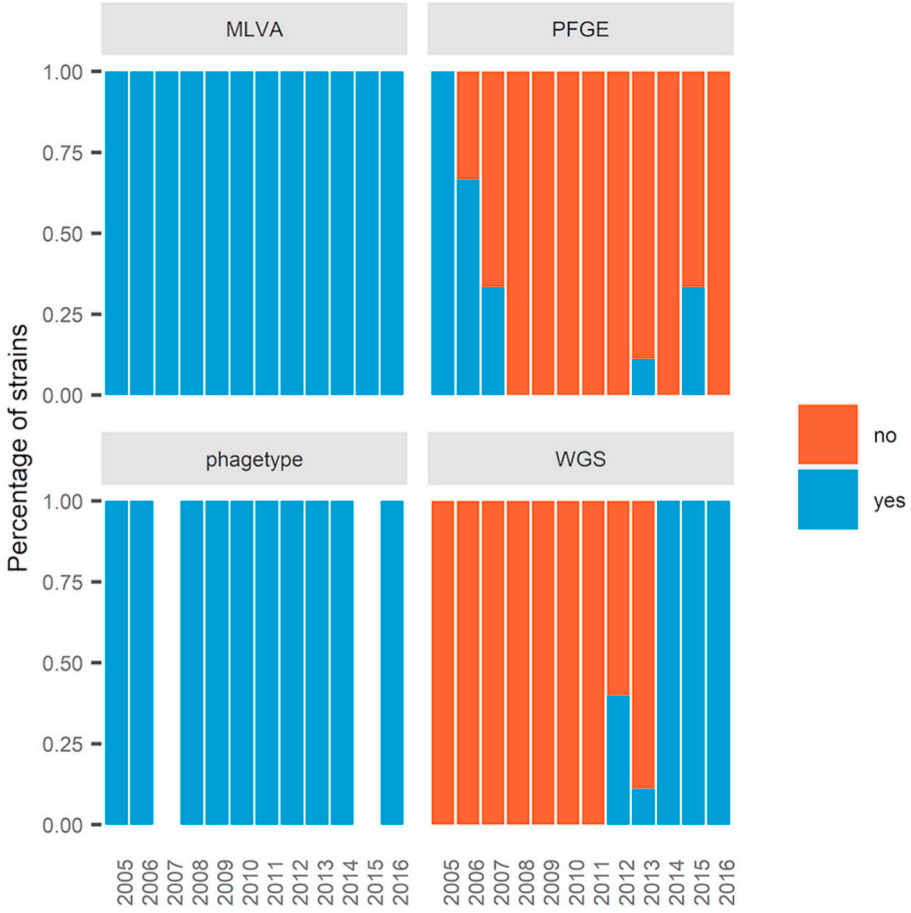

Fig. 2. Typing methods used in foodborne outbreaks of *Salmonella enterica* subsp. *enterica* over time. Percentage of *Salmonella enterica* subsp. *enterica* strains typed by MLVA, PFGE, Phagetyping and WGS by year, between 2005 and 2016, in Denmark.

