Promoting harmonization of life cycle inventory and food composition databases through semi-automatic standardization

Cédric Furrer[1], Daniel Sieh[2], Anne-Marie Jank[2], Grégoire le Bras[3], Moritz Herrmann[1], Alba Reguant-Closa[1], Thomas Nemecek[1]

[1] Agroscope, Life Cycle Assessment research group, CH-8046 Zurich, Switzerland
[2] themakers.ai GmbH, Chausseestrasse 8a, DE-10115 Berlin, Germany
[3] The Makers Food GmbH, Schreinerstrasse 58, DE-10247 Berlin, Germany
E-mail contact address: cedric.furrer@agroscsope.admin.ch

1. INTRODUCTION

Nutritional life cycle assessment (nLCA) is used to assess environmental impacts and nutritional quality of food combined to help identifying efficient food options (McLaren et al., 2021). However, the lack of standardization or an incomplete interlinkage of life cycle inventory (LCI) and food composition (FC) databases often limit combined analyses. Although many attempts have already been made in order to connect and standardize food items (FI) from different databases, variable database structure, different data availability, accessibility and incomplete data description have hampered a successful standardization. While fully automated procedures tend to be efficient (Isiprova et al. 2017; Eftimov et al. 2017), manual matching might be more accurate in some cases and more user friendly (Broekema et al., 2019; Hinojosa-Nogueira et al., 2021). Coupling automated and manual standardization with semi-automatic standardization has the potential to include the advantage of both methods: increasing accuracy while keeping the amount for manual work at a reasonable level.

2. METHODS

Data availability, data accessibility and the data structure of LCI and FC databases was analysed using the nutritional and environmental databases, EuroFIR and Agribalyse, respectively, as an example (European Food Information Resource (EuroFIR), 2023b; Asselin-Balençon et al., 2022). Results from the analysis were used to develop a food specific nomenclature for the semi-automatic standardization approach. Harmonized descriptors were created and collected manually beforehand. For that purpose, standardized names from the LanguaL™ system were used to properly classify FI (Møller & Ireland, 2018). Food entries in the databases were tagged with those descriptors subsequently and data interlinkage was achieved by comparing the descriptors.

3. RESULTS AND DISCUSSION

FI of both databases were found to be structured into glossaries with two main parts: the meta data storing descriptive information about the food (e.g., food name) and the base structure (e.g., nutritional parameters). “Food name”, “Food specification”, “Food recipe” and “Food processing” have been identified as key parameters for the standardization of food databases because they are required to uniquely identify the type of food . FI from LCI databases are also sensitive to parameters such as “System boundaries”, “Yield”, “Country of origin of food” and “Production system”. Information on parameters could only be accessed, if available, via the name field of a FI. Data connection was facilitated when excluding composite foods (e.g., pizza) from the standardization and only focusing on single foods (e.g., apple). Five categories (name, specification, treatment, processing, production system) were identified suitable for the food-specific nomenclature. Gathering synonyms and/or LanguaL™ codes manually in a connection list beforehand allowed for a standardized and automatic assignment and description of FI afterwards (Figure 1). Using the semi-automatic procedure in a case study showed that two entries out of 54 were incorrectly matched and had to be excluded manually.

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

*: little relevant or irrelevant; **: moderately relevant; ***: highly relevant
I: not provided; II: sometimes provided; III: fully provided

4. CONCLUSIONS

Applying semi-automatic standardization via the connection list showed to be a user friendly and accurate approach for standardization. Augmenting data quality by collecting additional meta data for the description of a FI would allow for a more correct matching of the same FI. Providing and agreeing on general guidelines for the structure, accessibility and format of food databases would increase the efficiency of data standardization between food databases.

5. ACKNOWLEDGEMENTS

This work is supported by the project OptiSignFood, which received funding from the European Union’s Horizon 2020 Framework programme under grant number 971242.

6. REFERENCES

McLaren, S. et al. (2021). Integration of environment and nutrition in life cycle assessment of food items: opportunities and challenges. https://doi.org/10.4060/cb8054en

Isiprova, G., et al. (2017). Mapping Food Composition Data from Various Data Sources to a Domain-Specific Ontology. 9th International Joint Conference on KEOD, Funchal, Portugal.

Eftimov, T. et al. (2017). StandFood: Standardization of Foods Using a Semi-Automatic System for Classifying and Describing

Foods According to FoodEx2. Nutrients, 9(6).

