Journal of Cleaner Production 470 (2024) 143198

Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Interlinking environmental and food composition databases: An approach, potential and limitations

Cédric Furrer[a,*], Daniel Sieh[b], Anne-Marie Jank[b], Grégoire Le Bras[c], Moritz Herrmann[a], Alba Reguant-Closa[a], Thomas Nemecek[a,**]

[a] Agroscope, Life Cycle Assessment research group, CH-8046, Zurich, Switzerland
[b] Themakers.ai GmbH, Chausseestrasse 8a, DE-10115, Berlin, Germany
[c] The Makers Food GmbH, Schreinerstrasse 58, DE-10247, Berlin, Germany

ARTICLE INFO

Handling Editor: Maria Teresa Moreira

Keywords:
Data interlinkage and standardization
Life cycle inventory databases
Food composition databases
Agribalyse
EuroFIR
Nutritional LCA

ABSTRACT

Connecting entries from environmental and nutritional databases of food products is needed to identify sustainable food options. To overcome the hurdle for a successful data standardization, this paper aimed to explore the general structure of food composition and life cycle inventory (LCI) databases and to provide a semi-automatized approach on how to successfully interlink data of two example databases.

The structure, the data availability and accessibility of food entries (FI) from the LCI database Agribalyse and selected food composition EuroFIR databases were analyzed. Harmonized food names from LanguaL™ codes from food classification systems were gathered and validated manually in order to use as descriptors to tag database entries in an automated way.

Both databases, EuroFIR and Agribalyse, provided sufficient amount of meta data to interlink FI with the standardization approach proposed. Information on food name, food specification, food processing and the type of production was used for data interlinkage purposes. Manual validation of data interlinkage showed that two out of a sample of 54 entries were found to have incorrectly assigned descriptors.

Agreeing on common principles (e.g., use of a specific classification system, common database formats) and improving meta data availability would facilitate database interlinkage and improve both, accuracy and efficiency. Developing solutions to increase meta data availability and accessibility of FI in food databases should become a key area of research in order to transition into more reliable database connection systems.

1. Introduction

Driving food systems more sustainable is widely discussed due to the various dramatic implications of the intensification of food production on environment (Alemu, 2022; Sirdey et al., 2023; Zhu et al., 2023). Efforts are increasing to direct food production towards products with low environmental impacts and high nutritional value (Mazac et al., 2023; Poore and Nemecek, 2018). This is especially true for products with high environmental impacts such as meat (Van Mierlo et al., 2022). To achieve such a transition of the food system, comprehensive knowledge and data about both, the environmental and the nutritional dimensions of foods is equally needed to identify suitable food options (Hallström et al., 2018; McLaren et al., 2021). Combining analyses of the food composition (FC) with a life cycle assessment (LCA) for food products, also called nutritional LCA (nLCA), offers the possibility to optimize both dimensions at the same time and allows for in-depth assessments (Heller et al., 2013; McLaren et al., 2021; Saarinen et al., 2017). It is expected that the method of nLCA will be used more widely for future studies, in order to assess food products sustainability more effectively.

Whereas most of the databases consist either purely of FC data or purely of life cycle inventory (LCI) data for foods, there is a lack of publicly available databases including both data combined. To our knowledge, there is only one connected database, matching entries from the CIQUAL database (the French food composition table) to entries of Agribalyse, the French LCI database for the agriculture and food sector (Anses, 2020; Asselin-Balençon et al., 2022). However, several other food composition databases (FCDB) (e.g., EuroFIR, USDA) and LCI

* Corresponding author.
** Corresponding author.
E-mail addresses: cedric.furrer@agroscope.admin.ch (C. Furrer), thomas.nemecek@agroscope.admin.ch (T. Nemecek).

https://doi.org/10.1016/j.jclepro.2024.143198
Received 23 October 2023; Received in revised form 30 June 2024; Accepted 17 July 2024
Available online 18 July 2024

databases with food entries (e.g., ecoinvent, AgriFootprint or World Food LCA Database) exist and could be connected (Becker et al., 2008; Becker et al., 2007; European Food Information Resource (EuroFIR), 2023b; LanguaL, 2023; Nemecek et al., 2019; van Paassen et al., 2019; Wernet et al., 2016; Westenbrink et al., 2019). Such interlinkage of data is expected to yield more comprehensive nLCA simply because the availability of matched data will be enhanced. Incorporating more interlinked data into future analyses is expected to increase overall data quality and yield more accurate results.

Many attempts have been made to connect and standardize data of food items (FI) to a common language in order to interlink data from different sources. The International Network of Food Data Systems (FAO/INFOODS) has also initialized guidelines for food standardization and harmonization (Charrondiere et al., 2016; Stadlmayr et al., 2012). Koroušic Seljak et al. (2018) proposed a computer supported food matching algorithm which automatically assigned codes from the EFSA FoodEx2 classification system. Assigned codes were validated manually by experts and used as trained datasets for further assignments. Additionally, a quality rating was given for each assigned code. Results showed that the quality of assignment overall was sufficient to good with some exceptions. Isiprova et al. (2017) analyzed normalization of nutritional parameters (e.g., energy or dry matter) from two databases based on several text similarity measures and a method including "Part of Speech" tagging probability weighting. The aim was to map the data to a food domain ontology. They showed a correct linking for 167 instances with an additional 23 no-match instances with an overall accuracy of 87.9%. Similarly, Eftimov et al. (2017) tried to automate the assignment of FoodEx2 classification codes to FI names of different FCDB using machine learning and natural language processing (NLP) approaches. Combining both approaches with post-processing rules yielded 79% correctly classified and described instances. The remaining 21%, incorrectly described, were mainly due to insufficient FoodEx2 codes or incomplete food description. Wolongevicz et al. (2010) created an algorithm to match USDA food codes to their food frequency questionnaire data. However, the authors did not describe the details of the algorithm thoroughly. Extra manual validation by experts was needed to validate matching entries. Martinez-Victoria et al. (2015) used the information system developed by Farran Codina (2004) for the LanguaL™ codification of foods. The information system "*added basic information about the food […]*". However, it is not described in detail how the matching of food was achieved. Although the approaches proposed by Isiprova et al. (2017) and Eftimov et al. (2017) yielded promising results, a fully automated process for food coding was shown to be difficult to achieve because of partial incorrect code assignment. Wolongevicz et al. (2010) and Martinez-Victoria et al. (2015) did not further elaborate on the validation of the matching.

Others relied purely on a manual match combining data of FI on a case-by-case decision including a substantial amount of time for data processing (Bertoluci et al., 2016; Broekema et al., 2019; Gurinović et al., 2016; Hinojosa-Nogueira et al., 2021). In all cases it became particularly clear, that manual work remained an essential part to ensure data validity independently of the level of automatization. Additionally, due to limitations in the accessibility of software and models or the lack of proper documentation of the matching procedure, it was difficult to follow and reproduce the procedures proposed. It is assumed that many databases in the food sector are still standardized manually mainly due to the complexity of such a task.

Whereas fully automated approaches tend to be efficient, transparent and reproducible, they might be more inaccurate than manual matching. Manual matching of food data in contrast generally shows to link data more precisely but is more time-consuming. Additionally, results from manual matching are not as easily reproducible as with fully automated procedures if not documented properly. Therefore, this paper investigates a semi-automated approach which aims at combining manual and automated matching in order to increase accuracy while keeping the amount for manual work at a reasonable level.

Due to the relevance of the topic, several areas were identified suitable for further investigation. First, to identify and summarize previous knowledge and challenges. This included for example the aspect, that FCDB and LCI databases differ in how data is presented to the user, which (meta)data is provided for FI and which data file format is used to access datasets. The available amount and the difficult accessibility of meta data was seen as a major limitation hampering the process for a successful standardization. Due to the many aspects to be considered for data connection, already available procedures were identified being only partly user friendly when it comes to implementation.

The objective of this study was to cover three major relevant topics within the field of nutritional and environmental database standardization. First, explore and analyze the general structure of nutritional, more concretely FCDB, and LCI databases in order to understand data organization and availability of meta data. Secondly, develop a specific and easily-applicable semi-automated procedure for the connection of data from both databases. Ultimately, highlight and discuss the current issues and limitations regarding the connection of data from FCDB and LCI databases.

Given the objectives of this study, the following research questions arise.

1. Do LCI and FCDB databases (Agribalyse and EuroFIR, respectively) provide enough (meta) data to successfully describe and interlink food entries between environmental and nutritional databases?
2. Can manual and automatized matching procedures be coupled into a semi-automatized standardization approach to facilitate data interlinkage?

2. Methods

Database users need to locate, identify and extract nutritional and environmental data of FI together in order to perform nLCA. Preprocessing of data is needed to ensure a successful standardization of data from individual databases. Accessing and interlinking data in a simple way was given the highest priority. Four major key themes have been identified relevant by end users regarding the work with databases and were considered especially relevant for this study (Clancy et al., 2015). Apart from the procedure for interlinkage, this included 1) database structure and technological components; 2) food classification; 3) data accessibility and 4) data availability.

