Food | Consumer | Health
Designing a world-class infrastructure to facilitate research

Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package 1
Deliverable D1.2

Position and final paper of RICHFIELDS

Date delivered:
M36

Authors:
Marc Jeroen Bogaardt, Giacomo Copani, Javier de la Cueva, Paul Finglas, Anouk Geelen, Charo Hodgkins, Barbara Korousic, Bent Mikkelsen, Krijn Poppe, Golboo Pour Abdollahian, Jos van den Puttelaar, Monique Raats, Trond Selnes, Lada Timotijevic, Hennie van der Veen, Pieter van 't Veer, Karin Zimmermann

Deliverable lead beneficiaries:
Stichting Wageningen Research

www.richfields.eu
#RICHFIELDS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Project	
Project acronym:	RICHFIELDS
Project full title:	Research Infrastructure on Consumer Health and Food Intake for E-science with Linked Data Sharing
Grant agreement no.:	654280
Project start date:	01.10.2015
Document:	
Title:	Position and final paper of RICHFIELDS
Deliverable No.:	D1.2
Authors:	Marc Jeroen Bogaardt, Giacomo Copani, Javier de la Cueva, Paul Finglas, Anouk Geelen, Charo Hodgkins, Barbara Korousic, Bent Mikkelsen, Krijn Poppe, Golboo Pour Abdollahian, Jos van den Puttelaar, Monique Raats, Trond Selnes, Lada Timotijevic, Hennie van der Veen, Pieter van 't Veer, Karin Zimmermann
Reviewer:	Karin Zimmermann – Project Coordinator prof. dr. ir. Pieter van't Veer – Scientific Coordinator
Start date:	1.10.2015
Delivery date:	30.09.2018
Due date of deliverable:	30.09.2018
Dissemination level:	PU
Status:	Final

Change history:		
Version	Notes	Date

Karin Zimmermann
Project Coordinator

prof. dr. ir. Pieter van't Veer
Scientific Coordinator

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package 1
Deliverable 1.2

Position paper of RICHFIELDS

Authors:
Marc Jeroen Bogaardt, Anouk Geelen, Karin Zimmermann, Paul Finglas, Monique Raats, Bent Mikkelsen, Krijn Poppe and Pieter van 't Veer

Bogaardt et al. Designing a research infrastructure on dietary intake and its determinants. Nutrition Bulletin, 43, 301–309.
https://onlinelibrary.wiley.com/doi/pdf/10.1111/nbu.12342

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Nutrition Bulletin

EMERGING RESEARCH

DOI: 10.1111/nbu.12342

Designing a research infrastructure on dietary intake and its determinants

M.-J. Bogaardt*, A. Geelen†, K. Zimmermann*, P. Finglas‡, M. M. Raats§, B. E. Mikkelsen¶, K. J. Poppe* and P. van't Veer†

*Wageningen Economic Research, The Hague, The Netherlands;
†Wageningen University, Wageningen, The Netherlands;
‡Quadram Institute Bioscience, Norwich, UK;
§University of Surrey, Guilford, UK;
¶Aalborg University, Aalborg, Denmark

Abstract

Research on dietary intake and its determinants is crucial for an adequate response to the current epidemic of diet-related non-communicable chronic diseases. In order to respond to this challenge, the *RICHFIELDS* project was tasked with designing a research infrastructure (RI) that connects data on dietary intake of consumers in Europe, and its determinants, collected using apps and wearable sensors, from behavioural laboratories and experimental facilities and from other RIs. The main output of the project, an RI design, describes interfaces (portals) to collect data, a meta-database and a data-model to enable data linkage and sharing. The *RICHFIELDS* project comprises three phases, each consisting of three work packages, and an overarching methodological support work package. Phase 1 focused on data generated by consumers (*e.g.* collected by apps and sensors) relating to the purchase, preparation and consumption of food. Phase 2 focused on data generated by organisations such as businesses (*e.g.* retail data), government (*e.g.* procurement data) and experimental research facilities (*e.g.* virtual supermarkets). Phases 1 and 2 provided Phase 3 with insights on data types and design requirements, including the business models, data integration and management systems and governance and ethics. The final design will be used in the coming years to build an RI for the scientific research community, policy makers and businesses in Europe. The RI will boost interdisciplinary multi-stakeholder research through harmonisation and integration of data on food behaviour.

Keywords: big data, consumers, diet, food, public health, research infrastructure

Identifying the need for research infrastructures

Diet-related, non-communicable chronic diseases, such as obesity and cardiovascular diseases, have been identified as a key European societal challenge as they pose a significant threat to the health of the population of the European Union (EU) (WHO 2012). To respond to this challenge, recent EU initiatives have been funding relevant research (JPI HDHL 2012; European Commission 2017). Dietary habits are determined by physical, biological, psychological, economic and sociocultural factors (Sobal 1991), which all operate simultaneously and interactively (Sobal *et al.* 2014).

Correspondence: Marc-Jeroen Bogaardt, Senior Researcher, Wageningen Economic Research, Alexanderveld 5, 2585 DB The Hague, The Netherlands.
E-mail: marc-jeroen.bogaardt@wur.nl

A robust and dynamic scientific evidence-base on dietary determinants is needed for the research community, governments, civil society organisations and the private sector to effectively respond to the urgent diet-related public health and sustainability challenges.

The EU's Seventh Framework Programme (FP7) project *EuroDISH* previously mapped existing research infrastructures (RIs) in the health and food domain (Brown *et al.* 2017; Snoek *et al.* 2018). The DISH-model was used to distinguish information about determinants of dietary behaviour (D), intake of food and nutrients (I), its relation to status and functional markers of the body (S), and health and disease outcomes (H) (Brown *et al.* 2017). The *EuroDISH* project confirmed a current state of disparate and fragmented health and food RIs (Brown *et al.* 2017). It found that fewer RIs exist in the area of food choice determinants compared to the food intake, status and health areas, and that RIs linking food choice determinants with food intake are also lacking (Snoek *et al.* 2018). The resulting knowledge gaps are hindering evidence-based research, the design of effective public health nutrition strategies and the reformulation food products by the food industry (Brown *et al.* 2017).

The open data movement in research and innovative ways of collecting data, including user-generated (big) data, provide new opportunities to study diet, lifestyle and their determinants. Data can be collected real-time [*e.g.* with geographic information system sensors] at the individual and group level, and this could provide valuable information on associations between determinants of food choice and dietary intake. Data to study food consumption patterns can be collected through new media platforms such as Twitter (Abbar *et al.* 2014; Fried *et al.* 2014) and Instagram (Mejova *et al.* 2015; Sharma & De Choudhury 2015). Weber and Achananuparp (2016) used data from public food diaries collected using the app *MyFitnessPal* to construct models to predict whether users will or will not meet their daily caloric goals.

The 3-year *RICHFIELDS* RI design project commenced in October 2015 with funding from Horizon 2020's EU Research Infrastructures (including e-Infrastructures) Work Programme. The project was tasked with producing a design for a RI for data on food-related consumer behaviour. This will serve as a data platform to facilitate the efficient alignment, linkage and sharing of scientifically reliable and technically sound data in the domains of food choice determinants and intake, while simultaneously accounting for ethical, legal and social considerations key to being able to conduct breakthrough research, develop innovative solutions to societal challenges, and enable policy makers and food industries to develop, evaluate and implement effective food and health policies, products and services.

EuroDISH's conceptual design as starting point

The conceptual design of the RI (Fig. 1) builds on the *EuroDISH* project (Snoek *et al.* 2018) and illustrates how different data sources of legally autonomous organisations can interact to enable the European research community to collaborate more effectively.

The conceptual design encompasses interfaces (portals) to collect data, a meta-database that provides information on the availability and accessibility of the data, and a data model that safeguards data comparability through methodology standardisation and calibration to enable data linkage and sharing.

The *RICHFIELDS* project explored the possibilities of using and combining different types of data: consumer-generated data, mostly real-time and *in situ*; business-generated data; and research-generated data from research laboratories, experimental facilities and from existing and developing RIs. Users of the data platform will be the scientific research community and also consumers, civil society, policy makers and the private sector. The services offered by the RI will include data sharing, standardisation, linking and quality assessment. Services for consumers could include diet advice, special offers and shopping list advice.

Structure of the RICHFIELDS project

RICHFIELDS comprises three phases (or design elements), each consisting of three work packages. The parallel Phases 1 and 2 each focused on different data types and together form the basis of the RI design developed in Phase 3 (Fig. 2). The specific aims of the three Phases were to:

- collect data generated by consumers when engaged in activities related to the purchase, preparation and consumption of food (Phase 1);
- identify data generated by business and research from laboratories and experimental facilities and other related RIs on purchase, preparation and consumption of food (Phase 2);
- design the RI including the business model, data integration and management, and governance and ethics (Phase 3).

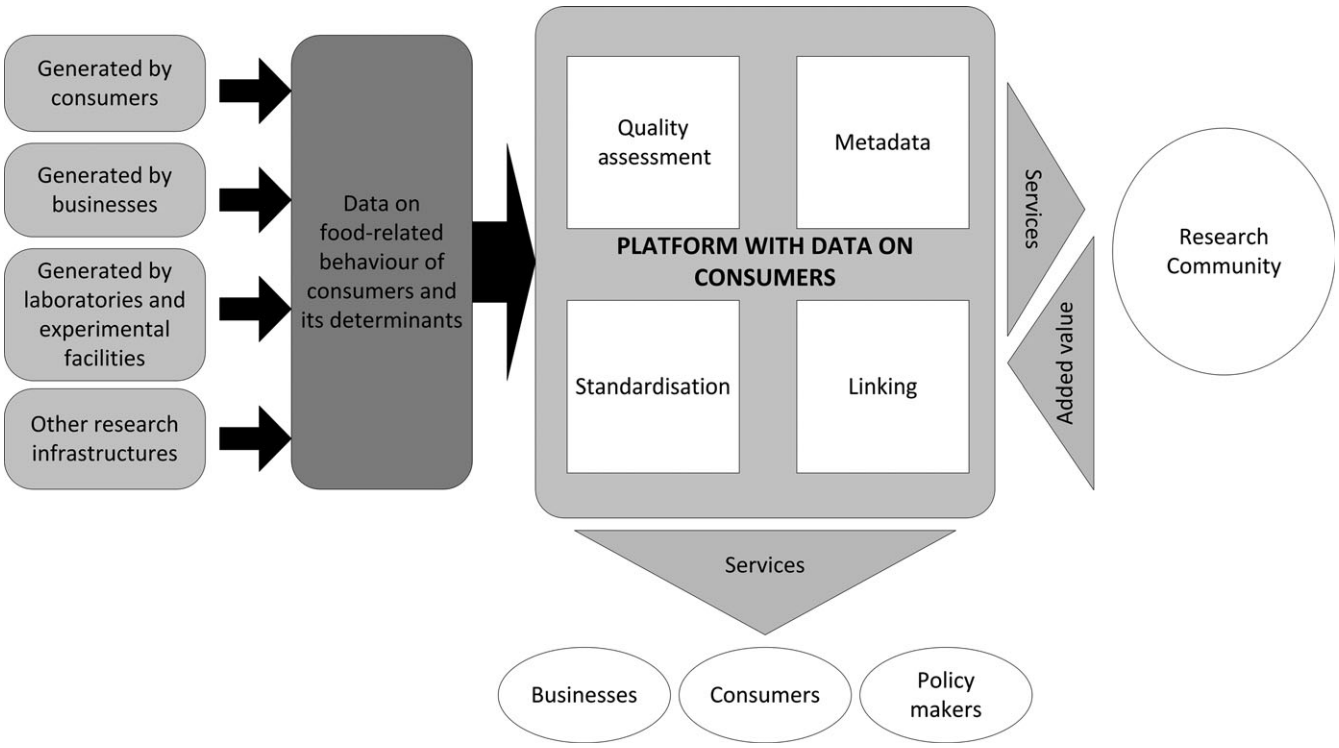

Figure 1 Conceptual design of the research infrastructure on dietary intake of consumers and its determinants.

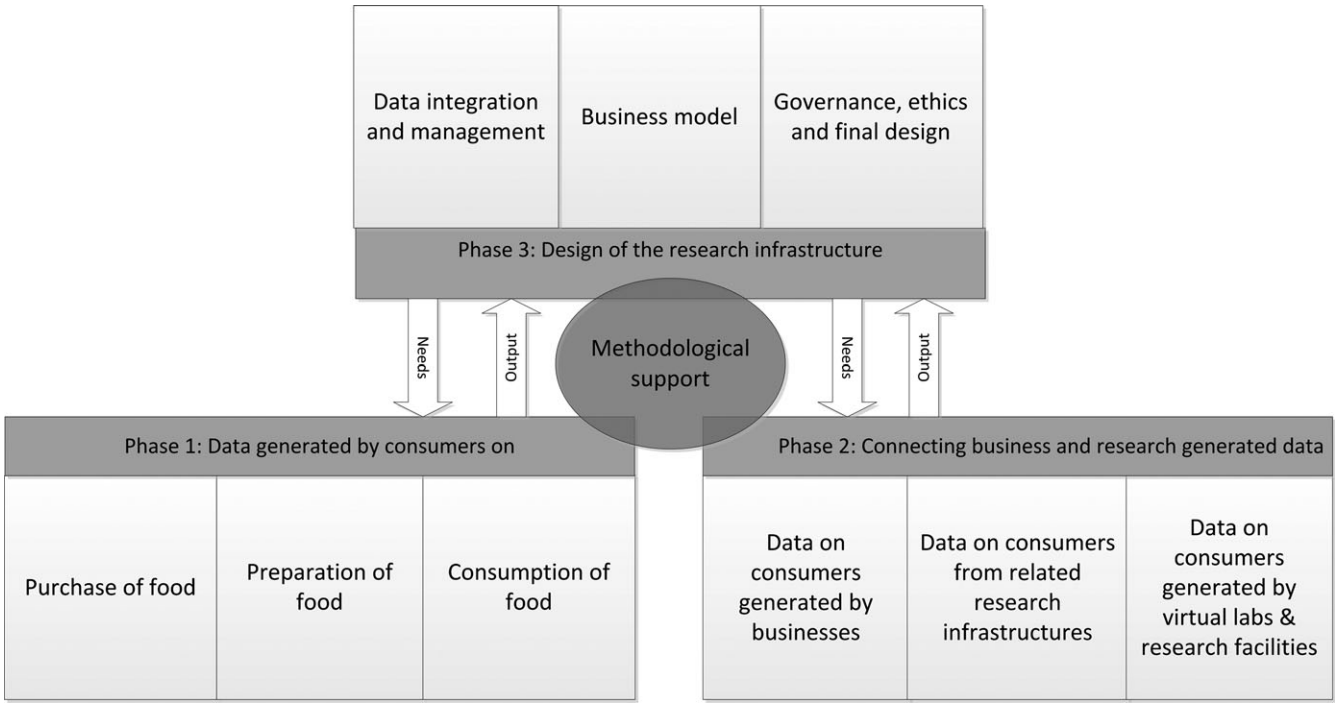

Figure 2 Structure of the RICHFIELDS project.

To ensure methodological consistency across Phases 1 and 2, a specific work package provided methodological support (see Fig. 2) including defining and harmonising concepts and methods to facilitate integration.