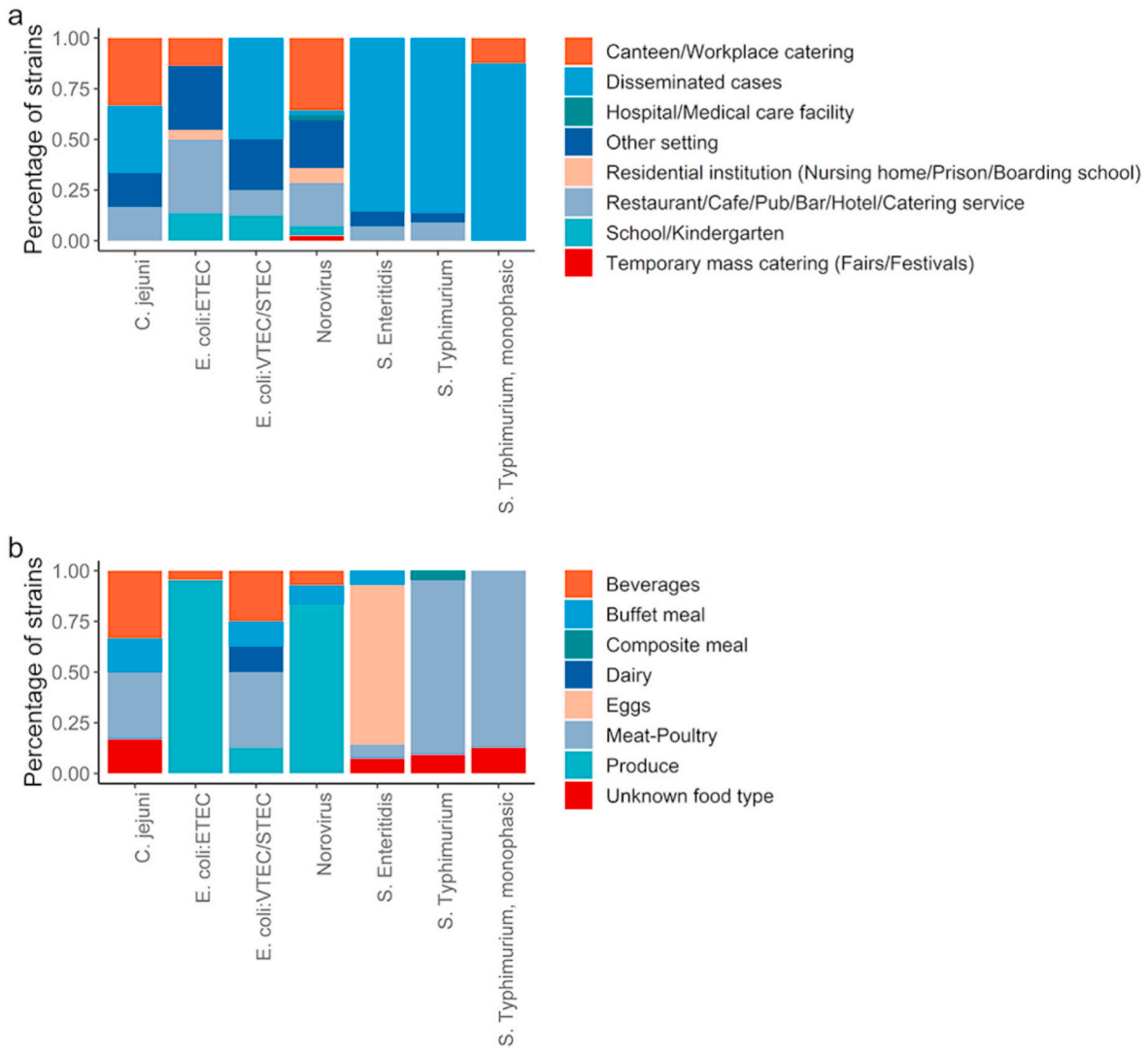

Fig. 3. Outbreak settings and major food sources of most frequent causative agents. Relative distribution of outbreak settings (*a*) and food sources (*b*) for the most common causative agents of foodborne outbreaks occurred in Denmark, between 2005 and 2016.

dropdown menu for the user to choose from. The list of protocols was built based on those used in a Danish setting and it can be reviewed and updates requested by the users as different protocols are applied; *(iii)* whenever applicable, a specific nomenclature format was selected as the GFI standard for reporting of a given typing method, in order to ensure data harmonization and data comparability. For Pulsed-field Gel Electrophoresis (PFGE) and whole genome sequencing (WGS) the user can report whether the outbreak isolate was PFGE typed and with which protocol, and whether it was whole genome sequenced. A PFGE profile or WGS sequence can be easily accessed by users by either contacting the member who reported the outbreak or by consulting the data and resources attached to the record.

Food source related attributes were categorized based on a food categorization scheme from the Interagency Food Safety Analytics Collaboration (IFSAC). The scheme is a five-level hierarchy and its categories take into account the implication of different processing, preparation and consumption practices on the microbiological environment of food (Richardson et al., 2017). A detailed description of the food categorization scheme is presented in Appendix A16. When reporting data to the catalogue, food sources must be assigned to the most specific possible food category.

Tables 2–4 summarize the outbreaks entered in the catalogue at the time of its public launch. The catalogue included at the time of writing 102 FBO with 166 outbreak stains, which originated 13,466 disease cases spanning over 12 years (Table 2). *Salmonella enterica* subsp. enterica, *Norovirus* and Enterotoxigenic *Escherichia coli* (ETEC) were the most common causative agents. Reported typing results across all causative agents included serotype, PFGE type, phage type, multiple-locus variable number tandem repeat analysis (MLVA) profile, multi-locus sequence type (MLST), WGS, antimicrobial resistance (AMR) profile, virulence profile and genogroup (Table 3).