The relevant research areas covered in this study are summarized in Fig. 1. In a first step, the general structure of LCI and FCDB databases has been analyzed and differences (especially what data is available) were highlighted. Results from step 1 were then used together with information from literature to identify the relevant meta data which could be used for database interlinkage. Assessing the availability and accessibility of meta(data) of datasets as well as providing a broad overview of how the datasets were structured yielded relevant insights which contributed to key areas 1, 3 and 4 according to Clancy et al. (2015). The analysis of the meta(data) also revealed the status of food classification of FI and provided insights into key area 2. In step 3, a procedure for a successful harmonization and classification of database entries was elaborated. Finally, a concept for semi-automatic data linkage was created and applied to a selection of entries from Agribalyse and EuroFIR in a case study. The outcome of the interlinkage of a selected sample of FI was manually validated by comparing the found database entries individually. A further analysis of nutritional values and environmental impacts of interlinked FI was not part of the study.

2.1. Food composition and life cycle inventory databases

FCDB provide data on the nutritional content and quality of foods (Delgado et al., 2021). The International Network of Food Data Systems (INFOODS) managed by the Food and Agriculture Organization of the United Nations (FAO) provides an overview of available databases

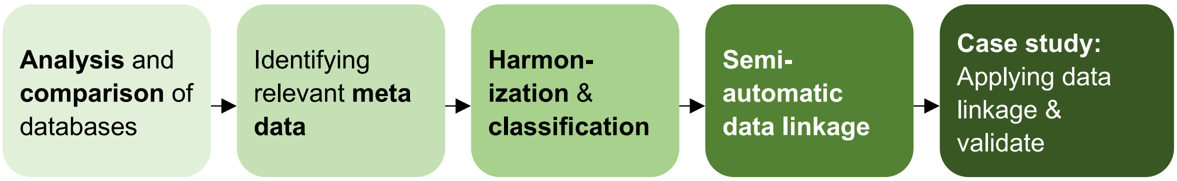
Analysis and comparison of databases
Identifying relevant meta data
Harmon-ization & classification
Semi-automatic data linkage
Case study: Applying data linkage & validate

Fig. 1. The different methodological steps for database analysis and interlinkage considered in this study.

(International Network of Food Data Systems (INFOODS), 2023). Nutritional databases or food composition tables from more than 100 regions are currently available worldwide differing in the amount of data available, in data quality and in data structure. EuroFIR provides food composition data for many food products with a focus on European countries. EuroFIR data was selected for the study because 1) data is provided in easily accessible Excel sheets, 2) nutritional values have been harmonized and standardized to consistent component names and 3) data is available for 31 different countries. EuroFIR data of the countries Switzerland, France, Estonia, Slovenia, Denmark and the United Kingdom was considered and was chosen as an approximation for the region of Europe (Becker et al., 2008; Becker et al., 2007; European Food Information Resource (EuroFIR), 2023b; LanguaL, 2023; Westenbrink et al., 2019). Additionally, data of these countries were the data most up to date available. Licensed data from EuroFIR consisted of a total of 11911 individual FI. Information on up to 187 nutritional parameters per database were provided (Table 1).

EuroFIR data was complemented with additional meta data from the component and the value type thesaurus provided online (European Food Information Resource (EuroFIR), 2023a). The component thesaurus provided a classification scheme for nutritional parameters (also referred to as components in EuroFIR). Thus, it was possible to group nutritional parameters such as e.g., "Glucose" and "Fructose" in the same category ("Monosaccharides"). Additionally, the value type thesaurus provided information on whether to in- or exclude nutritional parameters with value 0 and how to treat missing values.

To assess the environmental impact of food, life cycle assessment (LCA) is often used because 1) it assesses food products throughout their entire life cycle, usually from cradle-to-grave and 2) it is defined by the norms of the "International Standard Organization" (ISO) No. 14040 and No. 14044 (International Standard Organisation, 2006a, 2006b).

The LCI database Agribalyse version 3.1 has been selected because it is currently the most comprehensive and elaborated LCI database within the food sector (Asselin-Balençon et al., 2022). Agribalyse has been developed by the French government to assess the environmental impacts of foods in France. Thus, data has been constructed and is primarily valid for the geographical boundaries of France. However, certain inventories are imported from other LCI databases and can also be valid for other geographical regions. The database has already been linked to CIQUAL, the French food composition database but not to EuroFIR.

2.2. Database analysis

The databases Agribalyse and EuroFIR were used as example databases to study the generic structure of LCI and FCDB databases. A specific focus was laid on how data in the databases is organized. Data from databases was accessed in its raw format by using Excel or specific LCA software. Database structure was thoroughly examined, available meta data summarized and similarities and differences in meta data between LCI and FCDB highlighted. Meta data was classified as relevant if data was identified as minimal required information for a successful interlinkage of FI.

2.3. Meta data relevance

Available meta data from FI was priorized based on the guidelines for harmonization from FAO/INFOODS. According to Stadlmayr et al. (2012), food items need to be identified by "*its form and preparation*". Thus, all meta data fields providing information in relation to the identification of the food were considered relevant and were therefore selected for data interlinkage. No other meta data than the one directly provided by the datasets was used.

2.4. Data selection

FI representing composite foods were excluded from data interlinkage and only single FI were kept due to complexity. For that purpose, certain EuroFIR food categories (e.g., "Miscellaneous or undefined") were excluded and predefined exclusion terms for filtering were used (e.g., FI names containing "pizza"). Single foods were defined as "*food items […] available in the market, ready for human consumption and requiring either no or minimal preparation before eating*" (McLaren et al., 2021). Additionally to this definition, FI were only considered as single foods if they were 1) "*intended to be blended or processed with other items to make a complex food*" (McLaren et al., 2021) or 2) to be consumed alone (e.g., apple). That means, FI such as "ready-to-eat frozen pizza" were not considered as single foods but rather as composite foods because they are a mixture of various FI even if they would need only minimal preparation before eating. This leads to the definition that composite foods were defined as a mix of multiple single foods (McLaren et al., 2021).

For Agribalyse, single food inventories were mainly found in categories "Agricultural, Food, Consumption mixes" or "Agricultural, Food, Transformation". Selecting inventories of those categories also ensured to interlink inventories with the same life cycle stage, including emissions from agricultural production (e.g., nitrogen or field emissions), from processing of foods (e.g., electricity) and from transports (e.g., on farm transports).

2.5. Harmonization and classification

LanguaL™ has been used to capture, describe and classify FI. The LanguaL™ thesaurus "*provides a standardised language for describing foods, specifically for classifying food products for information retrieval*" (Møller and Ireland, 2018). Additionally, EuroFIR data has already been attributed LanguaL™ codes, which could be used for classification purposes. The main advantage of LanguaL™ is that it uses a set of

Table 1
Amount of available FI and nutritional parameters per country database for licensed EuroFIR data.

	Switzerland	France	Denmark	Slovenia	Estonia	United Kingdom
Amount of FI [n]	1080	3185	1190	405	3164	2887
Maximum amount of nutritional parameters available [n]	38	63	187	191	59	152

harmonized, controlled terms connected to specific codes. This allows a consistent standardization and classification. Individual codes such as “A01DJ” for example always represent one unique FI (here “Apple”). Because LanguaL™ provides a structured categorization of foods, more general data connected to a certain code can be retrieved. Information on the broader food group can therefore be extracted without initially specifying it. Banana (code “A0DQK”) and apple (code “A01DJ”) are automatically classified as “fruits” (A04RK).

2.6. Semi-automatic data interlinkage

The procedure for semi-automatic data interlinkage used harmonized descriptors together with a respective LanguaL™ code to merge FI from different databases. Linking FI by harmonized descriptors was automated using a standardization algorithm built with Python programming language (van Rossum and Drake, 2009).

To successfully assign harmonized descriptors to FI, a connection list was set up manually as a supporting file. The connection list acted as a fundamental integral part because it contained the mapping of relevant terms and synonyms to harmonized descriptors. A descriptor was defined as a word fragment which described an aspect of a food item (e.g., the type of food (e.g., apple or banana) or the processing stage of food (e.g., dried)). Descriptors were classified into a food specific nomenclature which was based on the five most relevant descriptor categories for foods (name, specification, treatment, processing, production system).

Finding descriptors for the connection list was done in a manual and semi-automated way comparing entry names from LCI and FCDB databases with names from the EFSA classification in the LanguaL™ thesaurus. Because checking for descriptors manually involved substantial amount of work, different approaches were evaluated to speed up the process for finding relevant synonyms.