Phase 1: Data generated by consumers

Due to the heterogeneity of the food supply and consumers lifestyles across European sub-regions, gathering data on dietary habits and health-related consumer behaviours is scientifically challenging (Stefler & Bobak 2015). Questionnaires, focus groups, observational methods and interviews are widely used research tools for collecting food-related consumer behaviour data. New technology-driven research tools are slowly on the rise using, for example, the TwitteR software package (Vidal *et al.* 2015), tracking technologies (in tourism studies) (Shoval & Ahas 2016), and brain imaging (in sensory sciences) (Horska *et al.* 2016; Reichert *et al.* 2018).

The *RICHFIELDS* RI design project considered three important food-related behaviours: purchase, preparation and consumption. Key research questions include: what food do people eat, in what quantity and what frequency? What food-related behaviours are associated with which dietary patterns? What are the demographic and personal characteristics of people with different diets? What are their attitudes, normative beliefs and social motivations, reasoning, emotions, towards health and sustainability? What is the social and built environment in which the behaviour is carried out?

As well as providing insights regarding food-related behaviour *per se*, the consumer-generated data can be used to derive health-related dietary data; for example, energy and nutrient intakes, dietary quality (nutrient density, energy density), which in turn may be related to energy balance (sedentary behaviour, physical activity, body size and composition), health status (blood lipids, blood pressure, overweight, chronic diseases) and lifestyle (sleep, stress) factors. Consumer data on purchase, preparation and consumption of food can be generated real time and *in situ*, using innovative information and communication technology (ICT) technologies (*e.g.* apps). Tools for consumer-generated data, including wearable technology, are expected increasingly to become an integral part of society (Research 2 Guidance 2015).

Phase 1 identified food-related data that is being actively or passively generated by consumers through the use of tools such as apps and sensors. Examples include banking transactions from which food-related purchase can be estimated, food-related search internet behaviour (*e.g.* recipes, restaurant reviews) and the use of apps to record food intake or disclose food-related images or text. The large scale generation of such data has the potential to provide data for the purpose of research. In order to determine consumers' willingness to share their food-related data, quantitative research was conducted in eight European countries (France, Germany, Italy, The Netherlands, Slovenia, Spain, Sweden and the UK) to provide insights as to the type of food-related data being generated, and the extent to which people are willing to share data with scientists, government and business that produce or sell foods and drinks. The survey also collected data on determinants of willingness to share data.

RICHFIELDS developed a set of quality criteria for the evaluation of consumer-generated data in terms of its scientific relevance and technical and legal governance. This includes the legal limitations, organisational restrictions, confidentiality and privacy concerns related to the collection, integration and dissemination of consumer-generated data and the technical protocols and standards for data access and data processing. Information about these topics is crucial for developing the blueprint of a data platform, such as *RICHFIELDS*, as well as for its data governance structure.

Phase 2: Data generated by business and research

Phase 2 identified and investigated how the data platform could be connected with data generated by businesses and the research community (see Fig. 2).

Business-generated data

The use of business-generated data was examined through interviews with representatives from businesses and agencies that are already collecting data on different aspects of food consumption. Two types of business-generated data were investigated in case studies, namely data generated in business-to-business interactions, where consumers purchase foods in retail stores, and data generated in business-to-government interactions, in which the food is sold by wholesalers to governments for use in welfare catering. The first is referred to as purchase and the second as procurement. The cases studies focussed on how ICT (*e.g.* software applications for data import and export, smartcards, near field communication tools, data meshes) is being and could be used to

make data collection more convenient. The feasibility and the ethical issues of the data capture were also examined.

Research-generated data

This work package focused on how data generated through research in smart lab settings can be included in a future data platform, with a particular emphasis on how new technologies and devices are being used in physical laboratories and research facilities across Europe to study behavioural nutrition under experimental conditions. Such new facilities have spread in the wake of the increased interest in studying dynamics of food choice and the behavioural design of food environments. The challenges and opportunities associated with extracting laboratory data were explored in three case study food laboratories and facilities: Restaurant of the Future located at Wageningen University (www.wur.nl/en/Expertise-Services/Facilities/Restaurant-of-the-Future-4.htm), Foodscape Lab located at Aalborg University, Copenhagen (www.capfoods.aau.dk/technical/FoodScapeLab/), and the Fake Food Buffet at ETH Zurich (Bucher *et al.* 2012). The experimental research settings in these laboratories and facilities add important scientific value by enabling data exchange and cross-validation between the research settings. They can also be used to test hypotheses about how consumers behave in real-life consumer environments (*e.g.* supermarkets, restaurants, home kitchens) and investigate bio-psychological mechanisms of food choice. All three facilities are controlled laboratory settings that allow for data collection under well-defined conditions, two of which, the Restaurant of the Future and the Foodscape Lab, also provide options for collecting data in real-life eating environments (campus canteens). In the Foodscape Lab and the Restaurant of the Future, experiments can be carried out using real food as well as virtual food environments such as virtual supermarkets and virtual buffets, where virtual reality (VR) technology is used and behavioural data is collected digitally in real-time (Mikkelsen *et al.* 2016a,b). At the Fake Food Buffet in ETH Zurich, food replicas, as well as real food, are used in experiments (Mikkelsen *et al.* 2016a,b). As developing and maintaining such kinds of lab facilities is rather costly and technology intensive, it was important to determine how protocols, devices, skills, and data can best be exchanged across the facilities and how a RI might play a role in this. An inventory of other experimental research facilities generating data concerning purchase, preparation and consumption of food has also been compiled.

The potential for delivering data to the platform from other relevant existing and developing RIs related to consumers' food intake, health, and lifestyle in Europe was explored in a separate work package, focusing first on food composition and food attributes data (Finglas *et al.* 2014). This included approaches to and challenges of integrating data on non-nutrient bioactives and food allergens, as well as the possibility of including data on branded foods. This was followed by a focus on linking to data related to food intake created within the framework of standardised dietary monitoring systems using agreed standard methodologies, such as GloboDiet (Dietary Exposure Assessment Group 2016). The work also included studying links to data collected in the context of clinical interventions [*e.g.* by European Clinical Research Infrastructure Network (ECRIN; www.ecrin.org) and European Commission (EC)-funded projects such as the *European Nutritional Phenotype Assessment and Data Sharing Initiative* (*ENPADASI;* www.enpadasi.eu) and *QuaLiFY* (www.qualify-fp7.eu)], and data on lifestyle factors, such as exercise, stress and sleep behaviour [*e.g.* EC-funded project *PREventive Care Infrastructure based On Ubiquitous Sensing* (*PRECIOUS;* www.thepreciousproject.eu)].

Phase 3: Design of the research infrastructure

Phase 3 designed the business model, the data integration and data management, and the governance of the RI (see Fig. 2). In general, designing data platforms with many data suppliers and data users involves highly complex sets of network externalities between and within different user groups (Reuver *et al.* 2015). In designing data platforms, different methods are applied; for example, the Design Science Research Methodology (DSRM) for open data platforms (Alexopoulus *et al.* 2014), Architecture Analysis and Design Language (AADL) for big data driven physical systems (Zhang *et al.* 2015) and sometimes a completely new architectural design is developed (Simmon *et al.* 2015). Taking a stakeholder and technology-oriented perspective, that accounts for both data providers and end-users, as well as technical restrictions associated with different data sources, is key to successful platform design (Schreieck *et al.* 2016). It is important to engage stakeholders throughout the design process (Michener *et al.* 2012) as their willingness to share data will determine the success of the data platform.

Business model

To develop a sustainable business model for the RI that provides value to its users, the benefits of various business models have been explored in terms of their value proposition (*i.e.* the services provided to the different users of the RI, such as scientists, consumers and businesses), the supply chain configuration (*i.e.* the way services are produced and provided to users), and the revenue model (*i.e.* the financial mechanisms that determine and regulate economic flows among all stakeholders). These business models were then subjected to a socioeconomic performance assessment, estimating the order of magnitude of investment needed, long-term turnover, and turnover impact for different participating organisations (private companies, research institutes). In order to assess their feasibility, these alternative business models were presented to relevant groups of stakeholders. Based on their feedback the final business model was further developed and subjected to performance forecast analysis focusing on indicators, such as net present value and payback time, to estimate its sustainability.

Data integration and management

The proposed data integration and data management procedures are based on state-of-the-art ICT for collecting big data from consumers, such as sensors, digital pictures, videos, purchase transaction records and GPS signals. Interfaces (portals) for different groups of users of the data platform, tailored to their specific needs, are key elements of the RI design. These user requirements were considered through evaluating similar multi-sided data platforms, based on innovative cloud and big data technology, such as Future Internet space (FIspace; www.fispace.eu), Just Eat (www.just-eat.com), Big Data Public Private Forum (www.big-project.eu), and evidence-based European RIs and projects such as the *European Food Information Resource* (*EuroFIR*; www.eurofir.org 2016), *NuGO* (an Association of Universities and Research Institutes focusing on the joint development of the research areas of molecular nutrition, personalised nutrition, nutrigenomics and nutritional systems biology) (www.nugo.org.), and *ECRIN* (www.ecrin.org), which mostly use relational databases to store data. In order to link different types of data, the *RICHFIELDS* project's new semantic data model is based on existing standard ontologies and incorporates aspects from the domains investigated in Phase 2. Together with a data provenance concept (*i.e.* who provided the data, in what context and how the data were dealt with), the architecture of the RI has been designed to enable full data integration. Functional and technical standardisation will ensure that apps can communicate with the proposed data platform.

Governance and ethics

The governance of the proposed RI encompasses the use of institutional and authority structures and forms of collaboration to allocate resources (*e.g.* money, people) to coordinate activities and control joint action across the network of participating organisations (Provan & Kenis 2007). The success of the RI based on the *RICHFIELDS* project's design will depend on the appropriate governance of all involved organisations with their datasets and resources (apps, sensors), their facilities (research laboratories, experimental facilities) and related services (cloud, interfaces). The RI governance structure deals with privacy, data protection, RI ownership of data, ownership of the RI, intellectual property rights, transparency and trust. In particular, consumer concerns about the (mis)use of their personal data, which includes their food-related behaviour, needs to be considered (European Commission 2015). The design of the proposed RI's governance structure also considers developments in digital technology and scientific research (e-science) software and the European Commission's ambition to make all scientific data open. Stakeholder views on the different governance models were sought and used to shape the final proposed RI governance structures and their alignment with the business model and the data model. The final design of the RI will be accompanied by a roadmap (including the financial strategy) for the actual building of the RI.

Challenges

Food-related data generated by consumers is of inherent interest to researchers and currently remains largely inaccessible and disconnected from the scientific community. Consequently, an important legal issue is whether consent has been obtained for use, for example, in research, and if not, how it might be obtained. It must be clear for which purposes the research community and businesses will use the data generated by consumers (Umhoefer *et al.* 2015).

A second challenge is whether each participating data provider will be willing to share its data with (some of) the other parties as data users. The ambition

to develop the European public-private RI is challenging due to the different 'cultures' of academia, food enterprises and European consumers.

The quality of the data is another concern. A key question is whether the data are fit for purpose. Another challenge is designing a sustainable, operational RI. This means that the RI must be flexible enough to adapt to new emerging technical data collection tools, such as implantable devices (*e.g.* microchips inserted into the human body), and to different forms of self-monitoring occurring in society: private, pushed, communal, imposed and exploited self-tracking (Lupton 2014). The RI also needs a sustainable business model that can withstand financial setbacks in the future. The ambition is to design a RI that copes with these challenges and overcomes the current data fragmentation between individual level and its environment in research, business and policy, and provides adequate services to tackle the societal challenge of diet-related non-communicable chronic diseases.

Stakeholder participation, consortium management, and dissemination

Generation and use of (big) data on food-related behaviours of consumers depends on the willingness to share data from a broad range of stakeholders: consumers, researchers, app providers and developers, food retailers, food and beverage industries, restaurants and caterers. Therefore, the engagement of these stakeholders in the design is crucial. Key stakeholders have provided input to *RICHFIELDS* through stakeholder platforms and workshops.

The *RICHFIELDS* consortium comprises 16 organisations from 11 EU Member States and one organisation from a non-EU Member State. The coherency and scientific quality of the work packages within each phase was coordinated and overseen by a scientific coordinator. To align the work and ensure coherency between the phases, the scientific phase coordinators regularly met as part of the Scientific Coordination Team. Together with the overall project coordinator and the project manager they formed the Project Management Team. The project also benefited from the expertise and networks of the members of the external Project Advisory Board, all active in the private and public sectors or scientific community, who provided input on stakeholders' needs and feedback about the progress and (preliminary) results.

Scientific papers (*e.g.* on measuring food choice and consumption behaviour, on linked data sharing) will document and disseminate the project's scientific results. In addition, the results have been, and will be, discussed with peers and stakeholders at conferences during and after the project's lifetime. A website (www.richfields.eu) provides project partners, stakeholders and other interested parties with information about progress and outputs. As well as the use of Twitter and LinkedIn, an annual electronic newsletter provides partners and stakeholders with project updates. Finally, in September 2018, the final RI design and roadmap will be presented to researchers, businesses and policy makers at a conference in Brussels.

Acknowledgements

The Consortium partners are: Wageningen Economic Research (NL), the German Institute of Food Technologies (DE), European Food Information Resource AISBL (BE), Jožef Stefan Institute (SI), Wageningen University (NL), University of Surrey (UK), RISE Research Institute of Sweden, Aalborg University (DK), Javier de la Cueva & Asociados (ES), Swiss Federal Institute of Technology (CH), European Food Information Council (BE), Quadram Institute Bioscience (UK), Institute of Industrial Technologies and Automation-National Research Council of Italy (IT), Centre for European Nutrition and Health (FR), Aalto University (FI), GS1 Denmark (DK). The Scientific Coordination Team are: Prof. Pieter van't Veer (chair; Wageningen University), Professor Monique Raats (coordinator Phase 1; University of Surrey), Professor Bent Egberg Mikkelsen (coordinator Phase 2; Aalborg University), Mr. Krijn Poppe (coordinator Phase 3; Wageningen Economic Research) and Dr. Paul Finglas (accessory member; Quadram Institute Bioscience). The members of the Project Advisory Board, appointed in June 2016, are: Dr. Inge Tetens (Chair; University of Copenhagen), Dr. Igor Spiroski (IPH Macedonia), Mrs. Anneke van Kollenburg (ENECO), Dr. Harriet Teare (University of Oxford), Mr. Fred van Alphen (IT manager), Mr. Christian Graversen (Welfare Tech) and Dr. Marijntje Bakker (JPI HDHL).

Conflict of interest

The authors have no conflict of interest to declare.

Funding

The *RICHFIELDS* project has received funding from the European Union's *Horizon 2020* Research and Innovation Programme under grant agreement No 654280 (www.richfields.eu).

Authors' contributions

MJB was responsible for the integration of the contributions of the co-authors. KZ and PvtV contributed to the background and the conceptual design section. AG and MR formulated the section about Phase 1 concerning data generated by consumers; BM and PF contributed the section about Phase 2 focusing on data generated by business and research; and KP and MJB were responsible for the content of the section about Phase 3 concerning the design of the RI. KZ contributed to the section about consortium management, stakeholder participation and dissemination. All authors were responsible for the final discussion section. All authors read and approved the final manuscript.

References

Abbar S, Mejova Y & Weber I (2014) You tweet what you eat: studying food consumption through twitter. *Computing Research Repository. DOI* **10**: 2702153.