In Table 4 the distribution of food sources causing the reported outbreaks is presented. The major food type 'meat-poultry' was reported in 28% (29/102) of the outbreaks, which were mainly caused by beef, 28% (8/29), and pork, 55% (16/29). Only 10% (3/29) of those outbreaks were caused by poultry. The major food type 'produce' was reported in 37% (38/102) of the outbreaks. Of chose 24% (9/38) were caused by fruits and 76% (29/38) by vegetables.

The R studio notebook that is available to all VRE members in GFI's shared workspace was used to illustrate the visualization of typing methods, outbreak settings, epidemiology and food sources associated with the most common causative agents (*Salmonella* Enteritidis, *Salmonella* Typhimurium (including monophasic), *Escherichia coli* (ETEC and Shiga toxin/verocytotoxin-producing (VTEC/STEC)), *Norovirus* and

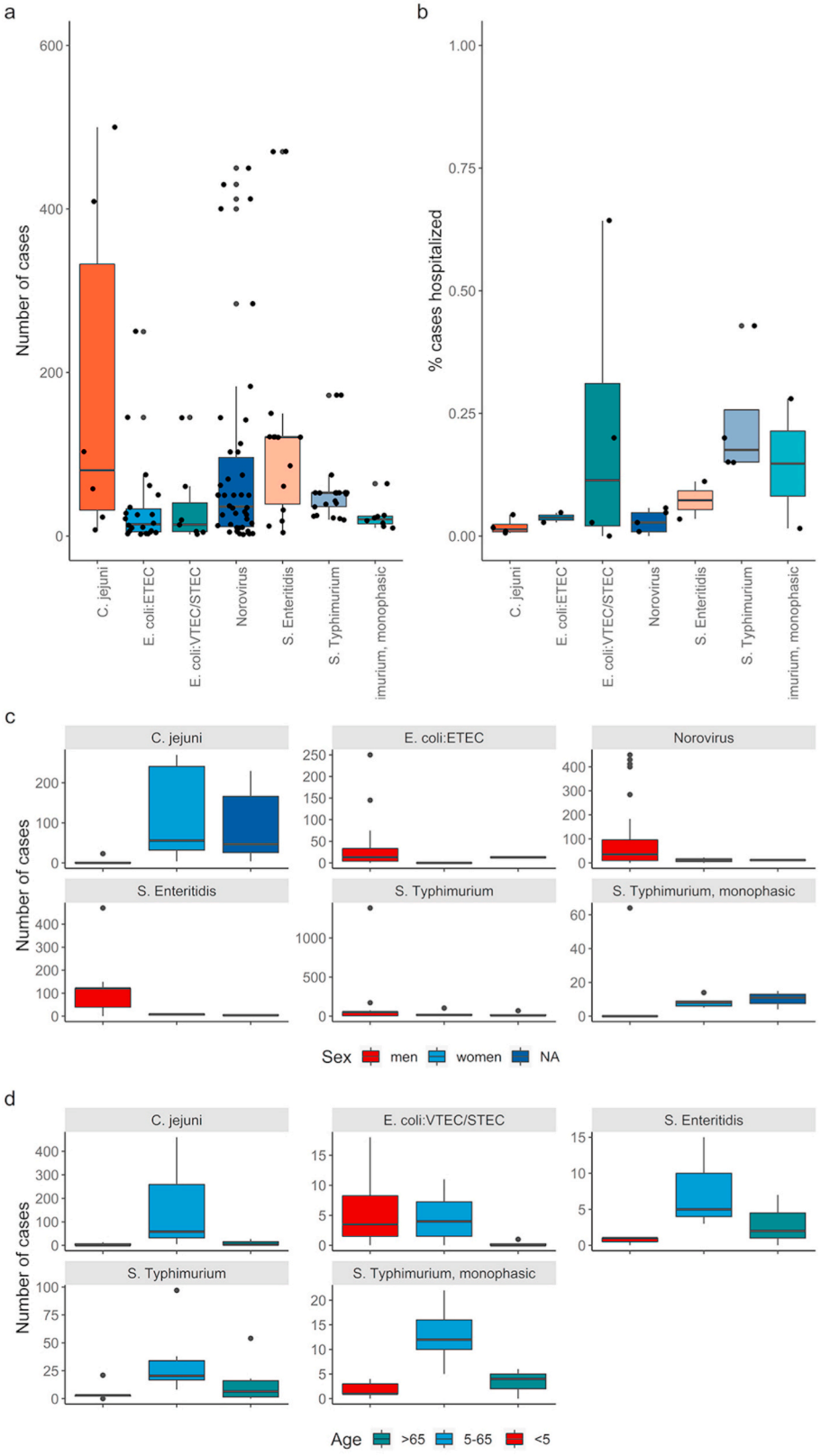

Fig. 4. Epidemiology of outbreaks caused by the most frequent causative agents. Distribution of the number of cases per outbreak (*a*), percentage of cases hospitalized per