- “Regular expression” (Regex) was used to extract similar words or word patterns of food names. If a certain pattern was not found in a food name, nothing was returned. Regex successfully extracted “apple” from “apples” for example.
- The “Levensthein distance” was used to compute the “alphabetical” distance between two food names returning a “similarity” value (Miller et al., 2009). The method returns a number which indicates the amount of letter replacements needed to convert one word into another word. In the case of “apple” and “apples”, the Levensthein distance is 1 because there is one change needed (adding or removing the ‘s’). Thus, the lower the value, the closer and more related are food names. Foods with “apple” in their name were for example marked as closer to foods with “apples” (distance = 1) in their name instead of “pear” (distance = 4).
- Using sentence embeddings based on Siamese BERT-Networks (SBERT) returned the distance between two food names as a cosine similarity value (Reimers and Gurevych, 2019). Because SBERT tried to assess the “meaning” of a word, it was able to successfully identify “corn” and “maize” as the same food although the food names were completely different.

The names of the codes in the LanguaL™ thesaurus were used as harmonized descriptor names. All relevant synonyms found by Regex, Levensthein distance and SBERT from entries in LCI and FCDB databases were attributed to the harmonized descriptor names if they were appropriate based on manual evaluation.

2.7. Case study

FI from both, Agribalyse and EuroFIR, were tagged with harmonized descriptors from the connection list. FI between the databases were then automatically interlinked by comparing the assigned descriptors. Validation of the procedure was done manually by checking the appropriateness of attributed descriptors. To ensure a representative validation sample, six foods covering the most relevant types were chosen (wheat, vegetables, nuts, meat, cheese and oil).

3. Results

Fig. 2 shows the scheme of the main structure of the FCDB and LCI databases. For LCI databases, an item is referred to as an inventory. Items always (in the case of EuroFIR) or partially (in the case of Agribalyse) refer to a certain food. Both have similarly structured glossaries with two main parts: the meta data, which stores descriptive information about the item as a whole (e.g., food name or unit) and the base structure. Base data in LCI databases quantifies inputs, emissions and outputs from or to the environment from the production of the food along its life cycle and reflects activities from agricultural production, food processing, storage, transportation or food preparation. Base data in FCDB describes the nutritional composition of a food (e.g., sugar content).

A further analysis of the structure of items from Agribalyse and EuroFIR databases is shown in Table 2. Each item contains several fields of meta data. The fields “name”, “unit”, “category” and “country” were found in items of both databases. In comparison to Agribalyse, EuroFIR additionally provided an identifier (ID) and LanguaL™ codes describing the FI. Agribalyse further included fields “data quality” and “included processes”, which indicated additional information about the underlying base data.

3.1. Relevant meta data for interlinkage

Depending on the type of database, either FCDB or LCI database, different meta data is required when standardizing foods and has been summarized in Table 3. Key parameters such as “food name”, “food specification”, “food recipe” and “food processing” are required to uniquely identify the type of food and need to be provided in food databases. Additional information on parameters such as “system boundaries”, “yield”, “country of origin of food” and “production system” are especially important for LCI databases for standardization purposes.

Comparing the required meta data (Table 3) to available meta data from databases (Table 2) shows that key parameters “food name”, “food specification” and “food processing” were provided in field “name” by both Agribalyse and EuroFIR. Additional meta data for those fields could also be extracted using the “LanguaL™ code(s)” in EuroFIR if not yet provided by the field “name”. The general lack of data for “food recipe” shows that standardization of composite FI (e.g., lasagne or pizza) is challenging. Parameter “country of origin of food” is fully available in Agribalyse but not in EuroFIR. Although parameters such as “system boundaries”, “yield” and “production system” are important for LCI databases, there are not always provided in the case of Agribalyse.

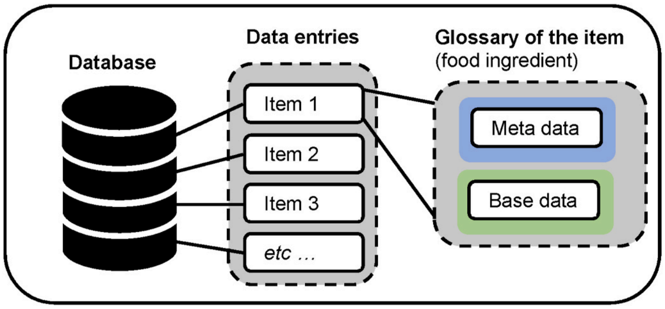

Fig. 2. Scheme of the main structure of FCDB and LCI databases using Agribalyse and EuroFIR as an example. In both cases, databases consist of several entries where each entry is structured into two parts, the meta and base data. Whereas meta data describes the FI in general, base data provides the nutritional or LCA data.

Table 2
Meta data fields for two selected items of Agribalyse and EuroFIR.

Description	Agribalyse	EuroFIR
ID		0211051
Name	Apple, conventional, electric platform, at orchard/kg/FR	Apple, fresh
Unit	Per kilogram	Per 100g edible portion
Category	AGRIBALYSE/Plant production/Fruits	Fruit or fruit product
Country	France	Switzerland
LanguaL™ code(s)		A0833, B1245, G0003, H0003, J0003, K0003 …
Data quality	Technological representativeness = 2, Geographical representativeness = 2, Technological representativeness = 2, Completeness = 2, Precision/uncertainty = 2, Methodological appropriateness and consistency = 2	
Included processes	(1) the processes of soil preparation and cultivation, sowing, weed control, fertilisation, pest and pathogen control, harvest; (2) the machines and shed or surface used to park them; (3) all inputs as seed, fertilizers (mineral and organic), active substances, water for irrigation, fuels as well as the transport to the farm; (4) the direct emissions of the fuel combustion, the abrasion of tyres and the direct emissions on the field.	

3.2. Food harmonization and classification

The classification and harmonization of food has been identified as a fundamental and essential integral part in the development of a database connection system. Food classification systems not only help to group, consistently identify and structure foods but also provide harmonized food names. LanguaL™ has been used as a source which provided information of 11 different systems for food classification (Table 4).

Although the food classification systems were found to be consistent within the same system, they were not fully comparable between each other. Five different codes were for example found for “bread”, where each one belonged to a different system (Table 5). Thus, it is difficult to standardize FI from different sources where different food classification systems were used.

Additionally, it has been found that classification systems differed in the level of detail. In some cases, classification was based upon broad food group(s) only (e.g., vegetables or fruits), whereas in other cases systems classified many foods with a high level of detail (e.g., carrots or salad and not only vegetable). Some systems included many and other few codes (Table 6). The EFSA FoodEx2 was identified as the system with the most codes, followed by the Global Product Classification (GS1 GPC) (Table 6) (EFSA, 2020; Global Standards One (GS1), 2023)). Whereas Agribalyse did not provide LanguaL™ codes for inventories, EuroFIR has implemented LanguaL™ codes from all food classification systems without considering the EFSA FoodEx2 classification system. In comparison to FoodEx2, these classification systems have shown to provide only a limited number of codes (Table 6). Thus, it was not possible to efficiently use the available LanguaL™ codes in the EuroFIR database for data interlinkage.

Table 4
Available food classification systems and their abbreviation in LanguaL™.

Name of food classification system	Abbreviation
CIAA Food Classification for Food Additives	CIAA
Classification of Products of Plant and Animal Origin, European Community	EC
EFSA Food Classification and Description System for Exposure Assessment	EFSA FoodEx2
Eurocode 2 Food Classification	Eurocode2
EuroFIR Food Classification	Eurofir
European Food Groups	EFG
Food Classification for Food Additives	Codex Alimentarius
Global Product Classification	GS1 GPC
U.S. Code of Federal Regulations	US CFR
USDA Standard Reference	USDA SR

Table 5
Available LanguaL™ codes for “bread” for five food classification system and indication whether the food classification system has been used in EuroFIR.

	Food classification system				
	EFG	EFSA FoodEx2	GS1 GPC	EuroFIR	US CFR
LanguaL™ code for “bread”	A0691	A004V	A0943	A0817	A0178
Food classification system used in EuroFIR?	True	False	False	True	True

Table 3
Relevant parameters for interlinkage of FCDB and LCI databases and their availability and accessibility in EuroFIR and Agribalyse.

Parameter	Example	FCDB databases (e.g., EuroFIR)	LCI databases (e.g., Agribalyse)	Additional info
Food name	*“Apple”*, *“Mango”*, etc.	[c] (III)	[c] (III)	Information needs to be extracted from title of a database entry
Food specification	*“Juice”*, *“Oil”*, etc.	[c] (II)	[c] (II)	Information needs to be extracted from title of a database entry. Often inconsistently accessible information (e.g., “sunflower oil” vs. “oil, sunflower”)
Food recipe	Percentage of water added to apple juice	[c] (I)	[c] (I)	Information, if provided, only in base data. Difficult to extract.
Food processing	*“pasteurized”*	[c] (III)	[c] (II)	Information needs to be extracted from title of a database entry
System boundaries	*“at farm”* or *“at processing”*	[a] (I)	[c] (II)	Not always provided in the database entry in Agribalyse
Yield	Yield of apple from agricultural production	[a] (I)	[c] (II)	Information only provided in base data. Difficult to extract.
Country of origin of food	*“Germany”*, *“France”*, etc.	[b] (I)	[c] (III)	
Production system	*“conventional”*, *“organic”*, etc.	[a] (I)	[c] (II)	Information needs to be extracted from title of a database entry

I: not provided; II: sometimes provided; III: fully provided.