Alexopoulus C, Zuiderwijk A, Charapabidis Y *et al.* (2014) Designing a Second Generation of Open Data Platforms: Integrating Open Data and Social Media. *International Federation of Information Processing 2014*, M. Janssen *et al.* (Eds.): EGOV 2014. LNCS 8653, pp. 230–41.

Brown KA, Timotijevic L, Geurts M *et al.* (2017) Concepts and procedures for mapping food and health research infrastructure: new insights from EuroDISH project. *Trends in Food Science & Technology* **63**: 113–31.

Bucher T, van der Horst K & Siegrist M (2012) The fake food buffer. A new method in nutrition behaviour research. *British Journal of Nutrition* **107**: 1553–60.

Dietary Exposure Assessment Group (2016) *Biennial Report 2014/2015*. Available at: www.iarc.fr/en/publications/pdfs-online/breport/breport1415/10SectionOfNutritionAndMetabolism.pdf (accessed 1 December 2016).

European Commission (2015) *Special Eurobarometer 431. Data protection. Report*. https://doi.org/10.2838/552336.

European Commission (2017) Food & Healthy Diet. Available at: https://ec.europa.eu/programmes/horizon2020/en/area/food-healthy-diet (accessed 10 May 2017).

Finglas PM, Berry R & Astley S (2014) Assessing and improving the quality of food composition databases for nutrition and health applications in Europe: the contribution of EuroFIR. *American Society for Nutrition. Advances in Nutrition* **5**: 608S–14S.

Fried D, Surdeanu M, Kobourov S *et al.* (2014) *Analyzing the language of food on social media*, *IEEE International Conference on Big Data*, October 27–30, 2014. Washington, DC, USA. https://doi.org/10.1109/BigData.2014.7004305.

Horska E, Bercik J, Krasnodebski A *et al.* (2016) Innovative approaches to examining consumer preferences when choosing wines. *Agricultural Economics*. **62**: 124–33.

JPI HDHL (Joint Programming Initiative - A Healthy Diet for a Healthy Life) (2012) *Strategic Research Agenda 2012–2020 and beyond*. 1st edition. June 2012. Joint Programming Initiative A healthy diet for a healthy life. The Hague, The Netherlands.

Lupton D (2014) *Self-tracking modes: reflexive self-monitoring and data practices*. http//dx.doi.org/10.2139/ssrn.2483549.

Mejova Y, Haddadi H, Noulas A *et al.* (2015) #*foodporn: Obesity patterns in culinary interactions*, *Proceeding DH 2015 Proceedings of the 5th International Conference on Digital Health*, pp 51–8.

Michener W, Allard S, Budden A *et al.* (2012) Participatory design of DataOne. Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics* **11**: 5–15.

Mikkelsen BE, Bucher T, Hieke S *et al.* (2016a) *Measuring food choice and consumption behaviour with real, fake or virtual food realities – a comparative approach from the RICHFIELDS program*. In: *Proceedings of Measuring Behavior 2016*. (A Spink, G Riedel, L Zhou *et al.* eds), 10th International Conference on Method and Techniques in Behavioral Research, 25–27 May 2016. Dublin, Ireland.

Mikkelsen BE, Høeg ER, Mangano L *et al.* (2016b) *The Virtual Foodscape Simulator – gaming, designing and measuring food behavior in created food realities*. In: *Proceedings of Measuring Behavior 2016*. (A Spink, G Riedel, L Zhou *et al.* eds). 10th International Conference on Method and Techniques in Behavioral Research, 25-27 May 2016. Dublin, Ireland.

Provan KG & Kenis P (2007) Modes of network governance: structure, management, and effectiveness. *Journal of Public Administration Research and Theory* **18**: 229–52.

Reichert JL, Postma EM, Smeets PAM *et al.* (2018) Severity of olfactory deficits is reflected in functional brain networks. An fMRI study. *Human Brain Mapping* **39**: 3166–77.

Research 2 Guidance (2015) *mHealth App Developer Economics 2015. The current status and trends of the mHealth app market. 5th annual study on mHealth app publishing based on 5,000 plus respondents*.

Reuver M, Haaker T, Nikayin F (2015) Designing Viable Multi-sided Data Platforms: The Case of Context-Aware Mobile Travel Applications. *International Federation of Information Processing 2015*, M. Janssen *et al.* (Eds.):I3E 2015. LNCS 9373, pp. 354–65.

Schreieck M, Wiesche M, Kromar H (2016) *Design and governance of platforms ecosystems. Key concepts and issues for future research*. Research Papers 76, 24th European Conference on Information Systems, 2016.

Sharma S, De Choudhury M (2015) *Detecting and Characterizing Nutritional Information of Food and Ingestion Content in Instagram*, *Proceeding WWW 2015 Companion Proceedings of the 24th International Conference on World Wide Web*, pp 115–6.

Shoval N & Ahas R (2016) The use of tracking technologies in tourism research: the first decade. *Tourism Geographies* **18**: 587–606.

Simmon E, Sowe SK & Zettsu K (2015) Designing a cyber-physical cloud computing architecture. *IT Professional* **17**: 40–5.

Snoek H, Eijssen LMT, Geurts M *et al.* (2018) Advancing food, nutrition and health research in Europe by connecting and building research infrastructures in a DISH-RI: results of the *EuroDISH* project. *Trends in Food Science and Technology* **73**: 58–66.

Sobal J (1991) Obesity and socioeconomic status: a framework for examining relationships between physical and social variables. *Medical Anthropology* **13**: 231–47.

Sobal J, Bisogni CA & Jastran M (2014) Food choice is multifaceted, contextual, dynamic, multilevel, integrated, and diverse. *Mind, Brain, and Education* **8**: 6–12.

Stefler D & Bobak M (2015) Does the consumption of fruits and vegetables differ between Eastern and Western European populations? Systematic review of cross-national studies. *Archives of Public Health* **73**: 29.

Umhoefer C, Rofe J, Lemarchand S *et al.* (2015) *Earning Consumer Trust in Big Data: A European Perspective*. DLA Piper & Boston Consulting Group.

Vidal L, Ares G, Machin L *et al.* (2015) Using Twitter data for food-related consumer research: a case-study on "what people tweeting about different eating situations". *Food Quality and Preference* **45**: 58–69.

Weber I, Achananuparp P (2016) *Insights from machine-learned diet success prediction*. In: *Pacific Symposium on Biocomputing, 2016*. (RB Altman, AK Dunker, L Hunter *et al.* eds), January 4–8 2016. Big Island of Hawaii. https://doi.org/10.1142/9789814749411_0049, http://arxiv.org/abs/1510.04802.

WHO (World Health Organization) (2012) *Action Plan for Implementation of the European Strategy for the Prevention and Control of Non-communicable Diseases, 2012–2016*. WHO Europe.

Zhang L (2015) Specification and Design Method for Big Data Driven Cyber Physical Systems. In: *Progress in Systems Engineering: Proceedings of the Twenty Third International Conference on Systems Engineering. Advances in Intelligent Systems and Computing 1089* (H Selvaraj, D Zydek, G Chmaj eds). https://doi.org/10.1007/978-3-319-08422-0_124.

Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package 1
Deliverable 1.2

Final paper of RICHFIELDS

Authors:
Marc Jeroen Bogaardt, Giacomo Copani, Javier de la Cueva, Paul Finglas, Charo Hodgkins, Barbara Korousic, Bent Mikkelsen, Krijn Poppe, Golboo Pour Abdollahian, Jos van den Puttelaar, Monique Raats, Trond Selnes, Lada Timotijevic, Hennie van der Veen, Pieter van ’t Veer and Karin Zimmermann

Summary

The aim of this deliverable is to describe the main findings and lessons for the design of a Research Infrastructure (RI) on consumers' food intake and determinants, as part of a wider Food, Nutrition and Health Research Infrastructure (FNH-RI). These areas and their domains are described and detailed in its content and focus on the position and final RICHFIELDS' (Research Infrastructure on Consumer Health and Food Intake using E-Science with Linked Data Sharing) paper. The aim is to get the final paper published in Public Health Nutrition, which has already been contacted about this. They are interested in a paper about RICHFIELDS.

Diet-related, non-communicable chronic diseases such as obesity and cardiovascular diseases have been identified as key European societal challenges, as they pose a significant threat to the health agenda for the European Union (EU) population (WHO, 2016). The EU Horizon2020 Programme addresses healthy diets for the ageing European population, as well as the increasing relevance of environmental and social sustainability of these diets[1] . Research on determinants of diet and physical activity has been prioritised to align with research on healthy choices in diet and physical activity in an encouraging societal environment. Food choice operates at physical, biological, psychological, economic and sociocultural levels, which all operate simultaneously and interact. Therefore, to impact public health and disease prevention, a sound and dynamic scientific evidence-base on consumer eating behaviour and lifestyle is crucial (Bogart *et al*., 2018). Such evidence provides valuable insight for research, governments, civil society organisations and industries to adequately respond to the urgent health and sustainability challenges in the health and food domain.

The FP7 EuroDISH[2] has identified the need for research infrastructures (RIs) in the food and health domain that can advance research within, among and over-arching the so-called food system related to nutrition and health research domains. EuroDISH initial findings confirm the current disparate and fragmented health and food research infrastructure. Advanced research infrastructures have been identified as useful to help facilitate health and food research. However, none of these can be considered health and food specific. Based on European Strategy Forum on Research Infrastructures (ESFRI)[3] recommendations (ESFRI, 2016) for a food and health research infrastructure (RI), the Food, Nutrition and Health Research Infrastructure (FNH-RI) brings together existing food and health-related RIs, including the Consumer Data Platform RICHFIELDS" with the focus on Food and Health Consumer Behaviour and Lifestyle.

1 https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges
2 http://www.eurodish.eu/
3 https://ec.europa.eu/research/infrastructures/index.cfm?pg=esfri

Research on determinants of diet and lifestyle is crucial for an adequate response to the current epidemic of diet-related non-communicable chronic diseases. In order to respond to this challenge the objective of the RICHFIELDS project is to design a research infrastructure on food-related behaviour of consumers. The conceptual design of the RI illustrates a European data platform that connects data on food-related behaviour with its determinants in consumers collected by using apps and wearable sensors, from behavioural laboratories and experimental facilities, and from other RIs. It will contain interfaces to collect data, a meta-database, and a data-model to enable data linkage and sharing.

The RICHFIELDS project is made up of three phases, each consisting of three work packages, and an overarching methodological support work package. Phase 1 focuses on data generated by consumers (e.g. collected by apps and sensors) relating to the purchase, preparation and consumption of food. Phase 2 concentrates on data generated by organisations such as businesses (e.g. retail and procurement data) and experimental research facilities (e.g. virtual supermarkets), as well as existing food, nutrition and health data from other RIs/platforms. Phase 1 and 2 will provide Phase 3 with information and knowledge on data types and design requirements needed to elaborate on in the actual design. Phase 3 encompasses the design of the RI including potential business models, data integration and management systems, and governance and ethics.

The main findings that have been obtained under the project have been the necessity and the opportunity of the creation of a supranational RI related to FNH where all agents who intervene in this field could use a common scientific and technical language, sharing data in a platform under a consensual data structure which guarantee that scientific quality requirements and legal conditions are accomplished.

Contents

Introduction

Addressing multifactorial issues related to food security, climate change, urbanism and all other XXI century societal challenges requires easy access to high quality data from many different sources such as consumer behaviour, food availability, socio-demographics etc. Easily accessible open data based research infrastructures in the food, nutrition and health domain are a potential powerful source and tool for solving some of these challenges.

Achieving food and nutrition security is considered a crucial step towards addressing the global challenges of climate change, natural resource scarcity and demographic expansion (EC, 2016a). It is also central to addressing the growing threat of diet-related health outcomes such as obesity, cancer, cardiovascular diseases and undernourishment which have been identified as a key European societal challenge (WHO, 2012; JPI HDHL, 2015; MRC, 2017). The significance of food and nutrition security is highlighted within the United Nations' Sustainable Development Goals (UNSDG) (UN, 2015), which represent a commitment of the international community towards sustainable future. UN SDG 2 states the need to "end hunger, achieve food security and improved nutrition and promote sustainable agriculture" (https://sustainabledevelopment.un.org/?menu=1300), a goal which is unambiguously about eating. Addressing other goals also wouldn't be possible without explicitly examining the processes of food production and processing, distribution, consumption and waste management. The role of the consumer in promulgating the UN SGC and achieving food and nutrition security is recognised by the EU Horizon2020 Programme which addresses healthy diets for the ageing European population, as well as the increasing relevance of environmental and social sustainability of these diets, through its funding calls [https://ec.europa.eu/programmes/horizon2020/h2020-sections]. According to the high level conference for Food 2030 (European Commission, 2016a), addressing consumer behaviour in terms of food purchase, preparation, consumption and handling of food and related non-food waste streams must be a priority in tackling these global challenges.

It is therefore increasingly recognised that research on determinants of diet and physical activity and factors influencing adoption of healthy and sustainable diets must be a priority in order to achieve these policy goals [JPI HDHL]. Food choice is shown to be a manifestation of many factors operating at environmental, biological, psychological, economic and sociocultural levels within a dynamic and complex system of interactions (Marshall, 1995; Shepherd and Raats, 2006). Such complexity requires a multi-disciplinary focus and numerous high level EU research & innovation strategy documents have identified the need for a cross-sectoral approach to food and nutrition science –e.g. European Technology Platforms, multi-actor approach of Horizon 2020; Responsible Research and Innovation (Rome Declaration, EC, 2014). And yet, despite the urgent need for stronger and more integrated scientific base in this domain, the current state of knowledge informing our

understanding of food, nutrition and food choices is fragmented and in relative infancy (MRC, 2017). In particular, the recommendations that improvement and linkage of key infrastructure and platforms, such as national surveys and cohorts, omics including metagenomics, and deep phenotyping facilities and brain banks, to better support innovative nutrition research, as well as the establishment of internationally leading cross-disciplinary Centres of Excellence in integrative nutrition to strengthen both research and training in key challenge areas and enhance scientific networking and cooperation across institutions (MRC, 2017).

Even with the growing leadership in this domain being provided by, for instance, Joint Programming Initiative (JPI HDHL 2015), Food 2030 (EC, 2016a), World Health Organisation (WHO 2014), the barriers can be found in the relatively low status of the science (MRC, 2017), the lack of cross-disciplinary funding (EC, 2016a) and overall limited funding being allocated to the research into personal behaviour and environmental influences on food choices (McCarthy et al, 2011). Perhaps in recognition of these barriers, JPI HDHL (2015, 64) called for the formation of an international science hub which would strengthen the food and nutrition science base.

In order to foster international and sustainable research infrastructures, the European Strategy Forum on Research Infrastructures (ESFRI) is offering support for international scientists and innovators to develop top science research infrastructures (RI). The mapping of food and nutrition research infrastructures carried out under the auspices of an EC funded project EuroDISH confirmed the current disparate and fragmented status of food and nutrition science base: it has identified a clear gap in the infrastructure support for the research of determinants of eating (Snoek et al, 2018; Brown, 2017). Whilst there was evidence of well-established RIs in the domains related to food and nutrition (e.g. CESSDA, ESS, BBMRI), these were nevertheless not specific to food and nutrition; in addition, no dedicated RIs were identified in the domain of food choice determinants.