outbreak (*b*), number of cases by sex (*c*) and number of cases by age class (*d*) for the most common causative agents of foodborne outbreaks occurred in Denmark, between 2005 and 2016.

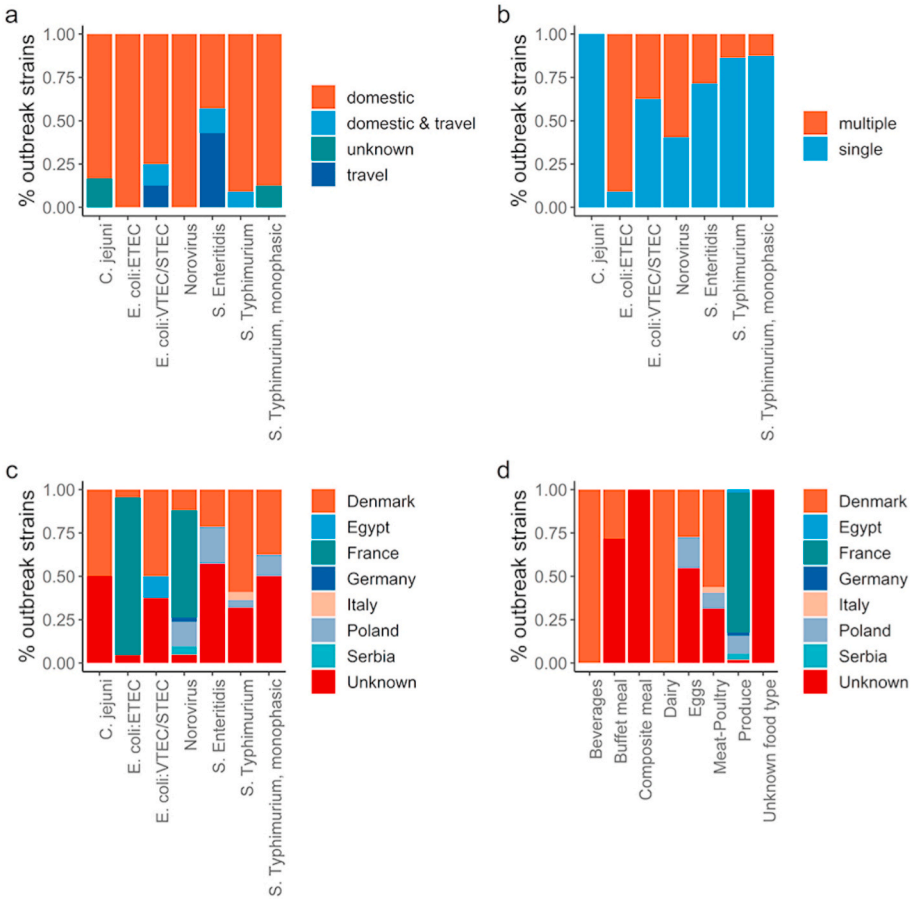

Fig. 5. Country of origin of most frequent causative agents and food sources. Relative distribution of travel- and domestic-related outbreaks (*a*) and of multiple- and single-country outbreaks (*b*), and relative distribution by country of origin (*c*) of the most common causative agents of foodborne outbreaks occurred in Denmark, between 2005 and 2016. Relative distribution by country of origin of the food sources of foodborne outbreaks occurred in Denmark, between 2005 and 2016 (*d*).

Campylobacter jejuni) (Figs. 2–6). It contains explanatory comments that clarify the different chunks of code, so it can be used by members who are non-proficient in R language. It can be executed in its present version on existing or future datasets extracted from the catalogue data, and it can be adapted by any VRE member to different analysis needs.