[a] little relevant or irrelevant.

[b] moderately relevant.

[c] highly relevant.

Table 6
Amount of FI tagged in EuroFIR grouped by the different food classification systems available.

		Use of classification system in selected EuroFIR	
Food classification system	Number of LanguaL codes available	System used by EuroFIR?	Number of FI tagged with classification system
CIAA	17	True	856
EC	49	True	145
EFSA FoodEx2	4524	False	0
Eurocode2	14	True	1024
Eurofir	120	True	13689
EFG	34	True	3078
Codex Alimentarius	17	True	1718
GS1 GPC	888	True	121
US CFR	183	True	3374
USDA SR	26	True	169

Although codes of the EFSA FoodEx2 classification system are not implemented in EuroFIR, it was identified as the most completed food classification system. EFSA codes therefore were used together with harmonized food names in the connection list.

3.3. Semi-automatic data interlinkage

Figs. 3 and 4 illustrate the concept developed to interlink data from LCI and FCDB databases semi-automatically. Setting up a consistent and food-specific nomenclature (connection list) for structuring and linking food data has been identified as a key aspect in the development of the standardization approach (Fig. 3). The aim is to keep standardization as simple and automated as possible while assuring the use of all relevant data provided by the databases.

In step 1 and 2 connections to the databases is established and all relevant FI is gathered in a list (Fig. 3). In step 3, composite foods are excluded and only single foods are kept. Connection of composite foods (e.g., pizza) between databases was found to be complex and expected to be unprecise due to limited information (e.g., missing recipe composition, Table 3). Excluding composite foods from the interlinkage and only focusing on single foods (e.g., apple) is expected to reduce the amount of work and facilitate data connection. For that purpose, name fragments (e.g., pizza) and LanguaL™ codes of composite foods are previously defined. FI are then excluded if they either contain a respective name fragment in their name or if LanguaL™ codes (if provided as meta data) match with one of the composite foods. For example, foods with “pizza” or “burger” in the FI name are excluded. As a result, a cleaned list of database entries is obtained (Fig. 3, step 4). Due to different names for food, distinguishing between single and composite foods can be a time intensive process. Missing a clear differentiation between single and composite foods can hamper the process in addition. The current filtering via name fragments identifies 2603 (21.9 %) and 23 (1.7 %) composite foods for EuroFIR and Agribalyse, respectively (Table 7).

The connection list is set up (Fig. 3, step 7) using food names and respective synonyms from the meta data of FI (Fig. 3, step 5) in combination with LanguaL™ names, respective synonyms and codes from the LanguaL™ thesaurus. For each item in the connection list, the information of step 5 and step 6 is summarized in five descriptor categories relevant for the description of food and the purpose of the matching of food data based on findings in Table 3. The category “name” describes the basic ingredient name, independent of a further specification of the food item (e.g., the variety). The category “specification” contains a

Table 7
Amount of identified single and composite foods for Agribalyse and EuroFIR.

Database	Single foods (#)	Composite foods (#)	Total (#)
Agribalyse	1298	23	1321
EuroFIR	9308	2603	11911

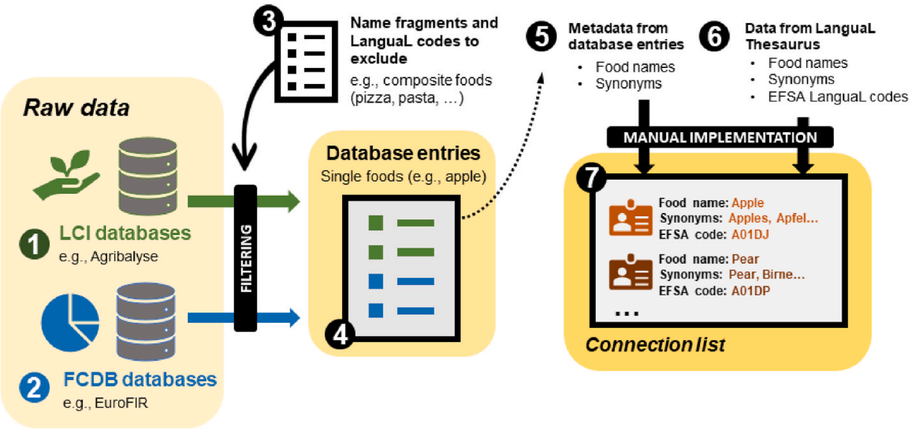

Fig. 3. The different steps needed to form the connection list ensuring a consistent and food-specific nomenclature.

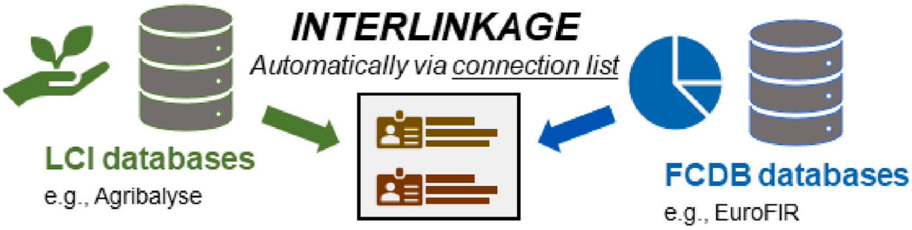

Fig. 4. Scheme of interlinking food items from both databases after having established the connection list. Descriptors in the previously defined connection list are assigned individually to FI of LCI databases on one hand and FI of FCDB databases on the other hand. Interlinkage is achieved if assigned descriptors match both, FI of LCI and FCDB databases.

Table 8
Example of the connection list for one descriptor with the five categories (n/a = not available).

Descriptor category	Harmonized descriptor	EFSA code	Synonym name(s) of descriptor	LanguaL™ code(s) of other food classification systems
Name	chickpea	A00PZ; A0BAV	garbanzo; bengal gram; chick bean; chick peas	A1484; B1172
Specification	iodized	n/a	iodized; with iodine; iodine added	A042R
Treatment	peeled	n/a	w/o peel; peel removed; skin removed; hulled	n/a
Processing	dried	n/a	dehydrated; water removed; dry	J0116; H0138; J0117; J0141; J0170
Production system	organic	n/a	organic prod	Z0253; Z0291; Z0210; Z0213

specification of the food and is applied in order to describe a FI in more detail, if needed. This includes for example a further distinction of the food or the variety (e.g., green bell pepper), the part of the plant or animal used (e.g., seed) or the form it is used if transferred from the original form to another (e.g., oil). The category "treatment" is used for all standard procedures which are applied during preparation of the food (e.g., cooking, drying). "Processing" indicates whether the FI has undergone additional measures enhancing the shelf life (e.g., canning, sterilizing). Lastly, the category "production system" describes how the food is produced (e.g., by means of conventional or organic production).

The main advantage of the connection list is to use relevant terms as harmonized descriptors in order to maintain consistency. Each descriptor is linked to synonyms and/or LanguaL™ codes with the same meaning to ensure unclear or different writings (e.g., "chickpea" and "chick peas") or meanings (e.g., "peeled" and "skin removed") (Table 8).

Once the connection list is established, it can be used to assign harmonized descriptors to FI by comparing synonym names to the names of database entries (Fig. 3, step 4) or LanguaL™ codes to the codes present in the FI. Finally, interlinkage of FI between databases is successful for all FI where identical descriptors are found (Fig. 4).

In the case of EuroFIR and Agribalyse, the connection list was initially created (Fig. 3, step 7 and Table S1). Harmonized descriptors were then automatically added to FI via the connection list (Fig. 4 and Fig. S2). In detail, this meant that FI from Agribalyse and EuroFIR were compared to all descriptor names and/or related synonyms from the connection list (Fig. 3). In case FI names matched one or several synonyms in the connection list, descriptors were added to the FI in the respective descriptor category. The same procedure was applied using LanguaL™ codes. Descriptors were assigned if LanguaL™ codes provided by the EuroFIR databases matched previously defined LanguaL™ codes in the connection list. Because Agribalyse does not provide any predefined LanguaL™ codes, descriptors could only be assigned through the comparison of synonym terms and not through the comparison of LanguaL™ codes.

3.4. Validation of data interlinkage

Six name-specification-treatment combinations were used to validate the data interlinkage procedure. Up to eight entries from seven different countries were found for the combinations in Agribalyse and EuroFIR (Table 9). The manual validation showed that two entries out of 54 entries were incorrectly matched and had to be excluded. The incorrectly matched entry "Tuna, in sunflower oil, canned U" from Agribalyse was catched by descriptors from "FI_3" (name = "sunflower", specification = "oil") because text patterns "sunflower" and "oil" appeared in the FI name. However, the entry primarily belonged to tuna fish and not to sunflower oil and needed to be excluded manually. When working with the data, it was observed that often the first words of a FI name were describing the food. Therefore, a possible solution to overcome such issues would be to weight words that appear at the beginning of FI names higher than words coming at the end. One entry was included but assignment has to be found only partially complete. The EuroFIR entry "EUR_3" in "FI_6" was erroneous chosen as single food entry because there was an additional cooking aid "cream". This underlines the importance of double-checking and continuously extending the filtering and final connection list.