The need for the RI, that would collate and collect, standardise and harmonise data necessary for the study of food choice (such as consumer generated data; business generated data; research generated data related to the study of food choice), is now being addressed through the international effort across the EU, which brings together a multidisciplinary and cross-sectoral team within the project RICHFIELDS. This EC-funded effort responds to the EU ambition towards fostering Open Science –European Cloud Initiative, 2016 (EC 2016b); EOSC Strategic Implementation Roadmap 2018-2020, Open Access (2012/417/EU), Digital Single Market (COM[2015] 192 final)– and aims to design a research infrastructure, to stimulate the development of unique services for collecting, aligning and sharing data related to the determinants of food choice. To this end, technologies to collect, align and share real-time data on food-related behaviour and lifestyle have been assessed across the spectrum of behavioural domains related to food

choice, which include purchase, preparation and consumption. Ultimately, the ambition is to overcome the current trend of food-related consumer behaviour and lifestyle being studied in disciplinary silos, within short time frames and a limited, reduced social and physical context.

The aim of this deliverable is to describe the main findings and lessons for the design of a RI consumers' food intake and determinants, as part of a wider FNH-RI, aligning data on consumers' food intake and determinants. The design-specific solutions and recommendations that will be proposed are based on the studies and activities developed in RICHFIELDS, which have been focused on four main areas: (1) research user needs and the emergence of various technical and innovative solutions for collecting real-time data from consumers; (2) consumers in the wider food systems environment sand the mix of public/government organisations and funding versus the private agri-food sector; (3) Ethical issues of privacy/ownership of consumer collected data versus commercial interests and (4) the business and governance of the Consumer Data RI and access right to the data made available via central data platform. The experience obtained in these studies and activities will be used in the final paper to critically review the main steps involved in the proposal of a RI focused on the researcher needs, which can be enumerated as (1) definition and typology of user, and assessment of user needs; (2) definition and assessment of technologies for collecting real-time data from consumers; (3) identification and classification of data for the study of food purchase, preparation and consumption; 4) study of multi-stakeholder needs; 5) development of the prototype design, backbone and ontology - technological solutions; 6) business model development; 7) governance model development and financial sustainability; 8) socio-legal and ethical considerations. The paper will outline the methodology followed to identify problems and address solutions, including the range of research studies, case studies and stakeholder engagement exercises. It will highlight the design-specific solutions, and conclude with the generic recommendations for the development of any future international RI.

Methods and approaches

The methodologies employed included a range of quantitative (inventory/mapping and surveys) and qualitative approaches (case studies and stakeholder engagement activities) performed with stakeholders, both as potential users and data providers, to elicit the necessary information to inform the design of the proposed data platform and RI from the perspective of the data that may be incorporated into it and the needs of the potential users.

Stakeholders included consumers, businesses and those operating in a research function in either academia, existing RIs or laboratories/experimental facilities. The data types explored

included those generated directly by consumers and those held within both the research and business communities.

Underpinning design

In order to design the business model, qualitative methods were applied. Interviews with industrial stakeholders were conducted to explore the potential ways to integrate business generated data. Dedicated sessions of stakeholders engagement were also organised, through RICHFIELDS stakeholders workshops, to identify the main needs of different types of users of the RI and services they expect from RICHFIELDS as well as condition to exchange data with the RI.

In the design of the new RICHFIELDS Data Platform (RDP), we firstly prepared user requirement's specification that was structured with respect to the ISO/IEEE 29148:2011(E) guidance, which proposes the elaboration of the following requirements-related processes: i) Stakeholder Requirements Definition Process, and ii) Requirements Analysis Process (ISO/IEC 15288) or System Requirements Analysis Process (ISO/IEC 12207). This specification included user epics and user stories specifying major functions of the RDP.

Next, we developed new concepts and tools for creating the RICHFIELDS semantic data model for linked (virtually integrated) food and nutrition data of different types. We introduced the RICHFIELDS ontology, which origins from the QuaLiFY ontology developed in the FP7 project QuaLiFY[4] and enables semi-automatic learning of new concepts directly from web services. Further, using the ontology learning methodology, the RICHFIELDS ontology can be updated when new concepts appear, and this will happen when new apps that provide data are connected to the RDP. Semantic annotation using the RICHFIELDS ontology will allow researchers access to harmonized data that can be further used for more advanced analysis. In order to support the enrichment of data with semantics, a new methodology based on the POS tagging-probability weighted method was presented (Eftimov et al. 2017b). Furthermore, we presented a new methodology for knowledge extraction from collected data without annotated corpus, named drNER, which combines a terminological-driven Named-Entity Recognition (NER) and a rule-based NER method (Eftimov et al. 2017a). Both presented methods are especially important for semantic enrichment that together with the RICHFIELDS ontology will allow researchers to have machine understandable data from different domains that further can be linked and analysed.

Then, we prepared an inventory of standardisation requirements for the functional and technical design of the RDP. Standards for collecting scientific data, business data and consumer data as well as standards for data linkage and harmonization were aligned with

4 http://quisper.eu

the user requirement's specification and the semantic data model. Existing European and global data standards for product codes, electronic health records, global positioning, physical exercise data etc. were taken in account. Special focus was given to work programmes of big data standardization consortia.

Consumer generated data

To explore the type and quality of consumer-generated data of potential interest to the proposed RI, the range of consumer data currently generated via health and lifestyle smart phone applications and tools was established across the three domains of purchase, preparation and consumption. The aim of developing this inventory was to provide the basis for identification of the scientific, technical and legal/ethical opportunities/issues regarding the potential value and integration of consumer-generated food behaviour data within the proposed RI.

To establish a deeper understanding of consumers' perspectives on willingness to share their data with the proposed RI, an online survey was performed in 8 EU countries (France, Germany, Italy, the Netherlands, Slovenia, Spain, Sweden, United Kingdom). The results from this study (n= 1000 per country) identified; (1) the extent to which consumers are willing to share their food and health related data with publicly funded researchers, governments and industry, (2) differences in willingness to share by country, age, gender, education or socio-economic status, (3) the relevant predictors to willingness to share their data.

Business generated data

An exploratory qualitative approach utilising case studies was adopted to gain an in-depth understanding of purchasing and procurement data held within the business community. Four case studies, represented by experts familiar with their respective organisation or institutional generated data were interviewed. The case study approach facilitated exploration of the complexities within a real-world setting and investigated how different institutions collect data about consumer behaviour. The main focus was on the business to consumer (B2C) interface and the business to government interface (B2G) in Sweden and Denmark. The studies explored three important topic areas; (1) best practices of collecting data, (2) ICT technology used for data collection (2) stakeholder perspectives for sharing of data in data pools (Ofei *et al*: 2017 p. 239; Hondo *et al*, 2017 p. 452).

Existing Research Infrastructures

Existing Research Infrastructures, networks and tools in the food and health domain were also studied utilising an exploratory qualitative approach. Four case studies focussing on (1) food composition and food attributes, (2) standardized food intake from population based

survey, (3) clinical intervention, (4) consumer diet, health and lifestyle were performed. Approaches to data access, data linking, governance and business models were explored with a view to defining the potential connection of these existing RI with the proposed RI.

Laboratories and experimental settings

The digital transition in the food, nutrition and health has offered new possibilities for measuring consumer behaviour in technology assisted ways in experimental settings. As a result a new generation of "smart food labs" has emerged. To better understand the type of consumer research being generated by which types of experimental facilities and laboratories and for what purpose, a mapping exercise was undertaken. This focussed on identifying laboratories and other research facilities across Europe used for studying consumer behaviour with sensoric technology/under controlled conditions. Two of the facilities identified were selected as case studies and in-depth interviews performed exploring their structures, purpose and technical specificities, to better understand their needs and wants, but also the potential for their connection to the proposed RI (Mikkelsen *et al*, 2017, p. 268).

Multi-stakeholder engagement

Throughout the project multi-stakeholder workshops were held to present interim findings from the various research streams within the project and to obtain feedback from stakeholders to help refine the design of the proposed RI. In addition, these activities were focussed on identifying potential motivators and barriers to future collaboration with the proposed RI, this being fundamental for its success and sustainability.

A quantitative user survey (N=100) was also performed targeted at researchers and other potential users of the proposed RI in order to establish how they might use such a resource and the extent to which they value the various tools and services that could potentially be offered. Utilising the learnings from EURODISH and data collected from expert interviews/consultations and the workshops held within RICHFIELDS, 8 Key potential customer benefits related to data, data collection, digital assistance and research support were identified and used as the base of this survey. The questionnaire was supported by a subset of mock-up webpages of the potential platform.

Results section: Scientific underpinning of the design based on RICHFIELDS results

Data and technical considerations

The scientific reach of the proposed RI/data platform is dependent on the diversity of data available to it and these include:

- Research data from other RIs, Laboratories and Experimental facilities.
- Business data (e.g. data from retailers, public procurement companies, statistical institutions and market organizations).
- Consumer-generated data from APPS (Smartphone and tablet applications) and sensors.

As a result of the studies performed, a number of opportunities and challenges for the proposed RI/data platform have been identified.

Research Data

Case studies have demonstrated that structures are in place to facilitate linking between some of the existing RIs in the food and health domain and the RICHFIELDS RI/data platform and therefore data from these sources is possibly the most accessible form of research data for RICHFIELDS. However, the development of a RICHFIELDS ontology and the harmonization of entities, food classification and description systems will be fundamental to facilitate future data access/exchange between existing and new RIs. The development of authoritative materials and standards must be a fundamental component of the RICHFIELDS offering to establish best practice and to help shape the research community moving forwards thus making future data sharing activities easier. Some of this work has already been established so RICHFIELDS needs to work with existing RI networks/experts to build on these.

Laboratory/Experimental facility Data

Thirty nine facilities in Europe involved in consumer research in the food and health domain were identified (Deliverable 10.1). Based on the subset of case studies performed it would appear that the majority of data collected in the past by these types of facilities is proprietary and typically not formatted, standardised or stored in a manner conducive to sharing outside the original purposes of the research study undertaken. In addition, the diversity of data generating devices including video and audio results in a wide variety of data types and thus increases the difficulty of post-hoc data integration. However, that is not to say that in the future data from these types of facilities could not be incorporated into the RICHFIELDS RI providing that sufficient support is given to standardise their future data collection in such a way as to be more easily shared with the wider research community via the RICHFIELDS RI. This would involve the development of harmonised Standard Operating procedures (SOPs), data management protocols, including calibration/standardisation protocols and improved approaches to obtaining ethical consent at the outset of the studies for future sharing with the wider research community. There is great potential if the proposed RICHFIELDS RI can develop a smooth and operational

infrastructure that allows the different labs to cooperate, optimise on the use of their expertise and share some of the burden of operating their high cost facilities.

Business data

In terms of business data, the results of the interviews and literature study performed identified that the ICT landscape is fast-paced, driven by an increasing connectivity of devices, increasing numbers of mobile devices used by consumers and cheaper and better sensors. The proposed RI/data platform must therefore be flexible enough to be able to respond to this dynamic ICT environment, however, careful consideration is needed on a case by case basis about the extent to which the data captured is reflective of the proposed research concepts and of sufficient quality to be treated as a useful variable for the RICHFIELD RI/data platform.

The Case studies also suggest that data collection may be significantly impacted by business purpose (e.g. policies to control suppliers or for organic procurement etc.) which may limit the potential usefulness of the data for scientific purposes within the proposed RICHFIELDS RI. However, there is clearly value in obtaining access to data from retailers or market research organisations as this type of data typically provides a broader consumer perspective on day to day food activities. The proposed RICHFIELDS RI therefore needs to ensure data source diversity but balance this with a clear understanding of the value of the difference types of data generated within businesses. Furthermore, a number of retailers may have already developed APIs (Application Programming Interfaces) for sharing data (e.g. Tesco in the UK) and these are potentially quick wins for the RICHFIELDS RI in terms of data acquisition from business that could be most readily exploited.

Sector	Type of technology	Data capturing technology	Devices facilitating data capture	Type of data collected	Case studies
Retail	Consumer location sensing technologies	Geo-fencing	Smartphones, GPS-devices	Location data involving a location-sensitive device (eg. smartphones with GPS)	RetailNext (Aurora, Mobile Engage), Euclid (Traffic, Insight), Shopkick (shopBeacon), Brickstream (Brickstream 3D+), Axper (3D vision, Sentinel), PathTracker
		Wi-Fi	Smartphones, tablets	Location data of smartphones connected to Wi-Fi	
		Bluetooth Low Energy (BLE)	iBeacon-compatible transmitters, smartphones	Proximity data to Bluetooth beacons of enabled smartphones	
		Visual systems	Analog or IP cameras, infrared cameras	Visual tracking data	
		RFID Technology	Smartphone RFID reader, RFID sensors	Consumer real-time product choice and purchasing data. Aggregated shopper tracking data to determine shopping speed, purchasing speed, and geography of trips.	
		Combination of technologies mentioned above	Several sensors available that combines different data capturing technologies. E.g., Aurora from Retailnext combines video technology with BLE and WIFI.		
	e-commerce and mCommerce	Online analytic tools for personal computers	Smartphone, personal computer, tablet	Web browsing patterns and online shopping patterns (Cookie data), online purchasing data	Adobe marketing cloud (Adobe), Virtual stores (Walmart)
		Online analytic tools for mobile devices	smartphone, personal computer, tablet	Mobile phone data	
	Social media			Social media sentiment analysis data	Kellogg's tweet shop
	Point of sale technologies	Barcode Technology	Digital barcode scanner, Smartphone barcode app (mobile point of sale), self-service checkouts, tablets, NFC tags	Consumer grocery shopping data	GfK ConsumerScan "Mini-Danmark, Mobile Point-of-Sale (SCANDIT), NFC tags in Casino supermarkets (France)
		Other point of sale hardware	Payment terminals, weighing sensors, cash registers	Amount owned, weight, money transactions	

Sector	Type of technology	Data capturing technology	Devices facilitating data capture	Type of data collected	Case studies
		Cloud based Point-of-sale software	uses data from devices mentioned in barcode technology and other point of sale hardware		Epos Now, Lightspeed Retail, Revel Systems, Lavu iPad POS
		Traditional point of sale software	uses data from devices mentioned in barcode technology and other point of sale hardware (except smartphone barcode scanners)		AIMsi, AmberPOS, RetailSTAR
Market Research Organization	Automated Voice Response and Voice Recognition	Interactive Voice Response survey	Touchscreen, freephone, post-call transfer to survey line, computer aided telephone interviews, web, email and SMS	Consumer feedback on product purchased and used	Vision OneTotalRecall
	Digital Observation and video	Digital diary and video recording	Webcam, smartphone, tablets, video camera, or some other type of digital audio/video recording device.	Consumer can either speak into the camera to describe a situation or feeling, or can take us on a tour, so to speak.	Olinger digital video diary
	Geo-location	GPS technology	Smart phone using apps with image, video capturing and survey questionnaire and integrated location	Photograph and record in-the-moment data in a specific location.	SSI's mobile QuickThoughts® 2.0 app. Geo-Intercepts app with features such as: GeoValidation, GeoIntensity and GeoNotification®.
	Neuromarketing research	Neuromarketing Techniques	Smart phone, tablet and laptops using facial recognition and other neuro analytics software	Captures the expressions and emotions people exhibited towards using a product	Face Reader-Noldus IREACT and Eye tracking-One vision

Consumer generated data

Whilst we typically talk about data collected via APPS and sensors (e.g. Fitbit) as being consumer-generated, in reality unless the data is being shared directly from the consumer to RICHFIELDS this type of data must also be considered business data. Use of this type of data for research purposes is somewhat fraught with the same limitations as that of other business data in that the purpose for its generation may impact its usefulness. However, again, a number of these APPS have developed APIs and there is the potential to capitalise

on these for data acquisition by the RICHFIELDS RI/data platform. In addition, there is significant potential to connect to developers/owners of AGGREGATORS already in the marketplace for the further development of the RICHFIELDS technical data infrastructure and to facilitate access to a wider breadth of consumer data. Careful consideration should again be given to identifying datasets that are of scientific relevance and sufficient quality to support the proposed research concepts in the RICHFIELDS science case since the research conducted within RICHFIELDS has identified a number of limitations but also opportunities with respect to consumer-generated data collected via commercial APPS:

Limitations

The main limitation from a scientific perspective with respect to consumer-generated data in the purchasing domain is that it does not identify whether the purchased food is consumed or not, nor does it identify the individual that may actually consume the food. The food may well be consumed by someone other than the purchaser e.g. family or friends. In addition, this data does not typically differentiate between intention and actual purchase and as such is not really a proxy of consumption. As a result, strong connections to public health outcomes at an individual level are limited if this type of consumer-generated data is utilised in isolation of consumption data. Whilst food purchase data is able to provide some understanding of consumer preferences/habits i.e. the types of foods, food retailers and restaurants that may be "on a user's mind" or that they utilise most frequently and can provide insight on food spend per week, month or year, its value is potentially limited: it cannot be used to track the behaviour associated with the purchase (e.g. the extent to which the purchased food is consumed, shared with others or results as waste).