Fig. 2 shows a clear increasing trend in whole genome sequencing of *Salmonella enterica* strains after 2012, with all strains being sequenced from 2014 onwards. Despite this trend, all strains continued to be MLVA-profiled. Fig. 3 shows the distribution of outbreak settings and major food sources for each causative agent. Compared to other agents, outbreaks of the three most common *Salmonella* serotypes were more often associated to disseminated cases (Fig. 3a). It is also obvious that for some causative agents there has been a predominant major food source of outbreaks – 'meat-poultry' for *Salmonella* Typhimurium, 'eggs' for *Salmonella* Enteritidis, and 'produce' for ETEC and Norovirus – whereas for other agents, the distribution of food sources is more heterogeneous (*Campylobacter jejuni* and VTEC/STEC) (Fig. 3b).

The case epidemiology of each causative agent is presented in Fig. 4. The two agents with a generally low number of cases per outbreak, VTEC/STEC and *Salmonella* Typhimurium (Fig. 4a), were responsible for outbreaks with high percentages of cases hospitalized (Fig. 4b), opposed to e.g. Norovirus which produced outbreaks with more cases but a low percentage of hospitalization. The distribution of cases by sex (Fig. 4c), shows that ETEC, Norovirus and *Salmonella* Enteritidis had a higher median number of cases among men, whereas *Campylobacter jejuni* and *Salmonella* Typhimurium monophasic had a higher median number of cases among women. Such differences could be confounded by the setting in which outbreaks occurred and a possible predominant presence of men or women, or the food sources involved and their eventual higher consumption by men or women, and thus it does not necessarily indicate an association between a specific sex and a causative agent. Further investigation is needed in order to determine the significance and cause of the observed differences. The same consideration applies to the distribution of cases by age category for each agent (Fig. 4d). For example, the high median number of VTEC/STEC cases among young children could tend to be attributed to an outbreak caused by unpasteurized milk in a 'school/kindergarten' (record *VTEC_STEC_DK_2016_001* in the catalogue). However, that outbreak caused a total of 5 cases in this age category, whereas an outbreak of disseminated cases caused by organic cured beef sausage caused 18 cases in young children (record *VTEC_STEC_MultiCountry_2007* in the catalogue).

Among the most frequent causative agents, *Salmonella* Enteritidis had the highest percentage of travel-related outbreak strains (Fig. 5a), and ETEC the highest percentage of strains associated to multiple-country outbreaks (Fig. 5b). Outbreaks of ETEC and Norovirus were mostly caused by strains originating in foreign countries, and in half or more of the outbreaks of *Campylobacter jejuni*, *Salmonella* Enteritidis and the monophasic variant of *Salmonella* Typhimurium the country of origin could not be determined (Fig. 5c). Determining the country of origin of outbreak strains was particularly challenging when the food source was a buffet meal, a composite meal or eggs (Fig. 5d). Whereas eggs could justify why the origin of more than half of the *Salmonella* Enteritidis strains was unknown, buffet- and composite-meals were not predominantly associated with any of the three causative agents (Fig. 3b).

The Sankey diagrams in Fig. 6 show the distribution of *Salmonella enterica* outbreak cases for the three most common serotypes, split by the three first levels of food source categorization, sequentially: food type, food subtype, food subtype variety. This sort of visualization is particularly useful when data is reported at different levels of detail, allowing to investigate both the distribution of records within each level individually and the relationship between sequential levels. Fig. 6a shows