4. Discussion

Main challenges for data interlinkage were found for the lack of available meta data and the meta data accessibility, the inconsistently given names of foods between databases and the handling of different data formats which is in agreement with other studies (Jennings-Dobbs et al., 2023; Koroušic Seljak et al., 2018; van Erp et al., 2021; Zeb et al., 2021).

Meta data serves as additional data for FI and supports a correct identification of the FI which in turn is crucial for a successful connection. As already stated by Koroušic Seljak et al. (2018), the most relevant requirement for efficient food matching lays in a high quality of raw data, especially when it comes to data documentation. EuroFIR and Agribalyse, provided sufficient amount of meta data to interlink FI with the standardization approach proposed. Although databases provided enough data for standardization purposes, limited meta data availability and a general lack of proper documentation was observed (Ferraz de Arruda et al., 2023; van Erp et al., 2021). The accuracy of database interlinkage could heavily be improved by enhancing 1) meta data availability and 2) meta data accessibility.

4.1. Meta data availability

For certain FI, further indication on composition would be relevant for a correct identification. For FI such as "Rice drink", relevant meta data (e.g., how much rice was used in the rice drink = ratio rice:water) was not provided. Database entries for "Rice drink" could therefore not further be differentiated by such ratios and were connected to other "Rice drink" entries independently. However, it is assumed that differentiating by such a ratio could affect environmental impact and nutritional composition strongly. The same holds true for composite foods (which were not considered in this work), where a further differentiation of FI by the respective recipe formulation is assumed to affect results.

Whereas the variable "foodexplorer_id" represents an identifier (ID) for EuroFIR, no ID is provided for Agribalyse. Thus, identification of FI is based solely on the names which makes data management difficult especially if FI names change when new database updates are released. This implies that already processed data has to be checked manually for its correctness and validity again, which involves repetitive and time-consuming work. We propose to introduce consistent unique universal ID's for both, LCI and FCDB data. The risk of confusion is thus minimized and the data becomes robust to changes in the name, as mentioned above.

A lack of data for the geographical representativeness of FI was also noticed. Due to this lack, differences in nutritional content due to diversity of different cultivars might not be properly reflected (Lupiañez-Barbero et al., 2018). In case of EuroFIR, each subdatabase relates to a country whereas Agribalyse is valid for the French market. However, only Agribalyse specifies the geographical region where each FI has been produced. For EuroFIR there is no such indication. That means, although considering a database entry for "strawberries" from the French EuroFIR database, it could be that the nutritional value of the strawberry entry would not reflect an average strawberry found in the French supermarket because it was imported from other countries such as Spain or Italy. Similarly, there is also no indication of the type of variety of a FI (i.

Table 9
Extracted FI from EuroFIR and Agribalyse databases for six name-specification-treatment-combinations.

Database	ID	Item	Geography	Valid?
FI_1 *(name = beef, specification = minced, treatment = cooked)*				
EuroFIR	EUR_1	Beef, ground, cooked (average)	France	True
	EUR_2	Beef, minced steak, 10% fat, cooked	France	True
	EUR_3	Beef, minced steak, 15% fat, cooked	France	True
	EUR_4	Beef, minced steak, 20% fat, cooked	France	True
	EUR_5	Beef, minced steak, 5% fat, cooked	France	True
	EUR_6	Burger, beef minced meat, cooked, without oil	Estonia	True
Agribalyse	AGB_1	Fresh ground beef production, industrial production, French production mix, at plant, 1 kg of fresh ground beef, for processing (PDi) {FR} U	France	True
	AGB_2	Ground beef, fresh, case ready, for direct consumption, at plant {FR} U	France	True
	AGB_3	Ground beef, fresh, for industrial tomato pasta products {FR} U	France	True
	AGB_4	Ground beef, fresh, for processing {FR} U	France	True
FI_2 *(name = “cashew”, treatment = “roasted”)*				
EuroFIR	EUR_1	Cashew nut, dry-grilled, unsalted	France	True
	EUR_2	Cashew nut, grilled, salted	France	True
	EUR_3	Cashew nut, grilled, unsalted	France	True
	EUR_4	Cashew nuts, dry roasted	Denmark	True
	EUR_5	Cashew nuts, oil roasted	Denmark	True
	EUR_6	Cashew nuts, kernel only, roasted and salted	United Kingdom	True
Agribalyse	AGB_1	Cashew nut, consumption mix {FR} U	France	True
	AGB_2	Cashew nut, unshelled, at processing {FR} U	France	True
FI_3 *(name = “cheese”, specification = “emmental”)*				
EuroFIR	EUR_1	CHEESE "EMMENTALER"	Slovenia	True
	EUR_2	Cheese, Emmental, 30% fat	Estonia	True
	EUR_3	Cheese, hard, Emmentaler, 45 % fidm.	Denmark	True
	EUR_4	Emmentaler cheese, at least 45% fidm	Switzerland	True
	EUR_5	Emmental cheese, from cow’s milk	France	True
	EUR_6	Emmental cheese, grated, from cow’s milk	France	True
	EUR_7	Hard cheese, emmental-type cheese, reduced fat	France	True
	EUR_8	Cheese, Emmental	United Kingdom	True
Agribalyse	AGB_1	Emmental cheese, from cow’s milk, at plant {FR} U	France	True
	AGB_2	Emmental cheese, from cow’s milk, consumption mix {FR} U	France	True
	AGB_3	Emmental cheese, grated, from cow’s milk, consumption mix {FR} U	France	True
	AGB_4	Hard cheese, emmental-type cheese, reduced fat, at plant {FR} U	France	True
	AGB_5	Hard cheese, emmental-type cheese, reduced fat, from cow’s milk, consumption mix {FR} U	France	True
FI_4 *(name = “rice”, specification = “flour”)*				
EuroFIR	EUR_1	RICE FLOUR	Slovenia	True
	EUR_2	Rice flour	France	True
	EUR_3	Rice flour	Estonia	True
	EUR_4	Rice flour	Denmark	True
	EUR_5	Rice starch	Denmark	True
	EUR_6	Flour, rice	United Kingdom	True
Agribalyse	AGB_1	Rice flour, at industrial mill {FR} U	France	True
	AGB_2	Rice flour, at industrial mill {IT} U	Italy	True
FI_5 *(name = “sunflower”, specification = “oil”)*				
EuroFIR	EUR_1	Sunflower oil	Denmark	True
	EUR_2	Sunflower oil	Estonia	True
	EUR_3	Sunflower oil	France	True
	EUR_4	Sunflower oil	Switzerland	True
	EUR_5	Sunflower oil HO (high oleic), refined	Switzerland	True
	EUR_6	Oil, sunflower	United Kingdom	True
Agribalyse	AGB_1	Sunflower oil, at plant {FR} U	France	True
	AGB_2	Sunflower oil, consumption mix, at plant {FR} U	France	True
	AGB_3	Tuna, in sunflower oil, canned {FR} U	France	False
	AGB_4	Sunflower oil, at plant {IT} U	Italy	True
FI_6 *(name = “sweet potato”, treatment = “cooked”)*				
EuroFIR	EUR_1	Sweet potato, boiled, without salt	Estonia	True
	EUR_2	Sweet potato, cooked	France	True
	EUR_3	Sweet potato, puree, cooked with cream	France	False
	EUR_4	Sweet potato, flesh only, boiled in unsalted water	United Kingdom	True
Agribalyse	AGB_1	Sweet potato, consumption mix {FR} U	France	True

e. variety of tomato) which can affect the nutritional composition (i.e. more or less content of vitamin C) of a FI (Micha et al., 2018).

Meta data regarding production systems (e.g., conventional or organic) were not always easily accessible or available for each FI in Agribalyse whereas for EuroFIR such information was not available at all.

Base data in LCI databases (here Agribalyse) could possibly be used as an additional source for the extraction of meta data because inventories might contain information of recipe formulations. For composite foods like hamburger or lasagna for example, the amount of specific food ingredients (e.g., beef meat) could be extracted even if it is not specified in the food name. However, such an extraction would be case specific and most probably not easily automatable due to the different structures of inventories provided by the different LCI databases. Thus, such an approach at the current state is expected a time-consuming process of manual work.

4.2. Meta data accessibility

As already stated by Ferraz de Arruda et al. (2023), assessing data in this study was found to be complicated as well. In the case of EuroFIR and Agribalyse, meta data was only provided as additional bulk text fragments in FI names, if available. Using such data efficiently was hindered mainly because of issues regarding the identification and extraction of such data due to inconsistent and unstructured documentation.

Currently, the title of a database entry often serves as a universal placeholder to store not only data about the type of food (e.g., apple) but also about whether food has undergone further processing (e.g., drying). Splitting the information into separately named fields would facilitate the accessibility of the data. Database owners have the best knowledge about the data and should therefore specifically focus on ways to extract and provide meta data in a structured way from current data. Additionally, they should make sure that their data collection systems offer the possibilities for data providers to insert meta data.