Similarly with consumer-generated food purchase data, the degree to which consumer-generated food preparation data can act as a 'proxy' for intake is questionable. Whilst the data reflect consumers' motivation to gain knowledge and to develop skills in food preparation, the degree to which this is translated into intake cannot be directly drawn from the data in its current form. At best, it describes an intention to purchase or intake certain foods and/or meals. Nevertheless, if it is possible to link food preparation and food consumption data for a single individual through the data linkages, this may provide an invaluable insight into the complex relationship between intention and actual behaviour.

In contrast to the consumer-generated food purchase and preparation data the majority of food consumption APPS do collect data at the individual level, on a daily basis, at a specific moment in time and over a period of time. Therefore, from a scientific perspective data collected by these APPS has the potential to provide insight into habitual food consumption behaviours and how these change over time at an individual level. However, the problem with many of these APPS is that similar to traditional food diaries, from a user perspective they rely on extensive commitment/high levels of individual motivation to maintain such a diary, good recall, time investment and a degree of expertise to identify and input

appropriate food categories / products into the system. So, the issues of reliability, validity, perhaps social desirability (cheating) and drop out still exist.

From a scientific perspective, the unknown quality and validity of the food composition databases used to underpin these APPS and the non-standardised procedures for portion size estimation means that conclusions with respect to the relationship between food consumption and nutrition related diseases may be limited. Detailed research on the associations between specific nutrients and health outcomes may also be limited since majority of APPS in this domain focus only on energy and macronutrients.

Finally, and perhaps the most fundamental issue with consumer-generated food consumption data is that there is a particularly high prevalence of APPS with the aim of behavioural change. This intervention focus is likely to limit the ability to develop a true picture of people's habitual or typical food consumption behaviour, because they have been primed towards a behaviour change goal that by definition, may change their habitual practices.

Opportunities

Whilst the previous section highlights the limitations for the potential scientific use of consumer-generated data collected via APPS particularly for the three domains studied, there are still many opportunities for use of this data to help better understand food behaviour. In particular these include opportunities for RICHFIELDS to link with the existing AGGREGATORS established in the public domain which are already linking consumer derived data across a range of different APPS into a personalised overview for a consumer.

Furthermore, the scientific limitations highlighted above for consumer-generated purchase and preparation data are potentially possible to overcome by linking to data from the consumption APPS allowing a more extensive mapping of food choice and eating behaviour from preparation through to consumption for an individual. Although, it must be recognised that protocols for performing such linkages would need to be carefully developed. Unstandardized or undocumented food intake assessment procedures, data exchange protocols and formats, terms of use and privacy regulations, limit possibilities to integrate, process and share user-documented food consumption data in a scientifically robust way and therefore best practice guidelines, quality standards and protocols are needed for the effective integration of consumer-generated food purchase, preparation and particularly composition data in a scientifically meaningful way.

Consumer-generated food purchase, preparation and consumption data are not typically collected in isolation of other potentially relevant data. A vital source for better understanding the possible drivers and barriers for people's food purchase, preparation and consumption behaviour is likely to come from associations between these data and other relevant social, health and lifestyle data. This undoubtedly has the potential to give a more

valid picture whereby different data sets corroborate each to her to create a fuller, more accurate picture overall and the interconnectedness of APPS/tools now presents new opportunities to further enrich the food-related data from external sources. For example, it may be useful to gain domestic food purchase, preparation and consumption data from dedicated APPS and link this with health and lifestyle APPS for an individual. This combined data could be further enriched with demographic, situational and social context data collected through APPS such as Facebook, Twitter and Instagram. However, it should be noted that the degree to which users would find this interlinkage acceptable and be willing to share this type of extensive data with the proposed RI will need to be carefully considered and governed.

Consumers' willingness to share data

Results from the survey indicate that consumers are on the whole willing to share their food related data with either (1) universities or publicly funded research organisations, (2) governments and (2) commercial companies. However, the data demonstrates that they are statistically more willing to share with universities and this is due to a significantly higher level of trust associated with universities and lower perceived risk of sharing their data with this type of organisation when compared to the other two stakeholder groups.

Exploration of the terms and conditions for the APPS we reviewed within our studies, Deliverable 6.2 frequently identified that although users (i.e. consumers) are defined as being the owners of their data, the APP vendors have retained the right to sell the data. In the commercial world this has facilitated the selling and buying of data between organisations. The recent implementation of the GDPR seeks to offer some protection to users by ensuring that their personal data cannot be traded in this way without their consent although, it does not curtail the selling of pseudo or fully anonymised data. However, the recent revelations about the degree to which users' various data could be linked despite being considered anonymous/pseudonymous and the insight this could generate (e.g. Cambridge Analytica scandal) highlighted the deeper ethical and legal implications of repurposing of such data (Deliverable 13.2).

Despite what is happening in the commercial world, within the research community, there are established ethical criteria that need to be met in relation to informed consent from data owners. Without this consent, the data is not acceptable for use in studies destined for scientific publication. This means that without significant scrutiny on a case-by-case basis of each existing commercial dataset, the data they hold is not readily useable by researchers. This will ultimately limit the value of any commercial data that RICHFIELDS incorporates into the proposed data platform for scientific purposes unless the issues associated with informed consent are fully addressed (Deliverables 13.1 and 13.2).

Current data source status

Currently, data sources use different systems for data description and classification, therefore, the RDP will provide a service for data linkage and harmonisation based on a new food ontology. To achieve high flexibility, we have proposed a semi-automatic ontology creation. Each new data set will be enriched with metadata in a form of machine-generated tags (terms and concepts) that need expert confirmation. Those tags will be used to provide a harmonised access to data from different sources. As details of the data linkage and harmonisation are out of the scope of this paper, more details can be found in (Eftimov et al, 2017a; Eftimov et al, 2017b; Mezgec and Koroušić Seljak, 2017; Mezgec et al, 2018).

Platform architecture

In Figure 1, the RDP architecture is presented. The architecture consists out of 3 main elements, the user interface (UI), the API management system and the gateway. The UI, provides information on knowledge repositories, research protocols, and ontologies. The UI also provides the portal for data providers. These three elements do not provide long term storage of the data, but rather make the interaction possible.

The second element, the API management system, manages connections between external systems and the internal ones, it basically takes care of harmonization.

The third element, the Gateway is a separate application server that actually transfers data and knowledge.

Furthermore the architecture shows additional parts, such as the information system which can store donated data, the information management system that stores the metadata and provenance information. Moreover, it performs the semi-automatic creation of the RICHFIELDS ontology and the corresponding harmonization web service.

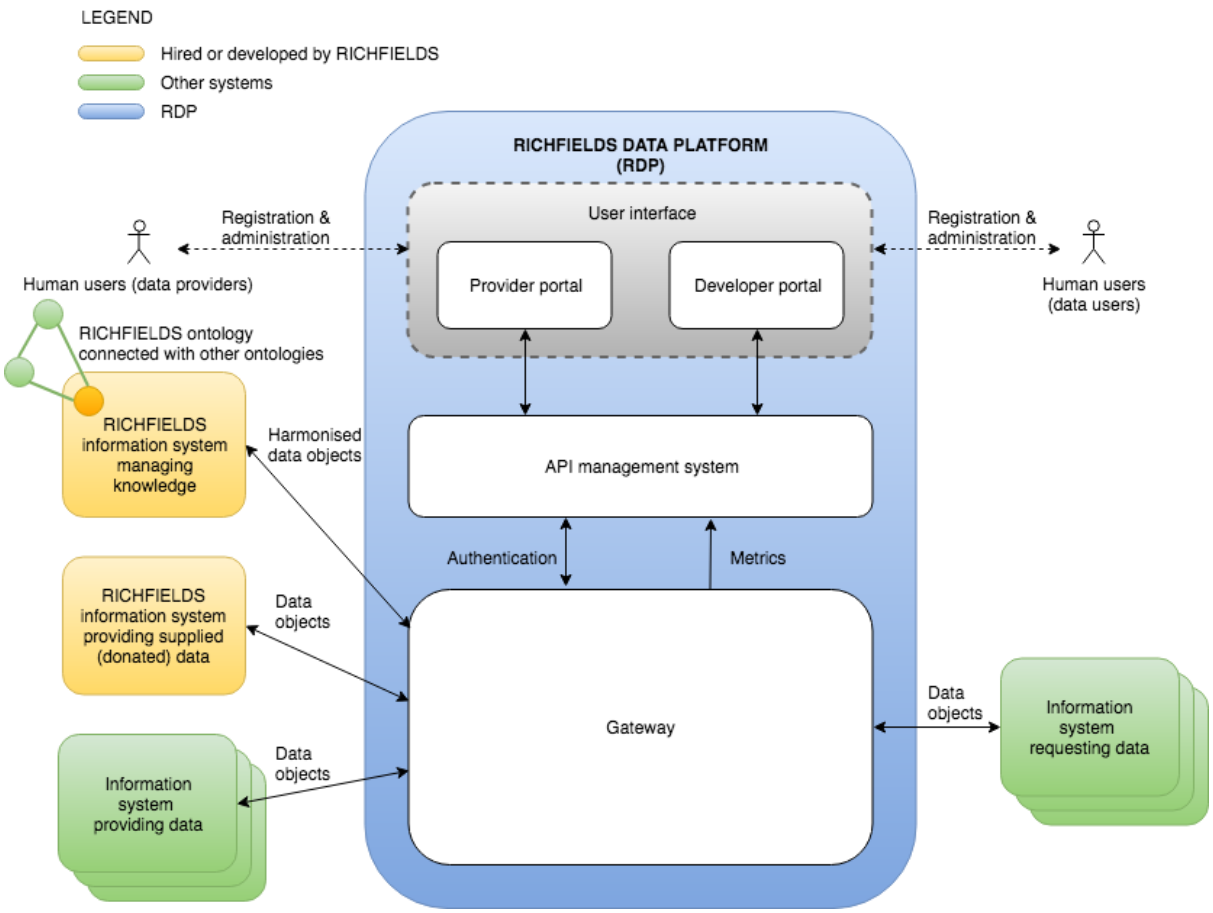

Figure 1. The RDP architecture.

Governance and ethical considerations

It is evident from the stakeholder engagement activities performed and study data collected that trust in the ability of the emergent RICHFIELDS governance structure to ensure that the data sharing activities it promotes are legally and ethically compliant, and that intellectual property rights and competitive advantage are not compromised is a fundamental requirement. Without this trust, the willingness to share data with the proposed RI/data platform will be severely compromised. This can be achieved by the establishment of a well-defined governance and management structure supported by scientific and business stakeholder advisory boards.

The variety of data sources potentially involved and the varying levels of consent they carry with them present significant challenges to the open access vision of RICHFIELDS. As previously explained, the datasets RICHFIELDS may obtain or connect to, will need to be evaluated on a case by case basis and the appropriate metadata assigned to them such that the possibility of non-compliant sharing from either a legal/ethical or data owner requirement is eliminated. Re-purposing of data needs to be carefully scrutinized and

controlled such that ethical compliance with the original participants' consent is always maintained.

Transparency is one means of engendering trust and it is recommended that the governance structure is fully transparent and the roles and responsibilities within it are well-defined. This is especially important in a public-private business model scenario when there is often differing drivers and a different set of guiding principles in terms of ethics. Similarly, processes for acknowledgement of the original data owner within any research publications arising from data acquired will be an important factor for data donators from the research community, but it may be that data donators from the business community prefer to keep their involvement less visible and therefore balancing all of these requirements in a robust and effective way will be an ongoing challenge.

In order to provide an appropriate organization able to manage the different decisions related to business, data sources, data-model or any other related to RICHFIELDS' inheriting organization, the resulting proposed structure is built up by three levels (Figure 2). The decision making level where the Assembly of Members takes decisions, with a Finance Committee and a Steering Committee and two advisory bodies; the Scientific and Ethical Board and the Stakeholder Forum. The executive level consists of the Director General and the staff, with a Management Committee where the nodes are members. The executive level has various groups and functions within the hub taking care of joint tasks, facilities and services, as ICT, a Data Protection Officer and trainings etc. The national nodes are working actively in these groups.