Hinojosa-Nogueira, D. et al. (2021). Development of an Unified Food Composition Database for the European Project "Stance4Health". Nutrients, 13(12). https://doi.org/10.3390/nu13124206

European Food Information Resource (EuroFIR). (2023b). EuroFIR. Retrieved 2023-06-30 from https://www.eurofir.org/about_eurofir/

Asselin-Balençon, A. et al. H. (2022). AGRIBALYSE 3 : la base de données française d'ICV sur l'Agriculture et l'Alimentation. ADEME. www.ademe.fr/mediatheque

Møller, A., & Ireland, J. (2018). LanguaL™ 2017 – The LanguaL™ Thesaurus. Danish Food Informatics. https://langual.org

Table 1. Relevant parameters for standardization 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.	*** (III)	*** (III)	Information needs to be extracted from title of a database entry
Food specification	*"Juice"*, *"Oil"*, etc.	*** (II)	*** (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	*** (I)	*** (I)	Information, if provided, only in base data. Difficult to extract.
Food processing	*"pasteurized"*	*** (III)	*** (II)	Information needs to be extracted from title of a database entry
System boundaries	*"at farm"* or *"at processing"*	* (I)	*** (II)	Not always provided in the database entry in Agribalyse
Yield	Yield of apple from agricultural production	* (I)	*** (II)	Information only provided in base data. Difficult to extract.
Country of origin of food	*"Germany"*, *"France"*, etc.	** (I)	*** (III)	
Production system	*"conventional"*, *"organic"*, etc.	* (I)	*** (II)	Information needs to be extracted from title of a database entry

*: little relevant or irrelevant; **: moderately relevant; ***: highly relevant
I: not provided; II: sometimes provided; III: fully provided

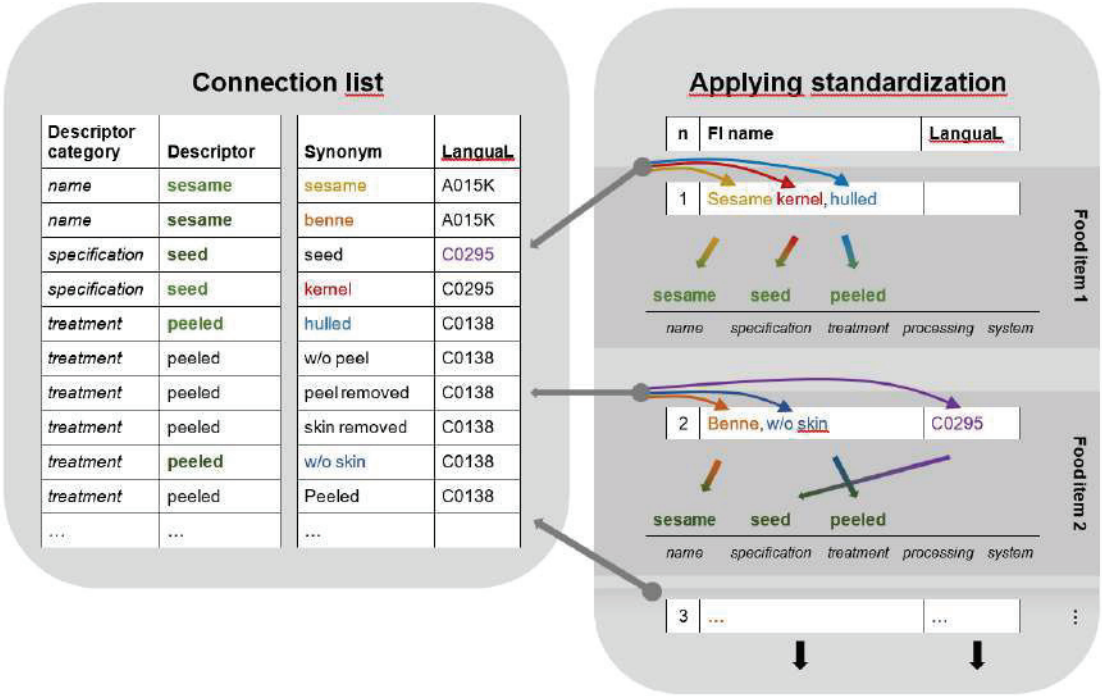

Figure 1. Scheme of the application of the standardization approach for two FI ("Sesame kernel, hulled" and "Benne, w/o skin"). For each FI where the standardization is applied, the name and the LanguaL™ codes (right), if available, are compared to the information of the connection list (left). If one or more terms or LanguaL™ codes appear in the connection list, the associated descriptor is assigned.