4.3. Classification systems

Similarly to Micha et al. (2018), a lack of international standards for documentation has also been found for EuroFIR. Improperly documented datasets limit data integration and therefore a successful data interlinkage (Jennings-Dobbs et al., 2023). Often, food is also improperly classified by databases and the different use of synonyms for food names poses problems (Lupiañez-Barbero et al., 2018). Thus, reducing food items to a common standard has been identified as key challenge. FI can only be correctly linked if they are successfully classified and identified (Koroušic Seljak et al., 2018). In regard to the implementation, each food needs to be uniquely identified. For that purpose, the name of a FI is often used. However, depending on culture and language, food might be named differently (Stadlmayr et al., 2012). It is also possible that for certain FI more than one term might be used for description (e.g., corn and maize). Depending on the context for which food data is used, names can also differ. Whereas food industry might use “dried grape”, consumers would more commonly refer to “sultana” instead. Thus, harmonization of FI names to a common standard remains a complex task especially because there is no general agreement on how FI should be named in food databases. Although FAO and INFOODS clearly states the importance of correct identification of food for proper matching of food entries, they do not propose guidance for harmonization of food names (Stadlmayr et al., 2012). Currently, there is no legal requirement on how to provide data from food databases in a standardized format (Zeb et al., 2021).

Food classification systems provide the advantage of standardized names for FI within the same system and are used in this study. Additionally, they provide a grouping scheme for foods with a systematic logic and thus classify foods with close characteristics (e.g., based on botanic classification or on nutritional content) together. Such grouping is especially useful to compare FI on a broader basis (e.g., comparing data of all FI classified as “meat”) or to approximate data of missing FI (e.g., use of FI data classified as “bovine meat” to approximate missing data of a FI named “beef steak”).

Many different classification systems are available and are used (LanguaL, 2023; Møller and Ireland, 2018). Unfortunately, classification systems are not necessarily standardized within all the classification systems available. Whereas in one system the term “banana” is used, terms such as “bananas” or “plantain” might be used in another system making it particularly difficult to combine the available systems easily.

Additionally, systems might group foods differently depending on the use case of the food classification system. Green peas are for example often considered as starchy vegetables due to their similar nutritional content with vegetables although they botanically would belong to legumes. Therefore, it is difficult that food classification systems combine and covers all aspects and characteristics. The choice of a system is mainly defined by the user applying the system to the specific use case. Whereas for Agribalyse, no food classification is provided by the database owner, EuroFIR provides a different number of LanguaL™ codes for each FI. Such codes are indirectly linked to classification systems. Although food classification systems from LanguaL™ were used, an efficient use for standardization purposes has been found difficult because codes from different classification systems were assigned. Thus, a successful connection of FI was not possible because codes for different classification systems were only partially comparable with each other (see chapter food classification). Although not clearly indicated, it is strongly assumed that such inconsistency derives from the individual assignment of LanguaL™ codes by the selected countries database managers. In some cases, LanguaL™ codes were also assigned incorrectly. To improve the quality of food classification, database managers should agree on a specific classification (e.g., EFSA FoodEx2) system to use before applying the classification to the FI. Additionally, setting up clear rules and, if data processing is not done in an automated way, training of staff on how to properly assign LanguaL™ codes to FI in a consistent way would avoid the assignment of wrong codes.

For the successful connection of FI from different databases, FI should be named according to the names given in the food classification systems in order to classify them efficiently. It is strongly recommended to use the names of existing classification systems because such systems are or might be used by other databases in the future and are updated regularly. There is a strong need to provide official standards and guidance on how to classify FI in food databases. FoodOn as an ontology providing a *“controlled vocabulary which can be used by both people and computers – to name all parts of animals, plants, and fungi […]”* could be used as a standard, especially because it is based on LanguaL™ and it could be incorporated into artificial intelligence (AI) technologies (Dooley et al., 2018; FoodOn, 2023).

4.4. Standardization approach

Coupling manual and automatized matching procedures into a semi-automatized standardization approach, assigned descriptors successfully to FI from both, EuroFIR and Agribalyse, databases. The standardization approach facilitated data interlinkage by reducing the amount of repetitive work. Having set up the connection list beforehand allowed for an automated processing of data, which has the advantage to be reapplied again once baseline data would be updated. There is also the possibility to share already gathered descriptors in order to interlink other food databases. Continuously increasing the amount of descriptors as well as improving the descriptor quality by adding more synonyms and LanguaL codes would not only pose a benefit for the interlinking status of EuroFIR and Agribalyse but also for other LCI and FCDB databases. Although descriptor assignment was correct in most of the cases, manual validation was needed to ensure that data connection was fully valid. It is expected that human expertise often remains an integral part in the whole process of data connection (Koroušic Seljak et al., 2018). Semi- or fully automated approaches therefore do not complete the task of standardization but might support in reducing the workload for manual assignment drastically. Additionally, databases are provided in different formats which complicates an efficient data accessibility.

The developed approach to interlink data of food databases allows to bring (existing) data from different sources together in order to provide a solid data basis for assessments of both, environmental and nutritional aspects of food production. The approach has been aligned with a food specific nomenclature to successfully interlink food data. Focusing on a further development of the connection list would allow for an efficient integration of other food LCI (e.g., World Food LCA Database or Agri-Footprint) and FDCB databases (e.g., USDA) (Blonk Consultants, 2014; Nemecek et al., 2019; Quantis, 2020; Haytowitz et al., 2023). Continuously increasing the amount of interlinked food databases (portfolio) enables to work with different types of food products given that one individual food database usually does not contain all types of foods. Additionally, the availability of nutritional values can also increase when interlinking foods. Whereas some FCDB might contain specific data for all different types of fatty acids, others might contain more detailed analyses of amino acids. Combining data sources therefore does not only increase the amount of data points, but also enables to work with additional data that otherwise would not have been provided.

4.5. Database format

Aiming at an efficient and successful linking of data requires a profound understanding of the organization and structure of data as well as of the technical possibilities for data connection. Basic knowledge of information technology systems often is required for a proper data management. Whereas end users of databases might have strong knowledge on the background data itself, they are often not experts or specialists in computer science. Transformation of data into a more useful format therefore is not always easily possible and might be time-intensive (Jennings-Dobbs et al., 2023). In regard to the technical implementation, combining data from different sources therefore often is a challenging task.

Data from different sources comes with different formats and often with different unique structures. In some cases, specific tools are needed to access data. Getting used to the structure of a database thus needs initial effort and time. This task becomes even more challenging if data structure is not clearly documented or very complex. Easily accessible and well-known formats such as Excel tables facilitate data management and were preferred by data managers (Clancy et al., 2015). However, often such formats are not necessarily built for handling big data and therefore reach their limits fast. Excel for example allow a maximal number of 1′048′576 rows and 16′384 columns with each cell limited to 32′676 characters. Going beyond such limitations requires other (often not commonly known) software such as Notepad++ to successfully manage data tables. However, in comparison to Excel, using other software might have limitations in the general user interface (e.g., less intuitive). Thus, it is up to the database processor to decide which software and which database format type to use in order to suit the needs for a proper processing of data the best. There is no general rule for the appropriateness of software to manage data. Providing different file formats would facilitate the use of data from databases (Jennings-Dobbs et al., 2023).

EuroFIR data was provided in an easily accessible format (Excel and XML) and was already in a state where the data could be linked without any further preprocessing. Inventory data from Agribalyse needed preprocessing where LCI data was translated into environmental impacts which required certain knowledge on how to run LCA. Data from Agribalyse could only be accessed via specific LCA tools such as SimaPro, openLCA or Brightway2 (Ciroth, 2007; Mutel, 2017; PRé Sustainability, 2012). Therefore, assessing data in Agribalyse was less convenient than in EuroFIR.

4.6. Techniques for database interlinkage

Developing and applying techniques and procedures for the standardization of data requires a comprehensive understanding of available methods and tools. As previously shown, many procedures especially in NLP have already been tested (Eftimov et al., 2017; Isiprova et al., 2017). Some of the approaches require comprehensive knowledge of programming as a prerequisite, which is not always available. Artificial intelligence (AI) might open up new possibilities for such a task. The application of such technologies has become more relevant currently. However, because AI techniques are complex, they require advanced knowhow for application. There is potential that such techniques might support the organization and structuring of available data, extract additional information on meta data, generate new data and subsequently move the development of efficient standardization procedures forward. Such potential should be elaborated in further studies.

5. Conclusions

Providing interlinked data from FCDB and LCI databases is key to ensure that analyses in the area of nLCA can correctly be conducted in future studies. Focusing on increasing the availability of more standardized data between different databases would facilitate the running of more complete analyses in order to enable better decision-making for suitable food options and to promote the overall sustainability of the food industry.

This study provides a promising approach for data interlinkage by efficiently automatizing the assignment of manually validated descriptors to database entries. In order to improve the quality of data interlinkage further, the following topics are considered relevant and should be addressed in future studies.