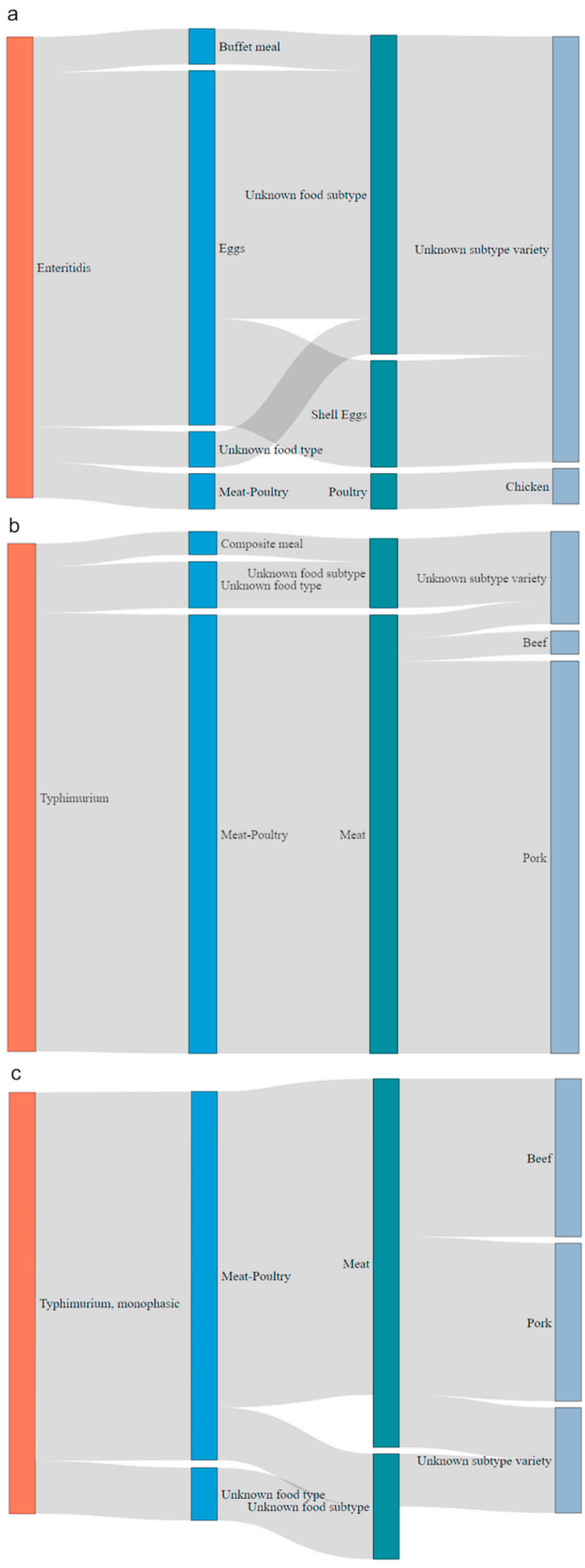

Fig. 6. Distribution of food sources of three *Salmonella enterica* subsp. *enterica* serotypes.
Sankey diagrams showing the distribution of food sources for foodborne outbreaks occurred in Denmark, between 2005 and 2016, by S. Enteritidis (*a*), S. Typhimurium (*b*) and S. Typhimurium monophasic (*c*).

that the largest proportion of *Salmonella* Enteritidis outbreaks could be identified as originating in eggs, but that the majority of the egg product subtypes was either unknown or unreported. Diagrams 6b and 6c show a clear difference in the proportions of meat-related outbreaks that were attributed to pork and to beef, with *Salmonella* Typhimurium outbreaks being predominantly caused by pork, and its monophasic variant having a more even distribution between beef- and pork-associated outbreaks.

The current membership policy of GFI is 'restricted', i.e. new members must request the VRE managers to join. GFI membership is currently managed by the Technical University of Denmark. In order to request membership, users must first create an account in D4Science.[4] Membership can then be requested at https://services.d4science.org/web/foodborneoutbreak.

4. Discussion

The data catalogue has been populated with records from Danish foodborne outbreaks to better illustrate the potential of GFI to new members and to encourage their contribution with their own records. We expect that the perception of the catalogue's structure and of the opportunities offered by other features of the virtual research environment are enhanced by providing the users with real outbreak records from a range of causative agents, typing methods and food sources. However, community support is one of the main challenges for the sustainability of any VRE (Candela et al., 2013).There is therefore the need for an engaging dissemination strategy, including demonstration of GFI features to the target community (i.e. international actors in foodborne outbreak investigation), and their motivation to contribute with data reporting. The potential of GFI is expected to grow significantly with the expansion of the content of the catalogue in terms of number of reported outbreaks and their diversity (e.g. geographical origin). Several VREs have been successfully implemented to date in the e-infrastructure of D4Science to serve a wide variety of scientific scenarios (Assante et al., 2019c). GFI is among the first implemented in the area of Food and Agriculture and therefore it is expected to attract many interested users. Users from countries where international network-wide communication is already practiced during outbreak investigation may show more resistance to join GFI, for considering that it offers no added value. However, GFI offers the opportunity to connect several presently distinct networks, into a unique, multidisciplinary (e.g. across causative agents), international community. The establishment of such a collaborative community is one of the present major challenges in the field of food safety.

An additional sustainability challenge is the attraction of further funding for the maintenance and continuous adaptation of the VRE to its community's needs (Candela et al., 2013). We believe that by introducing GFI and its catalogue to international research consortia studying foodborne diseases, we might identify potential partners interested in populating the catalogue and eventually performing future developments in the VRE. Research projects which aim at capacity building across countries in the area of food microbiology are of particular interest. Future developments of GFI may include features that enable private chats and conference calls among members. Such features will enhance the utility of GFI for its community, for example by enabling straightforward, instant creation of dedicated communication channels across institutes/countries during an outbreak investigation.