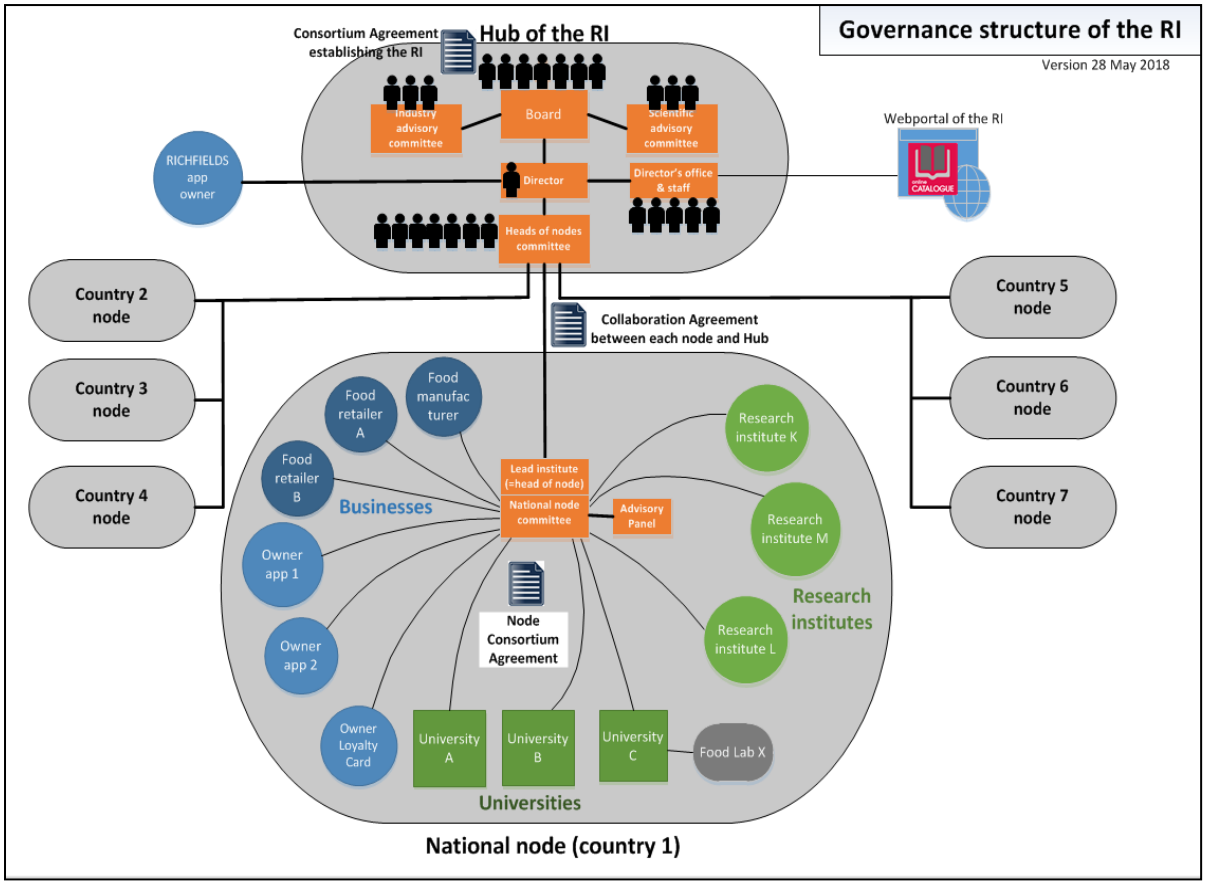

Figure 2. Governance structure of the RI.

Business Model considerations

Our stakeholder engagement activities have emphasized the importance of satisfying the needs of the potential data providers in the development of the proposed RICHFIELDS RI/data platform. Clearly, there are aligned objectives and therefore obvious benefits for other established and emerging RIs, laboratories and experimental facilities working in the public domain to share their data with the proposed RICHFELDS RI/data platform and these include; the potential for innovation by linking diverse datasets, standardization of protocols and data collection activities thus increasing on re-use and integration/enrichment of data/tools/models, increased visibility of and the potential to collaborate with other RIs, labs and experimental facilities operating in a similar research domain. However, it should be noted that many laboratories and experimental facilities undertake both publicly funded and industry/business funded research. It is unlikely that the business data they generate will be as readily accessible to the RI/platform.

Whilst many of the stakeholders from within the business community acknowledge their responsibility to improve public health as an incentive to data sharing it is unlikely that they will invest the necessary time and resources to actively share their data for purely altruistic reasons. Therefore, it is imperative that sufficient services are offered to further motivate

data providers from the various business stakeholder groups (retail, public procurement, market research, APPS and AGGREGATORS) to share their data with the proposed RI/data platform. This is especially important since access to data from the platform for them may well have to be quite limited compared to users from the publicly funded research community for ethical reasons.

The RICHFIELDS RI/data platform will need to establish itself as the authoritative, 'go-to' resource for data, tools and services within the domain of food behaviour determinants. This means that in order to be successful and appeal to the widest user base, the provision of authoritative and best-practice materials must be considered to be equally as important as the provision of the data connectivity and should therefore form an essential part of the service offering within the proposed RICHFIELDS RI business model.

Furthermore, to enhance its future potential to support high quality research it is important that RICHFIELDS provides training services either via online or physical courses and possibly even consultancy on a one-to-one basis to build the research community. This type of service offering, sharing expertise and best practice, will not only raise the quality of data being collected from consumers and by business for the future, but also enhance capabilities to perform high quality research within this domain. However, this will need to be costed in to the business model and the capacity built-in from the start.

Finally, the continued visibility of the proposed RI/data platform within the wider research and business community is key to its success and sustainability. The value of regularly engaging with and inviting feedback from users/stakeholders is an established way of ensuring a product or service to continue to satisfy ever-changing needs. It also helps to ensure continued engagement from data providers, data users and other stakeholders who are more likely to feel valued if they have a voice within the organization. Consideration within the business model should also be given to establishing an annual conference to disseminate the benefits of utilizing the proposed data platform in research activities. By communicating successful outcomes of research utilizing RICHFIELDS to the wider research community, the impact and credibility of the RI/data platform will be substantially increased.

Discussion section

Data and technical considerations

This design report is built on the RICHFIELDS project. It focuses on dietary behaviour and its determinants. This section on the scientific field however takes a broader scope, i.e. of the envisioned Food Nutrition and Health Research Infrastructure (Deliverable 13.3). It therefore also includes public health, environmental sustainability and food supply, as

elements that interact with consumer behaviour at the level of the food system. Thus, data content and its management by the platform must fit the needs of this scope.

The scientific field of the FNH-RI focuses on the consumers' dietary behaviour and as related to nutrition and (public) health, as well as environmental sustainability (see Figure 3). It ranges from the plate to the planet and public health. Dietary choices connect to their origin in agriculture and fisheries through the supply chain and to health outcomes via public health and prevention. The supply chain and consumer food choices are of direct relevance to public health, prevention, health promotion, and environmental sustainability. The scientific challenge is to create scientific breakthroughs and innovations that facilitate and enable the transition to a truly sustainable food system that delivers personal well-being, public health, environmental equilibrium, social justice and economic prosperity.

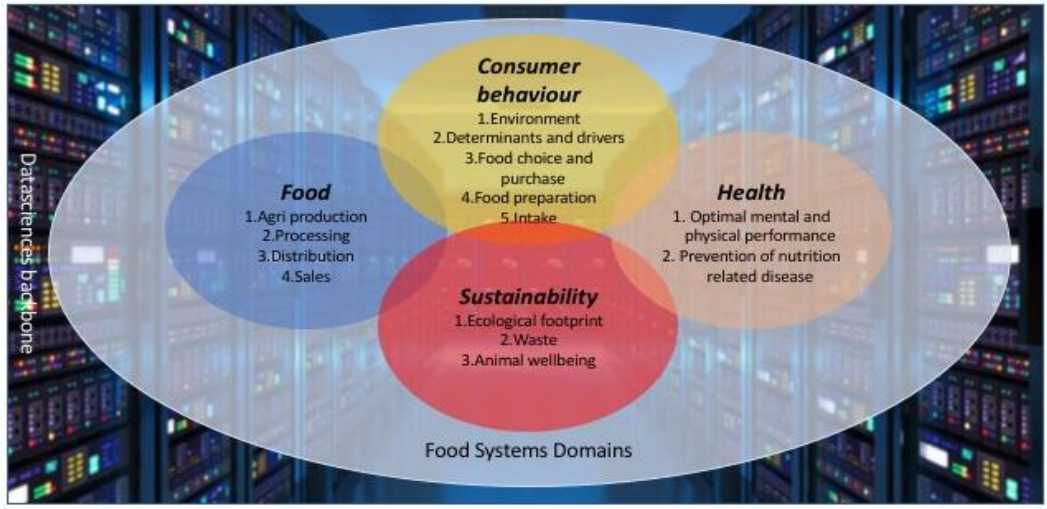

Figure 3. Simplified scheme of the four subdomains of the FNH-RI (Consumer Behaviour, Public Health, Environmental Sustainability and Food Supply). These subdomains are embedded within a food systems approach and their interconnectivity is achieved by an ICT-backbone connecting to the digitalisation of society and data sciences.

Currently fragmented data, information and knowledge prohibits policy makers and the private industry to develop effective interventions that help shifting current consumption patterns towards more healthy and sustainable diets. To create scientific breakthroughs and innovations, top level research is needed at the intersection of dietary behaviour, its determinants, public health and environmental sustainability:

1. Dietary behaviour, its determinants and food supply at the consumer level. This refers to in depth research on dietary behaviour of consumers in the context of their social and build food environment, as enabled by the RICHFIELDS-RI design in this report (Figure 4).

- Dietary behaviours relate to what, when, and where people eat their foods. This data usually derives from a dietary assessment methodology, preparation data (including recipes) and purchase data. This data is fundamental to subsequent assessment of nutrient intake profiles, nutritional adequacy, assessment of risk, and establishing norms (PRIs, by EFSA) and dietary guidelines (FBDGs, by national authorities). Moreover, it is crucial to assess environmental footprints (by LCA) and developing mitigations strategies to reduce GHGe.
- The drivers and determinants of dietary behaviour relate to the why and how of food consumption (Snoek et al. 2018, RICHFIELDS, 2015-2018). Consumer behaviour is determined by psychosocial and biological factors and embedded in the evolutionary heritage. The interplay between the behavioural and physiological determinants of dietary habits is at the centre of the RI. Knowledge of the interplay is crucial to develop effective strategies for behaviour change.

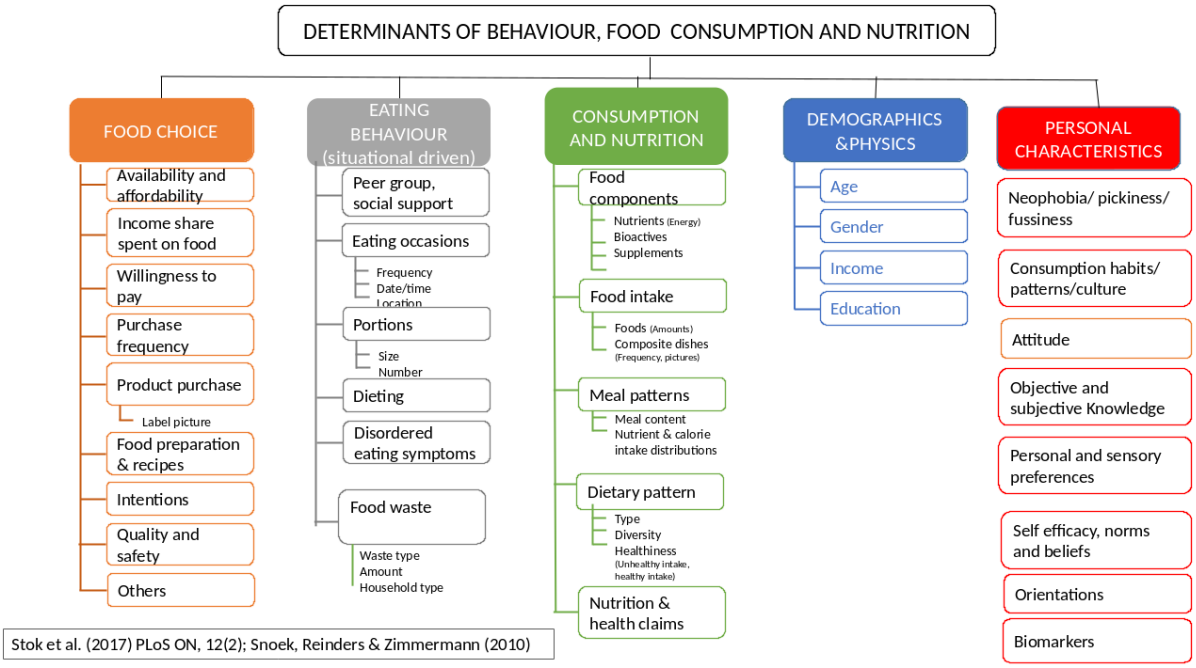

Figure 4. Determinants of behaviour, food consumption and nutrition.

The digital revolution creates unprecedented possibilities to assess dietary patterns and behaviours and patterns, its drivers and determinants and the food supply chain in a multidisciplinary way in real time. However, collection of data takes place in many different formats and as a result there is a great need for quality assurance and standardisation and incorporation of big data approaches in order for those data to create value for developers of apps, sensors and wearables. A research infrastructure has an important role to play.

2. Public health, environmental sustainability and food supply. This relates to the extension of the scientific domain from the scope in RICHFIELDS to the FNH-RI.

    - Public health and public health nutrition are built on the causal relationships between the diet, consumers' health, well-being and health as based in nutritional epidemiology, intervention trials, molecular, (epi)genetic and (patho)physiological mechanisms, and the cognitive and neurological basis of sensory sciences (Snoek et al. 2018, RICHFIELDS, 2015-2018). This data is crucial to long term strategies to combat the triple burden of malnutrition (micronutrients, obesity, non-communicable diseases) and to achieve healthy lives in an urbanizing and ageing society.

    - Environmental sustainability refers to food production within planetary boundaries (Rockstrom, 2009), environmental footprints from e.g., GHGs, land use, food processing and food waste. Moreover, mitigation strategies (alteinnovative production systems) may affect nutritional quality health, e.g. by altering essential trace elements via plant production systems; fats/carbohydrates in staple foods, or fatty acid composition of farmed fish. This data is crucial to identify strategies that balance health and environmental effects of food system changes.

    - Food supply. The food supply chain includes production of agricultural commodities, storage and distribution, processing and packaging, retail and marketing. In addition to its causal connection to public health and sustainability, the supply chain also leads to a redistribution of food over societal subgroups. The resulting accessibility and affordability of foods by farmers and citizens, in rural and urban areas across the globe impacts, food prices, food quality, food choice, dietary quality and the socio-economic distribution of health within populations.

Here, the digital revolution creates unique opportunities to connect data on health, sustainability and social disparities to dietary behaviour and its determinants. By taking advantage of connected data sets and the features of new types of sensors and wearable devices unique opportunities are offered to study the direct feedback loops of consumers behaviour simultaneously and their optimal interplay with long-term strategies for public health and the environment.

Data sources

Although the relation between food, health and sustainability is a very hot topic, finding or collecting the right data on food consumption and its socio-psychological determinants still is a hard job, either by re-using data or collecting data yourself. Data are often fragmented, national and project-driven, not standardized and not harmonized across countries and

therefore often of limited value. Data on food consumption often tends to focus on either dietary intake or socio-psychological determinants rather than a combination of both research domains and they are often collected in experimental settings instead of in real life settings. On the other hand, the choice to collect data yourself is not straightforward. It is a time-consuming process which might delay your research and for many researchers it is complex due to many regulations related to data collection and storage (FAIR, new regulations on privacy and data security, open access). To comply with all these rules is for many researchers not a topic that they get very enthusiastic about. Finally, for respondents filling in questionnaires on food intake is very detailed which makes it hard to find sufficient respondents.

The RICHFIELDS Data Platform (RDP) aims to collect, link and harmonise, analyse, store, and deliver food- and nutrition-related data (henceforth data) and information to various stakeholders (users). Data is going to be collected either by other information systems (ISs) connected with the RDP or by individuals collaborating with RICHFIELDS, and may be of any type, i.e., structured, semi-structured, unstructured; small, big; open, linked open; raw, processed; aggregated, disaggregated; with or without microdata (personal data about an individual), etcetera.