The presented insights into the structure of FCDB and LCI databases have revealed limited meta data availability and difficult meta data accessibility which poses significant challenges for data interlinkage. Essential meta data such as the food composition, the geographical origin of food or a universal unique identifier is not always provided by food databases and results in a major challenge when interlinking food databases. Due to the lack of meta data, FI could not be characterized and identified fully which in turn complicated data standardization. Collecting additional meta data for the description of a FI apart from the information provided in the FI name would augment data quality, allow for a clearer differentiation and a more accurate matching of the same FI. Thus, database owners should focus on providing special fields during data entry so that data providers are able to input as much meta data as possible. At the same time, the precision of algorithms should continuously be improved to ensure the maximum use of already available meta data in databases. Developing solutions to increase meta data availability of FI in food databases should become pivotal for research in order to foster transition into more reliable database connection systems.

Because each database has its own structure and is managed differently, additional knowledge on how to harmonize and facilitate data accessibility would also contribute to make data more interoperable. Research should focus on enhancing the documentation of data and providing a more common structure within the same type of databases. Meta data should also be stored in separate fields so that it can be accessed easily and does not need to be extracted beforehand. Providing structured meta data to FI which is technically easily accessible would significantly improve the correct identification, description and connection of FI between databases.

Providing and agreeing on general guidelines for the structure, accessibility and format of food databases would allow to work more efficiently within and between food databases.

The inclusion of existing food classification systems into the semi-automatized standardization approach has helped to find common names for food and group them accordingly. Because food classification systems are perceived as an essential integral part in database interlinkage and the applicability of such systems in nutritional and environmental databases as well as the implementation and further development of specific classification systems should be thoroughly investigated in future research. Currently, food classification systems are not yet fully implemented in food databases. Agreeing on common principles (e.g., use of only one classification system) would facilitate the task of connection. Pushing the development of the connection list (e.g., adding harmonized descriptors) also enables to interlink more data from other LCI and FCDB in the future.

Several solutions and techniques for database interlinkage have already been proposed and should be further elaborated. Current advancements in artificial intelligence (AI) applications could promote the development of standardization procedures and should be extensively investigated in future studies.

As stated by Ferraz de Arruda et al. (2023), working with big data is a task full of challenges. Collaboration between computer scientists and experts in the food system needs to be enhanced to integrate aspects from different models, combine already existing applications in the field and gather available knowledge. Future research should allow to bring expertise from experts and users together to provide answers to the challenges identified.

CRediT authorship contribution statement

Cédric Furrer: Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Daniel Sieh:** Writing – review & editing, Methodology, Funding acquisition. **Anne-Marie Jank:** Writing – review & editing, Methodology, Funding acquisition. **Grégoire Le**

Bras: Writing – review & editing, Software, Methodology, Data curation. **Moritz Herrmann:** Writing – review & editing, Methodology. **Alba Reguant-Closa:** Writing – review & editing, Methodology. **Thomas Nemecek:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: all authors report financial support was provided by Horizon (2020) European Innovation Council Fast Track to Innovation.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by the project OptiSignFood, which received funding from the European Union's Horizon 2020 Framework programme under grant number 971242. We thank four anonymous reviewers for the helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2024.143198.

References

Alemu, R., 2022. The race towards more sustainable food systems. Nature Food 3 (9), 679–680. https://doi.org/10.1038/s43016-022-00598-5.

Anses, 2020. Ciqual French food composition table. https://ciqual.anses.fr/.

Asselin-Balençon, A., Broekema, R., Teulon, H., Gastaldi, G., Houssier, J., Moutia, A., Rousseau, V., Wermeille, A., Colomb, V., Cornelus, M., Ceccaldi, M., Doucet, M., Vasselon, H., 2022. AGRIBALYSE 3 : la base de données française d'ICV sur l'Agriculture et l'Alimentation. Methodology for the food products. Initial publication Agribalyse 3.0-2020, update 3.1-2022 ADEME. https://doc.agribalyse.fr/documentation-en/agribalyse-data/documentation.

Becker, W., Møller, A., Ireland, J., Roe, M., Unwin, I., Heikki, P., 2008. Proposal for structure and detail of a EuroFIR Standard on food composition data II: technical Annex, Version 2008. https://www.eurofir.org/proposal-for-structure-and-detail-of-a-eurofir-standard-on-food-composition-data/.

Becker, W., Unwin, I., Ireland, J., Møller, A., 2007. Proposal for structure and detail of a EuroFIR standard on food composition data I: description of the standard. EuroFIR Technical Report. https://www.eurofir.org/proposal-for-structure-and-detail-of-a-eurofir-standard-on-food-composition-data/.

Bertoluci, G., Masset, G., Gomy, C., Mottet, J., Darmon, N., 2016. How to build a standardized country-specific environmental food database for nutritional epidemiology studies. PLoS One 11 (4), e0150617. https://doi.org/10.1371/journal.pone.0150617.

Blonk Consultants, 2014. Agri-footprint description of data, V 1.0. www.agri-footprint.com/assets/Agri-Footprint-Part2-DescriptionofdataVersion1.0.pdf.

Broekema, R., Blonk, H., Koukouna, E., van Paassen, M., 2019. Optimeal EU dataset - methodology and data development. https://blonksustainability.nl/tools-and-databases/optimeal.

Charrondiere, R., Rittenschober, D., Nowak, V., Stadlmayr, B., Wijesinha-Bettoni, R., Haytowitz, D., 2016. Improving food composition data quality: three new FAO/INFOODS guidelines on conversions, data evaluation and food matching. Food Chem. 193, 75–81. https://doi.org/10.1016/j.foodchem.2014.11.055.

Ciroth, A., 2007. ICT for environment in life cycle applications openLCA — a new open source software for life cycle assessment. Int. J. Life Cycle Assess. 12 (4), 209–210. https://doi.org/10.1065/lca2007.06.337.

Clancy, A.K., Woods, K., McMahon, A., Probst, Y., 2015. Food composition database format and structure: a user focused approach. *Plos ONE, 10*(11). https://doi.org/10.1371/journal.pone.0142137.

Delgado, A., Issaoui, M., Vieira, M.C., Saraiva de Carvalho, I., Fardet, A., 2021. Food composition databases: does it matter to human health? Nutrients 13 (8). https://doi.org/10.3390/nu13082816.

Dooley, D.M., Griffiths, E.J., Gosal, G.S., Buttigieg, P.L., Hoehndorf, R., Lange, M.C., Schriml, L.M., Brinkman, F.S.L., Hsiao, W.W.L., 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food 2 (1), 23. https://doi.org/10.1038/s41538-018-0032-6.

EFSA, 2020. FoodEx2: a standardised food classification and description system. Retrieved 2023-06-30 from. http://www.efsa.europa.eu/en/data/data-standardisation.

Eftimov, T., Korošec, P., Koroušić Seljak, B., 2017. StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. Nutrients 9 (6). https://doi.org/10.3390/nu9060542.

European Food Information Resource (EuroFIR), 2023a. eThesaurus Manager. Retrieved 2023-06-01 from https://ethesaurus.eurofir.org. .

European Food Information Resource (EuroFIR), 2023b. EuroFIR FoodEXplorer. http://www.eurofir.org/foodexplorer/login1.php.

Farran Codina, A., 2004. Desarollo y aplicacion de un sistema de informacion para la elaboracion de tablas de composicion de alimentos Barcelona, Spain. https://diposit.ub.edu/dspace/handle/2445/42489?mode=full.

Ferraz de Arruda, H., Aleta, A., Moreno, Y., 2023. Food composition databases in the era of Big Data: vegetable oils as a case study. Front. Nutr. 9 [Review]. https://www.frontiersin.org/articles/10.3389/fnut.2022.1052934.

FoodOn, 2023. FoodOn: a farm to fork ontology. Retrieved 2023-06-30 from. https://foodon.org/.

Global Standards One (GS1), 2023. GS1 general specifications standard, v23.0 (January 2023) [Standard]. https://ref.gs1.org/standards/genspecs/.

Gurinović, M., Milešević, J., Kadvan, A., Djekić-Ivanković, M., Debeljak-Martačić, J., Takić, M., Nikolić, M., Ranković, S., Finglas, P., Glibetić, M., 2016. Establishment and advances in the online Serbian food and recipe data base harmonized with EuroFIR™ standards. Food Chem. 193, 30–38. https://doi.org/10.1016/j.foodchem.2015.01.107.

Hallström, E., Davis, J., Woodhouse, A., Sonesson, U., 2018. Using dietary quality scores to assess sustainability of food products and human diets: a systematic review. Ecol. Indicat. 93, 219–230. https://doi.org/10.1016/j.ecolind.2018.04.071.

Haytowitz, D., Ahuja, J., Wu, X., Somanchi, M., Nickle, M., Nguyen, Q., Roseland, J., Williams, J., Patterson, K., Li, Y., Pehrsson, P., 2023. USDA national nutrient database for standard reference, legacy release. U.S. Department of agriculture, agricultural research service, beltsville human nutrition research center. https://agdatacommons.nal.usda.gov/articles/dataset/USDA_National_Nutrient_Database_for_Standard_Reference_Legacy_Release/24661818.