The main objective of GFI and its data catalogue is to facilitate foodborne outbreak investigations worldwide, by supporting risk based sampling strategies during traceback investigations. In practice, once an outbreak strain is identified and typed, it will be possible to search the data catalogue for previous outbreaks caused by the same or a similar causative agent, worldwide. The user will have access to both detailed records and the contact of the person(s) who reported outbreaks - both

[4] D4Science homepage: https://services.d4science.org/home.

may be of special relevance to help in an ongoing outbreak investigation. The result of a targeted catalogue search may indicate for example types of food previously identified as an agent source, a pathogen subtype involved in a similar age/gender case distribution, or the identification of the occurrence of a specific pathogen in a certain food product in a country from which a suspect food product is imported. All of these inferences may help directing an outbreak investigation. Once kept alive and up to date, GFI may also assist in the identification of emerging food safety threats, worldwide. Overall, GFI is a platform to encourage and facilitate international collaboration and data exchange between foodborne outbreak investigators.

5. Conclusion

We developed a data catalogue of foodborne outbreaks, to be shared and populated among an international community of outbreak investigators. The catalogue includes for every record detailed information of causative agent typing results, epidemiological data and the food source of the outbreak. It can accommodate in a harmonized way most common agents and food sources involved in foodborne infections, and reporting of a wide range of typing methods. It is hosted in the GFI virtual research environment, which offers a shared workspace, a social networking area, and open-source software for data visualization and analysis on the cloud. GFI hence represents a solution to the challenge of establishing a global collaborative community of actors in foodborne outbreak investigation, which accommodates members with various microbial typing practices.

Funding sources

This work was supported by the European Union Research and Innovation programme Horizon 2020 [grants nr 643,476 and nr. 731,001].

CRediT authorship contribution statement

Ana Sofia Ribeiro Duarte: Conceptualization, Project administration, Supervision, Formal analysis, Visualization, Writing - original draft. **Cecilie Liv Nielsen:** Investigation, Data curation, Conceptualization, Formal analysis, Writing - original draft. **Leonardo Candela:** Software, Writing - review & editing. **Lars Valentin:** Software, Writing - review & editing. **Frank M. Aarestrup:** Funding acquisition, Conceptualization, Writing - review & editing. **Håkan Vigre:** Project administration, Conceptualization, Writing - review & editing.

Declaration of competing interest

The authors have no competing interests to declare.

6. Acknowledgements

We thank Mathias Filter and Taras Günter (Federal Institute for Risk Assessment, Germany) for their support in the implementation of the VRE.

We thank Steen Ethelberg, Katrine Grimstrup Joensen, Mia Torpdahl and Eva Møller Nielsen (Statens Serum Institute, Denmark) for their assistance in data collection and data curation.

We thank Sophie Honoré (Technical University of Denmark) for introducing the outbreak records into GFI.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2020.107623.

References

Anon. (2006). *Annual report on zoonoses in Denmark 2005*. Copenhagen, Denmark: The Ministry of Family and Consumer Affairs. ISSN: 0909-4172.

Anon. (2012). *Annual report on zoonoses in Denmark 2011*. National Food Institute, Technical University of Denmark. ISSN: 0909-3837.

Anon. (2013). *Annual report on zoonoses in Denmark 2012*. National Food Institute, Technical University of Denmark. ISSN: 1600-3837.

Anon. (2017). *Annual report on zoonoses in Denmark 2016*. National Food Institute, Technical University of Denmark. ISSN: 1600-3837.

Assante, M., Boizet, A., Candela, L., Castelli, D., Cirillo, R., Coro, G., et al. (2019a). Realising a science gateway for the agri-food: The AGINFRA PLUS experience. In *11th international workshop on science gateways (IWSG 2019)*, 12-14 June 2019.

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., Panichi, G., Perciante, C., & Sinibaldi, F. (2019b). The gCube system: Delivering virtual research environments as-a-Service. *Future Generation Computer Systems, 95*, 445–453. https://doi.org/10.1016/j.future.2018.10.035.