We have identified several types of ISs that could provide data: scientific clouds (e.g., European Open Science Cloud, Zenodo, FigShare), server platforms (e.g., Quisper, EuroFIR, GS1 GDSN, EFSA, MetroFood, CORBEL), application servers (e.g., RICHFIELDS, PRECIOUS, FitBit, Twitter), application servers analysing big data, smart food labs, and Internet of Things/Foods platforms. These data sources are mainly focused on heterogeneous data about food intake and its determinants. The *good features* of data provided by the mentioned data sources include: high diversity of data related to food, nutrition and health of consumers; data collected in those data sources can be provided in standardized formats (e.g., MS Excel reports, pdf documents, xml files); there are constantly appearing new ISs that collect data; and public releasing of data has become a positive global trend. We identified also a few *gaps*, like in general consumers use apps and collect data for a limited period of time (there are only few apps in heavy use); there are not many data sources that provide data via web services (less than 15% of data sources identified by RICHFIELDS).

Currently, the GDPR (Deliverable 13.2) stipulates that personal data may be transferable if the legal conditions set forth in the norm are accomplished. With this respect, we suggest that a new RICHFIELDS mobile application is developed that could empower consumers and researchers alike to join the collection and harmonisation of food-, nutrition-, and health-related microdata. In this way, data collected by other ISs linked to the RICHFIELDS mobile app can be transparently collected and securely managed.

Laboratory data

The RF design study examined in detail the potentials and constraints in sourcing data on consumer behaviour and food choice using smart labs, business environments and consumer applications to capture patterns of consumer behaviour and food choice. The study revealed that the new connectedness that has followed in the wake of digitalization offers new opportunities for connecting datasets that have not previously been connected to create new insight in consumer behaviour and thereby contribute to better solutions to currents societal challenges related to the food system. The RF study has also indicated that consumers widely support the idea of sharing data with science in the form of data donation, data philanthropy and citizen science. The RF study also show that there are obvious opportunities related to sourcing valuable data through business environments such as food retail and food service. And the RF study also shows that smart food labs for the study of food choice and behaviour under controlled or semi controlled conditions can offer new insight and offer new solutions when it comes to design of food environments. The RF study however underlines the need for creating a firm and smooth infrastructure around these smart labs in order to get the most out of their offerings.

Metadata

In order to provide easy access to quality data, the RDP aims to enrich all data elements above mentioned with metadata (data about data), which includes provenance data. FAIR principles (Wilkinson et al, 2016) and Open Science principles (Elbaek and Nielsen, 2013) will be applied to organize and define the metadata and provenance information and to make them accessible to a broad range of researchers and research users. Metadata including provenance data is important for the data linkage and harmonisation. Metadata is generally used to describe any characteristic of the content of the dataset, including provenance information, which provides derivation, transformation history and places of origin.

Governance

General model: the hub and the nodes

Building on EuroDISH, the lesson from RICHFIELDS is that a hub and nodes model is the best way to work within such a distributed network where the participants are spread out geographically. The hub then manages and coordinates the operations of the RI, while the nodes are national representatives of European countries (EU member states and former EFTA countries) with a membership in the overall foundation. The Foundation, the RI is an independent legally non-profit organization for the purpose of serving the research infrastructure, bears the formal name “STICHTING Food, Nutrition and Health Research Infrastructure” and is based in Wageningen, the Netherlands.

Independence is essential as it entails the ethical and legal commitments of the scientific community. The hub is the central manager and network coordinator of the facilities, resources and services offered by the RI. But the hub will merely promote communication within and among the national nodes. The nodes are free to choose their own research and organization, allowing them to use an informal type of organization. The nodes can have members from outside universities and public research institutes such as private research institutes, research labs from food companies or service organisations and research funders (like ministries, funding agencies, patient organisations). The node must appoint a Head of Node, representing its country with the signature of a minister. This allows the nodes to be members and take part in the decision making. The node should be on the national Roadmap but in absence of a roadmap, the recognition of a relevant ministry is sufficient.

External issues

The essence of the external relationships (Figure 5) is to have a demonstrated stronghold in ethics and law. Both privacy and right of control are matters addressed by the European Union's General Data Protection Regulation (GDPR)[5]. The GDPR imposes obligations on both the controller and the processor of data when it comes to processing personal data. Informed consent is key to legitimizing processing of personal data. Informed consent is a mechanism widely used in research to legitimize the use of an individual's data for a particular purpose(s). Purpose limitation is an important aspect of consents: processing is not allowed unless further data processing is compatible with the initial purpose for collection or for scientific or historical research purposes. That does not accommodate multiple data sets that are being used for purposes other than those for which the consent was sought. As the GDPR actively encourages to pseudonymise the RI will have this as a leading principle. Through pseudonymisation the data cannot be attributed by its users to a particular data subject (person) without the use of additional information and this additional information has to be kept separately from the processed data by the controller. This makes pseudonymisation different from anonymisation, in which this key is not stored separately. In general, the RI will only work with adequate standards, with standard data protection clauses or Business Corporate Rules (BCRs) which will require authorisation by the supervisory authority.

The Scientific and Ethical Board is pivotal to external relations and it will for this end have the right to advise on its own initiative and publish the decisions it takes. Its advice should be sought by the RI management on all relevant issue, including the collection new types of data from individual consumers and making data available for some research questions that have specific ethical aspects. As the outside world (like consumers on the social media) does

5 https://eugdpr.org/

not make much difference between types of data managed by the DI-RI, be it individual consumer data or general statistics, it is important for the reputation of the DI-RI that the ethics committee monitors all the activities of research infrastructure, not only those where individual data of consumers are involved. The ethics committee could also advice on the protocols on matters relating to data security, transfer of data to third countries, assessing the genuineness of a request by data users and the rules of operation in the event of requests that may be ethically dubious or questionable, data subjects' requests, and complaints procedures.

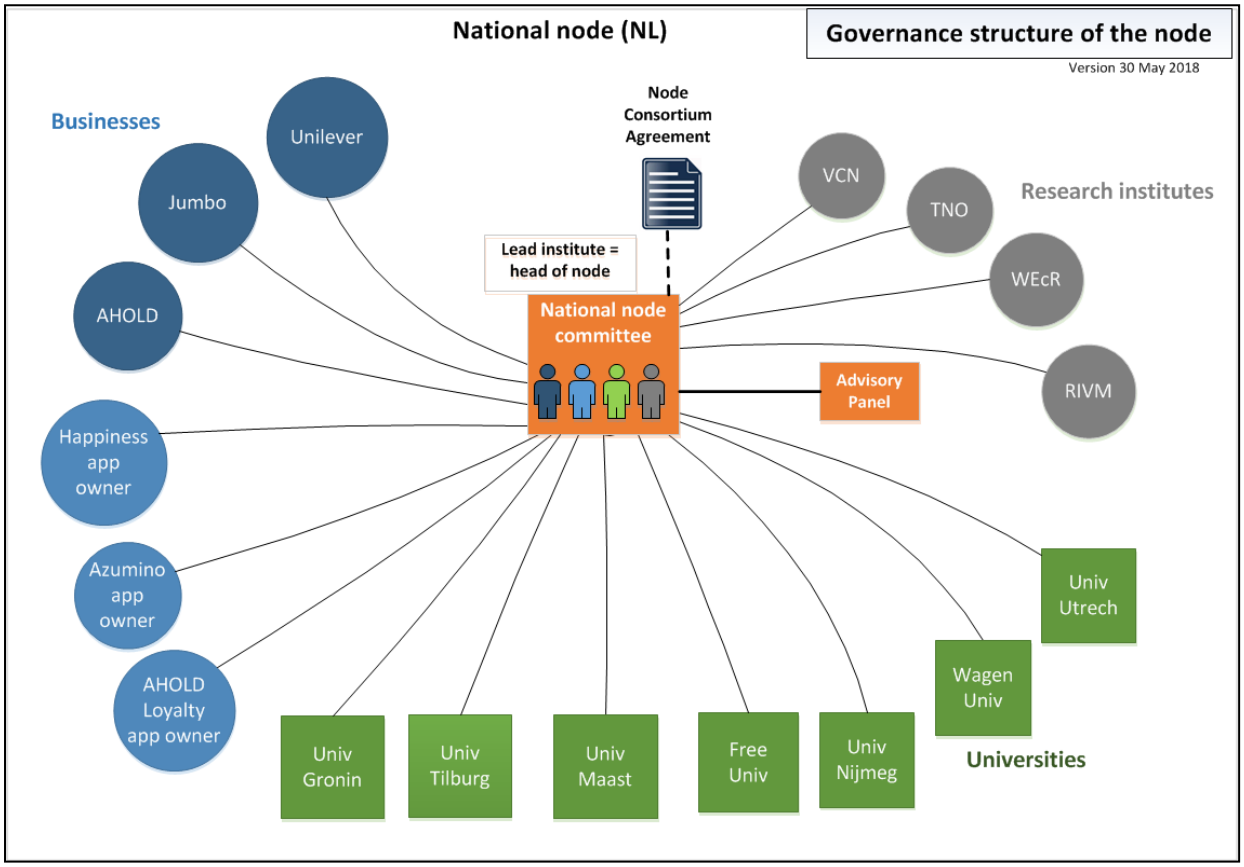

Figure 5. National node relationships

Business model

The business model of RICHFIELDS RI is designed based on a set of core value propositions that aim at satisfying the needs of researchers as the main RICHFIELDS users as well as other RICHFIELDS users including businesses, policy makers and consumers. There are five main categories of services offered by RICHFIELDS:

- Data-related services: Users of the RICHFIELDS platform can have access to high quality integrated data via a single-entry point. The RICHFIELDS platform offers access to data catalogues as well as various data sets that could be research data

from other existing RIs and business generated data sets from consumer apps and industries which are partners of RICHFIELDS. Moreover, RICHFIELDS offers its users access to consumer-generated data that are collected directly by RICHFIELDS app and are recognised as micro data sets. These micro data sets will be available to users under specific conditions through RICHFIELDS micro data labs. Figure 6 shows how users can have access to data in RICHFIELDS platform.

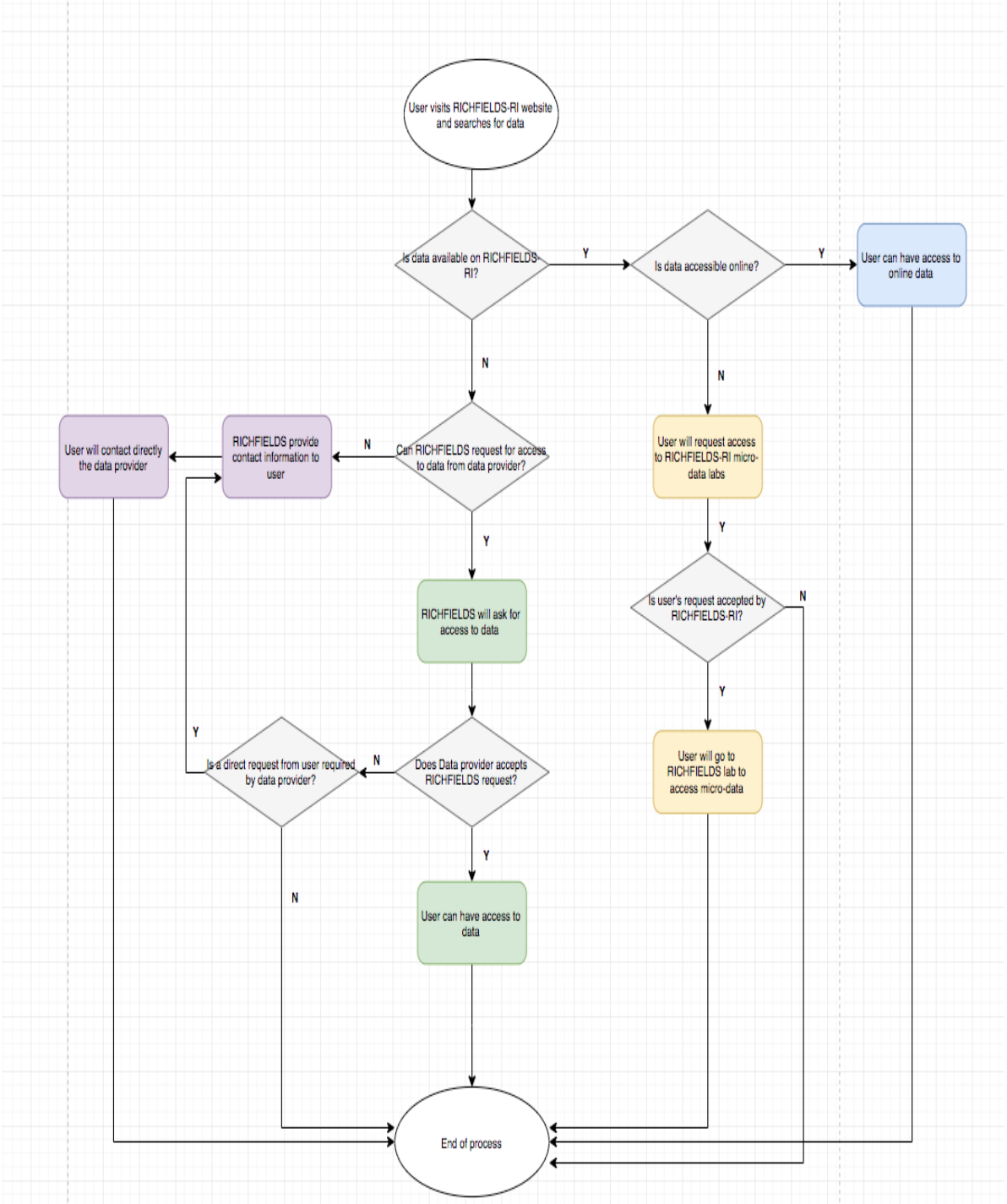

Figure 6. Process of access to data for users.

- Knowledge-related services and tools: The RI will provide access to standardised research protocols, ontologies, semantic models and vocabulary/thesauri to its users. To make data sets from different science groups and geographical areas compatible to each other it is necessary to develop standardized research protocols for data collection and preparation including ethical consent, legal compliance and ontologies. These standardised research protocols and data tools are major offerings especially for the scientific community since by implementing these tools, researchers will be able to use harmonised data collection methods and definitions to eventually generate standardised data that can be easily integrated with the other data sets. In addition, the RICHFIELDS infrastructure will give easier access to labs and physical facilities for RI Platform members through enhanced networking and community building, an RI internal member database with contact details and information about expertise/opportunities and, an RI agreement and guideline for Facility Sharing between facility owners and users.

- Consultancy services: The RI will offer consultancy services in particular for businesses and policy makers by performing on-demand analysis of data. Such a service targets the needs of users who do not need to have access to data but to the results of the data analysis based on their specific research question. In particular for small-medium enterprises which do not have a dedicated R&D section, consultancy services will be more practical.

- Training: Doing science drawing on the new potentials offered by RI represents a new mode of working for researchers and other RI users. Therefore, training such as extension services, summer schools, PhD courses is crucial. Thus, the RI will offer online and physical training sessions to key users of the RI around different topics such as using the RI data platform, using protocols and RICHFIELDS semantic, different analysis methods for data analysis and using micro data set labs.

- Community building and networking: The RI will provide networking opportunities to its users in particular the research community, by creating an online forum, establishing links between researchers and data suppliers, an annual RICHFIELDS conference and occasional workshops. Such a community will not only help to establish a consolidated relationship with the RI members but it also provides an opportunity to cement a more concrete link between different types of stakeholders and users such as researchers, businesses and policy makers.

Apart from these five categories other services will be offered to users of RICHFIELDS app (consumers) via provision of general and personalised advices on food, nutrition and health aspects. All service offering of the RI are depicted in figure 7.