Heller, M.C., Keoleian, G.A., Willett, W.C., 2013. Toward a life cycle-based, diet-level framework for food environmental impact and nutritional quality assessment: a critical review. Environmental Science & Technology 47 (22), 12632–12647. https://doi.org/10.1021/es4025113.

Hinojosa-Nogueira, D., Pérez-Burillo, S., Navajas-Porras, B., Ortiz-Viso, B., de la Cueva, S.P., Lauria, F., Fatouros, A., Priftis, K.N., González-Vigil, V., Rufián-Henares, J.Á., 2021. Development of an unified food composition database for the European project "Stance4Health". *Nutrients, 13*(12). https://doi.org/10.3390/nu13124206.

International Network of Food Data Systems (INFOODS), 2023. International food composition table/database directory. Retrieved 2023-06-30 from https://www.fao.org/infoods/infoods/tables-and-databases/en/. .

International Standard Organisation, 2006a. ISO 14040:2006. In: Environmental Management — Life Cycle Assessment — Principles and Framework. International Standard Organisation (ISO), pp. 1–20.

International Standard Organisation, 2006b. ISO 14044:2006. In: Environmental Management — Life Cycle Assessment — Requirements and Guidelines. International Standard Organisation (ISO), pp. 1–46.

Isiprova, G., Eftimov, T., Koroušić Seljak, B., Korošec, P., 2017. Mapping Food Composition Data from Various Data Sources to a Domain-specific Ontology. 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2017), 2017-11-01. Portugal, Funchal.

Jennings-Dobbs, E.M., Forester, S.M., Drewnowski, A., 2023. Visualizing data interoperability for food systems sustainability research—from spider webs to neural networks. Curr. Dev. Nutr. 7 (11), 102006 https://doi.org/10.1016/j.cdnut.2023.102006.

Korouŝic Seljak, B., Korošec, P., Eftimov, T., Ocke, M., van der Laan, J., Roe, M., Berry, R., Crispim, S.P., Turrini, A., Krems, C., Slimani, N., Finglas, P., 2018. Identification of requirements for computer-supported matching of food consumption data with food composition data. Nutrients 10 (4), 433–451. https://doi.org/10.3390/nu10040433.

LanguaL, 2023. The international framework for food description. Retrieved 2023-06-30 from http://www.langual.org/. .

Lupiañez-Barbero, A., González Blanco, C., de Leiva Hidalgo, A., 2018. Spanish food composition tables and databases: need for a gold standard for healthcare professionals. Endocrinología, Diabetes y Nutrición (English ed.) 65 (6), 361–373. https://doi.org/10.1016/j.endien.2018.05.011.

Martinez-Victoria, E., Martinez de Victoria, I., Martinez-Burgos, M.A., 2015. Intake of energy and nutrients; harmonization of food composition databases. Nutr. Hosp. 31, 168–176. https://doi.org/10.3305/nh.2015.31.sup3.8764.

Mazac, R., Järviö, N., Tuomisto, H.L., 2023. Environmental and nutritional Life Cycle Assessment of novel foods in meals as transformative food for the future. Sci. Total Environ. 876 (162796) https://doi.org/10.1016/j.scitotenv.2023.162796.

McLaren, S., Berardy, A., Henderson, A., Holden, N., Huppertz, T., Jolliet, O., de Camillis, C., Renouf, M., Rugani, B., Saarinen, M., van der Pols, J., Vázquez-Rowe, I., Vallejo, A.A., Bianchi, M., Chaudhary, A., Chen, C., Cooreman-Algoed, M., Dong, H., Grant, T., Green, A., Hallström, E., Hoang, H.M., Leip, A., Lynch, J., McAuliffe, G., Ridoutt, B., Saget, S., Scherer, L., Tuomisto, H., Tyedmers, P., van Zanten, H., 2021. Integration of environment and nutrition in life cycle assessment of food items: opportunities and challenges. https://doi.org/10.4060/cb8054en.

Micha, R., Coates, J., Leclercq, C., Charrondiere, R., Mozaffarian, D., 2018. Global dietary surveillance: data gaps and challenges. Food Nutr. Bull. 39 (2), 175–205. https://doi.org/10.1177/0379572117752986.

Miller, F.P., Vandome, A.F., McBrewster, J., 2009. Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?

Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press. https://dl.acm.org/doi/10.5555/1822502.

Møller, A., Ireland, J., 2018. LanguaL™ 2017 – the LanguaL™ thesaurus. Danish food informatics. https://langual.org.

Mutel, C., 2017. Brightway: an open source framework for life cycle assessment. J. Open Source Softw. 2 (12), 1–2. https://doi.org/10.21105/joss.00236.

Nemecek, T., Bengoa, X., Lansche, J., Roesch, A., Faist-Emmenegger, M., Rossi, V., Humbert, S., 2019. Methodological guidelines for the life cycle inventory of agricultural products. Version 3.5, december 2019. World Food LCA Database (WFLDB). https://simapro.com/wp-content/uploads/2020/11/WFLDB_MethodologicalGuidelines_v3.5.pdf.

Poore, J., Nemecek, T., 2018. Reducing food's environmental impacts through producers and consumers. Science 360 (6392), 987–992. https://doi.org/10.1126/science.aaq0216.

PRé Sustainability, 2012. SimaPro LCA software. https://pre-sustainability.com/solutions/tools/simapro/.

Quantis, 2020. World food LCA database WFLDB. https://quantis.com/who-we-guide/our-impact/sustainability-initiatives/wfldb-food/.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks. https://arxiv.org/abs/1908.10084.

Saarinen, M., Fogelholm, M., Tahvonen, R., Kurppa, S., 2017. Taking nutrition into account within the life cycle assessment of food products. J. Clean. Prod. 149, 828–844. https://doi.org/10.1016/j.jclepro.2017.02.062.

Sirdey, N., David-Benz, H., Deshons, A., 2023. Methodological approaches to assess food systems sustainability: a literature review. Global Food Secur. 38 (100696) https://doi.org/10.1016/j.gfs.2023.100696.

Stadlmayr, B., Wijesinha-Bettoni, R., Haytowitz, D., Rittenschober, D., Cunningham, J., Sobolewski, R., Eisenwagen, S., Baines, J., Probst, Y., Fitt, E., Charrondiere, R., 2012. FAO/INFOODS guidelines for food matching, Version 1.2. https://www.fao.org/3/ap805e/ap805e.pdf.

van Erp, M., Reynolds, C., Maynard, D., Starke, A., Ibáñez Martín, R., Andres, F., Leite, M.C.A., Alvarez de Toledo, D., Schmidt Rivera, X., Trattner, C., Brewer, S., Adriano Martins, C., Kluczkovski, A., Frankowska, A., Bridle, S., Levy, R.B., Rauber, F., Tereza da Silva, J., Bosma, U., 2021. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food [perspective]. Frontiers in Artificial Intelligence 3. https://www.frontiersin.org/articles/10.3389/frai.2020.621577.

Van Mierlo, K., Baert, L., Bracquené, E., De Tavernier, J., Geeraerd, A., 2022. Moving from pork to soy-based meat substitutes: evaluating environmental impacts in relation to nutritional values. Future Foods 5 (100135). https://doi.org/10.1016/j.fufo.2022.100135.

van Paassen, M., Braconi, N., Kuling, L., Durlinger, B., Gual, P., 2019. Agri-footprint 5.0 Part 1: methodology and basic principles. www.agri-footprint.com.

van Rossum, G., Drake, F.L., 2009. Python 3 reference manual. Create space. https://dl.acm.org/doi/book/10.5555/1593511.

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., 2016. The ecoinvent database version 3 (part I): overview and methodology. Int. J. Life Cycle Assess. 21 (9), 1218–1230. https://doi.org/10.1007/s11367-016-1087-8.

Westenbrink, S., Kadvan, A., Roe, M., Koroušić Seljak, B., Mantur-Vierendeel, A., Finglas, P., 2019. 12th IFDC 2017 Special Issue – evaluation of harmonized EuroFIR documentation for macronutrient values in 26 European food composition databases. J. Food Compos. Anal. 80, 40–50. https://doi.org/10.1016/j.jfca.2019.03.006.

Wolongevicz, D.M., Brown, L.S., Millen, B.E., 2010. Nutrient database development: a historical perspective from the framingham nutrition studies. J. Am. Diet Assoc. 110 (6), 898–903. https://doi.org/10.1016/j.jada.2010.03.019.

Zeb, A., Soininen, J.-P., Sozer, N., 2021. Data harmonisation as a key to enable digitalisation of the food sector: a review. Food Bioprod. Process. 127, 360–370. https://doi.org/10.1016/j.fbp.2021.02.005.

Zhu, Z., Duan, J., Dai, Z., Feng, Y., Yang, G., 2023. Seeking sustainable solutions for human food systems. Geography and Sustainability 4 (3), 183–187. https://doi.org/10.1016/j.geosus.2023.04.001.