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., et al. (2019c). Enacting open science by D4Science. *Future Generation Computer Systems, 101*, 555–563. https://doi.org/10.1016/j.future.2019.05.063.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2007). *Knime: The Konstanz information* miner. In W. Gaul, M. Vichi, & C. Weihs (Eds.), *Studies in classification, data analysis, and knowledge organization*. Springer. ISSN: 1431-8814.

Candela, L., Castelli, D., & Pagano, P. (2013). Virtual research environments: An overview and a research agenda. *Data Science Journal, 12*. https://doi.org/10.2481/dsj.GRDI-013. GRDI75-81.

ECDC. (2019). The European surveillance system (TESSy) Accessed at https://ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy.

EFSA & ECDC. (2015). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014. *EFSA Journal, 13*(12), 4329. https://doi.org/10.2903/j.efsa.2015.4329.

Gopinath, G., Hari, K., Jain, R., Mammel, M. K., Kothary, M. H., Franco, A. A., et al. (2013). The pathogen-annotated tracking resource network (PATRN) system: A web-based resource to aid food safety, regulatory science, and investigations of foodborne pathogens and disease. *Food Microbiology, 34*(2), 303–318. https://doi.org/10.1016/j.fm.2013.01.001.

Gormley, F. J., Little, C. L., Rawal, N., Gillespie, I. A., Lebaigue, S., & Adak, G. K. (2011). A 17-year review of foodborne outbreaks: Describing the continuing decline in England and Wales (1992-2008). *Epidemiology and Infection, 139*(5), 688–699. https://doi.org/10.1017/S0950268810001858.

Guzman-Herrador, B., Vold, L., Berg, T., Berglund, T. M., Heier, B., Kapperud, G., Lange, H., & Nygård, K. (2016). The national web-based outbreak rapid alert system in Norway: Eight years of experience, 2006-2013. *Epidemiology and Infection, 144*(1), 215–224. https://doi.org/10.1017/S095026881500093X.

Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M., & Scheutz, F. (2015). Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *Journal of Clinical Microbiology, 53*(8), 2410–2426. https://doi.org/10.1128/JCM.00008-15.

Krause, G., Altmann, D., Faensen, D., Porten, K., Benzler, J., Pfoch, T., et al. (2007). SurvNet electronic surveillance system for infectious disease outbreaks, Germany. *Emerging Infectious Diseases, 13*(10), 1548–1555. https://doi.org/10.3201/eid1310.070253.

Marler Clark LLP, P. (2018). Foodborne illness outbreak database Accessed at http://www.outbreakdatabase.com.

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., et al. (2017). PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveillance, 22*(23), 30544. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544.

Peters, T., Bertrand, S., Björkman, J. T., Brandal, L. T., Brown, D. J., Erdõsi, T., et al. (2017). Multi-laboratory validation study of multilocus variable- number tandem repeat analysis (MLVA) for *Salmonella enterica* serovar Enteritidis. *Euro Surveillance, 22*(9), 30477. https://doi.org/10.2807/1560-7917.ES.2017.22.9.30477, 2015.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, L. C., Bazaco, M. C., Parker, C. C., Dewey-Mattia, D., Golden, N., Jones, K., et al. (2017). An updated scheme for categorizing foods implicated in foodborne disease outbreaks: A tri-agency collaboration. *Foodborne Pathogens and Disease, 14*(12), 701–710. https://doi.org/10.1089/fpd.2017.2324.

RIVM. (2018). Databases Accessed at https://www.rivm.nl/en/noronet/databases.

RStudio Team. (2016). *RStudio*. Boston, MA: *Integrated Development for R*. RStudio, Inc.. http://www.rstudio.com/

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the public GenomeTrakr database for foodborne pathogen traceback. In A. Bridier (Ed.), *Methods in molecular biology: Vol.1918. Foodborne bacterial pathogens* (pp. 201–212). New York: Humana. https://doi.org/10.1007/978-1-4939-9000-9_17.

Wielinga, P. R., Hendriksen, R. S., Aarestrup, F. M., Lund, O., Smits, S. L., Koopmans, M. P. G., et al. (2017). Global microbial identifier. In X. Deng, H. den Bakker, & R. Hendriksen (Eds.), *Applied genomics of foodborne pathogens. Food microbiology and food safety* (pp. 13–31). Cham: Springer. https://doi.org/10.1007/978-3-319-43751-4_2.

World Health Organization (WHO). (2015). *WHO estimates of the global burden of foodborne diseases: Foodborne disease burden epidemiology reference group 2007-2015*. Geneva: World Health Organization. ISBN: 9789241565165.