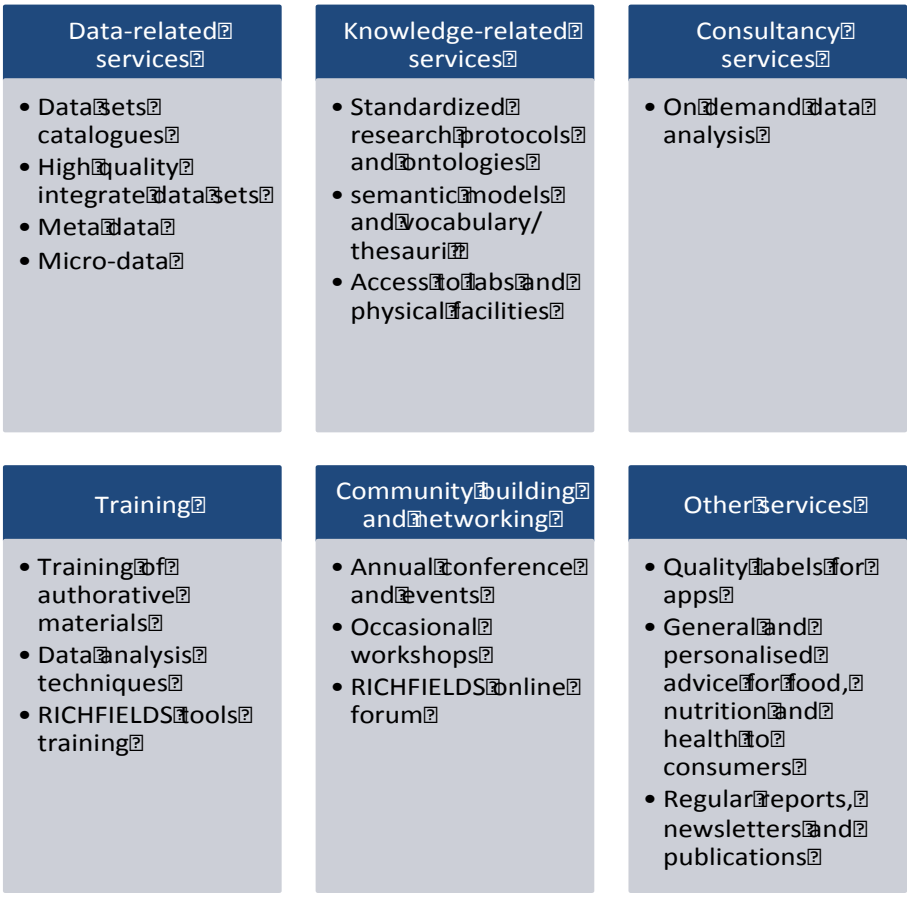

Figure 7. Service offerings of RICHFIELDS RI

To ensure its long-term sustainability, RICHFIELDS encompasses a public-private business model. The budget of the RICHFIELDS Central Hub will be provided by National Nodes (Member States) as well as public funding such as the ESFRI Roadmap programme from the European Commission. Financing of the national nodes and their activities is independent from the central hub. National nodes will be funded independently through national investments coming from public or private sources and they have their own governance and financing models. Full members of national nodes will have access to full service of the RICHFIELDS RI, which will also provide an initial core offer (Figure 8). Some services might be charged additionally for any member or user. Meanwhile, RICHFIELDS can have access to public grants provided by EU projects in case the RI gets involved in the projects as a partner.

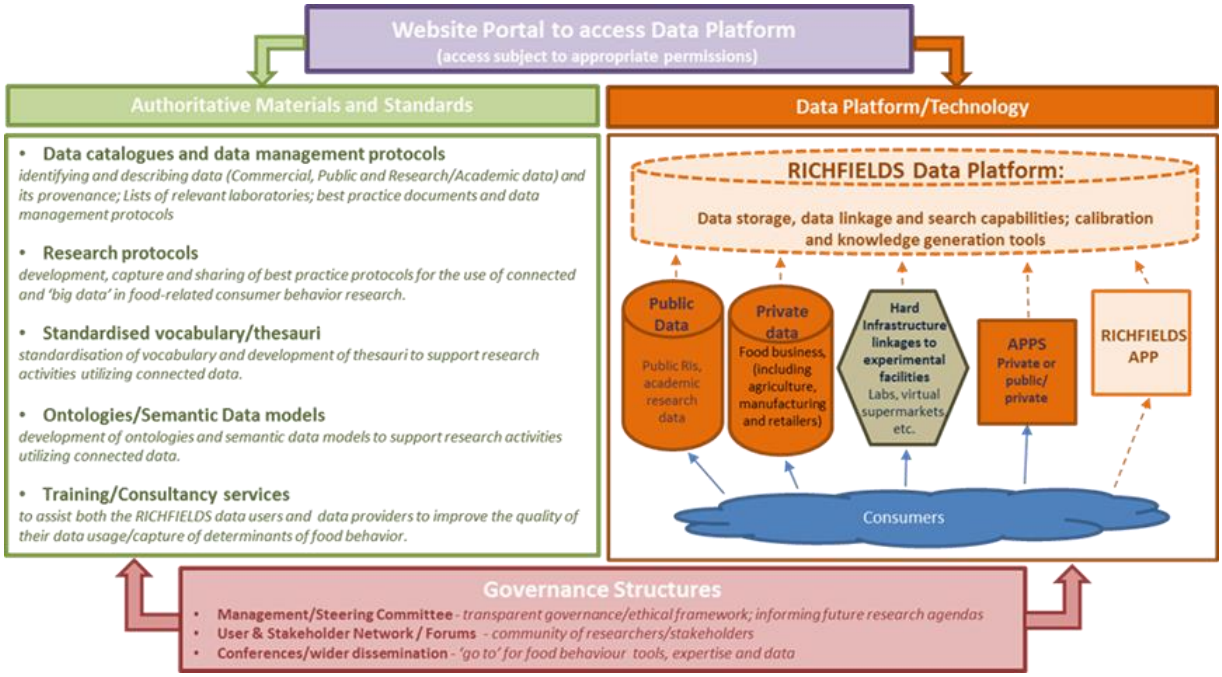

Figure 8. Core offering proposal for the Data Platform

While the initial funding to establish and run the RI in the first years will be mainly public, the private funding is the revenue streams that will ensure the RI sustainability in long-term after the first years of operation. Development of the RI will be gradual, according to its financial possibilities, as shown in Figure 9. Revenues will be generated through membership fees that industries pay to the RI in order to be able to be a member of the RI and use its services. In case a company is not a member of the RI, they can apply for particular services and thus will pay per requested service.

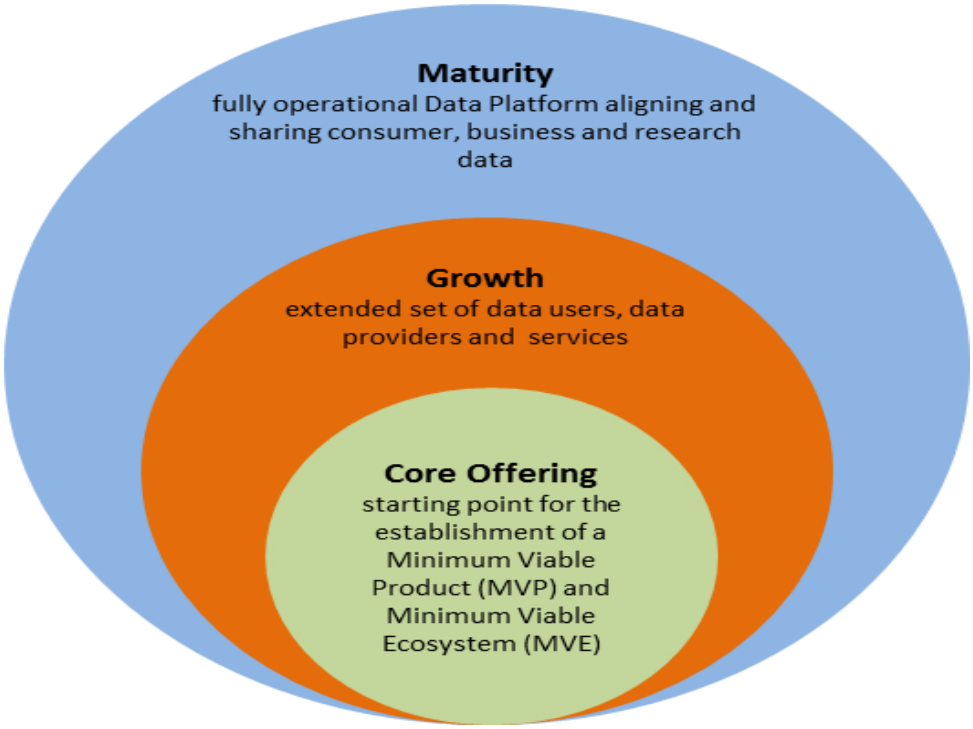

Figure 9. Proposed phases of development for the Data Platform

Conclusions

Through the efforts incurred in the design and development of RICHFIELDS' project, the main building blocks for a RI on consumers' food intake and determinants have been identified. As have been showed by the workshops held during the project and by empirical data, there is a need for a common language to be used by the scientific community in the transition from the analog to the digital age that has to focus in the usage of new interoperable digital tools and shareable data repositories. Simultaneously, the scenario where the RI lives is dynamic so the possibility of adaptation to new growths and potentials has been taken into account in the core of the designed model to allow its durability and sustainability.

References

Bogaardt, M.-J.; Geelen, A.;,Zimmermann, K. ;Finglas, P.; Raats, M. M.; Mikkelsen, B. E.; Poppe, K. J. & van't Veer, P. (2018). Designing a research infrastructure on dietary intake and its determinants. *Nutrition Bulletin*, https://onlinelibrary.wiley.com/doi/pdf/10.1111/nbu.12342

Brown K.A. et al. (2015). European food and health research infrastructures. EuroDISH: inventory and identified gaps and needs.

Brown, K, Timotijevic, L et al. (2017): Concepts and procedures for mapping food and health research infrastructure: New insights from the EuroDISH project. *Trends in Food Science & Technology* 63.p.113-131. http://dx.doi.org/10.1016/j.tifs.2017.03.006

Digital single market (2015) https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

Eftimov, T.; Koroušić Seljak B.; Korošec P. (2017a). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE* 12(6): e0179488. doi: 10.1371/journal.pone.0179488.

Eftimov, T.; Korošec, P.; & Koroušić Seljak, B. (2017b). StandFood: Standardization of Foods Using a Semi-Automatic System for Classifying and Describing Foods According to FoodEx2. *Nutrients*, 9(6), 542. doi: 10.3390/nu9060542.

Elbaek, M. & Nielsen, L. H. (2013, June 12). OpenAIRE Guidelines for Data Archive Managers v1.0. Zenodo. http://doi.org/10.5281/zenodo.6918.

EC. European Commission (2016a). European Research and Innovation for Food & Nutrition Security. Food 2030 High-Level Conference Background Document. DG for Research and Innovation Bioeconomy directorate.

EC. European Commission (2016b). Realising the European Open Science Cloud; Drafted by the Commission High Level Expert Group on the European Open Science Cloud. Directorate-General for Research and Innovation.

EC. European Commission (2014). Rome Declaration on Responsible Research and Innovation in Europe. https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf

ESFRI. European Strategy Forum on Research Infrastructures (2016). Strategy Report on Research Infrastructures. https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/esfri_roadmap_2016_full.pdf

Hondo, H; Kaunisto, E; Ofei, KT; Mikkelsen, B. E. & Hieke, S. (2017). Small devices for Big data – business driven smart technologies to collect data on consumer behaviour (#113), in

Mikkelsen, BE; Ofei, KT; Tvedebrink, TDO; Romani, AQ & Sudzina, F (editors): *Proceedings from 10th International Conference on Culinary Arts and Sciences, July 5-7th 2017 Aalborg University Copenhagen - Exploring Future Foodscapes*, p. 452.

JPI HDHL. Joint Programming Initiative. A healthy diet for a healthy life. (2015). Strategic Research Agenda 2012 - 2020 and beyond. https://www.healthydietforhealthylife.eu/index.php/hdhl-documents-2/key-documents/sra/download

Marshall, DW (1995). *Food choice and the consumer*. Blackie Academic and Professional, London.

McCarthy, M., Aitsi-Selmi, A., Bánáti, D., Frewer, L., Hirani, V., Lobstein, T., McKenna, B., Mulla, Z., Rabozzi, G., Sfetcu, R., Newton, R. (2011). *Research for food and health in Europe: themes, needs and proposals*. Heal. Res. Policy Syst. 9.

MRC. Medical Research Council, National Institute for Health Research. (2017). *Review of nutrition and human health research*. London. https://mrc.ukri.org/documents/pdf/review-of-nutrition-and-human-health/

Mezgec, S. & Koroušić Seljak, B. (2017). NutriNet: A Deep Learning Food and Drink Image Recognition System for Dietary Assessment. *Nutrients*, 9(7), 657. doi: 10.3390/nu9070657

Mezgec, S.; Eftimov, T.; Bucher, T. & Koroušić Seljak, B. (2018). Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment. *Public Health Nutrition,* Apr 6:1-10. doi: 10.1017/S1368980018000708.

Mikkelsen, B. E. & Ofei, KT. (2017). Measuring food behaviour the smart way - case insights from the implementation of Foodscapelab (#114), in Mikkelsen, BE; Ofei, KT; Tvedebrink, TDO; Romani, AQ & Sudzina, F (editors): *Proceedings from 10th International Conference on Culinary Arts and Sciences, July 5-7th 2017 Aalborg University Copenhagen - Exploring Future Foodscapes*, Published by Captive Food Studies. AAU, Copenhagen, p. 268.

Ofei, KT; Hondo, H; Kaunisto, E & Mikkelsen, B. E. (2017). Can business generated big food data be used to understand food consumption behaviour and can a research infrastructure be generated around such data?(#110). in Mikkelsen, BE; Ofei, KT; Tvedebrink, TDO; Romani, AQ & Sudzina, F (editors): *Proceedings from 10th International Conference on Culinary Arts and Sciences, July 5-7th 2017 Aalborg University Copenhagen - Exploring Future Foodscapes*, Published by Captive Food Studies. AAU, Copenhagen, p. 239.

Rockström, J. (2009, 26 October). Planetary Boundaries: Exploring the Safe Operating Space for Humanity [presentation] (PDF), Stockholm Resilience Centre, Biodiversity and Ecosystem Services, Master Class, Club of Rome General Assembly, 26 Oct 2009.

Shepherd R, Raats MM (2006). *The Psychology of Food Choice*, CABI Publishing.

Snoek H, Eijssen LMT, Geurts M. et al. (2018). Advancing food, nutrition and health research in Europe by connecting and building research infrastructures in a DISH-RI: Results of the EuroDISH project. *Trends in Food Science & Technology*. Volume 73, pp. 58-66.

UN. United Nations (2015). United Nations General Assembly Resolution A/RES/70/1 of 25 September 2015. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

Wilkinson, M. D. *et al*. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Sci. Data* 3: 160018, doi: 10.1038/sdata.2016.18.

WHO. World Health Organization (2014). *EU Action Plan on Childhood Obesity 2014-2020*.

WHO. World Health Organization (2012). *Action Plan for Implementation of the European Strategy for the Prevention and Control of Non-communicable Diseases, 2012-2016*.