Outline methodology for RICHFIELDS

Date delivered: M6

Author: Kerry Ann Brown & Lada Timotijevic

Deliverable lead beneficiaries: University of Surrey
<table>
<thead>
<tr>
<th><strong>Project</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Project acronym:</td>
<td>RICHFIELDS</td>
</tr>
<tr>
<td>Project full title:</td>
<td>Research Infrastructure on Consumer Health and Food Intake for E-science with Linked Data Sharing</td>
</tr>
<tr>
<td>Grant agreement no.:</td>
<td>654280</td>
</tr>
<tr>
<td>Project start date:</td>
<td>01.10.2015</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Document:</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>Outline methodology (facilitate knowledge transfer from phases 1 &amp; 2 to the RICHFIELDS final design in phase 3)</td>
</tr>
<tr>
<td>Deliverable No.:</td>
<td>D4.1</td>
</tr>
<tr>
<td>Authors:</td>
<td>WP 4 and all RICHFIELDS partners</td>
</tr>
</tbody>
</table>
| Reviewer: | Karin Zimmermann – Project Coordinator  
prof. dr. ir. Pieter van’t Veer – Scientific Coordinator |
| Start date: | 1.10.2015 |
| Delivery date: | 31.04.2016 |
| Due date of deliverable: | 31.04.2016 |
| Dissemination level: | PU |
| Status: | ongoing |

<table>
<thead>
<tr>
<th><strong>Change history:</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>Notes</td>
</tr>
<tr>
<td>001</td>
<td></td>
</tr>
</tbody>
</table>

Karin Zimmermann  
Project Coordinator  
prof. dr. ir. Pieter van’t Veer  
Scientific Coordinator
## CONTENTS

1. Executive summary ................................................................. 5
2. Project objectives .................................................................. 5
3. Work package 4 (methodology) ............................................. 6
   3.1 Objectives........................................................................ 6
   3.2 Partners ........................................................................ 6
3. Deliverable ............................................................................ 6
   4.1 Approach ......................................................................... 6
   4.2 User needs analysis ......................................................... 7
   4.3 Data architecture ............................................................ 8
     4.3.1 Concepts and definitions ........................................ 8
     4.3.2 Data quality ............................................................ 8
     4.3.3 Research question scope ....................................... 9
     4.3.4 Access and re-use of the data .................................. 10
   4.4 RICHFIELDS research activities ...................................... 16
     4.4.1 Inventories (WP 5-7 and WP 10) ............................ 16
     4.4.2 Interviews, focus groups and workshops ............... 18
     4.4.3 Case studies (or pilot actions) ................................. 21
     4.4.4 Workshop 1 ............................................................ 21
   4.5 Issues to address throughout RICHFIELDS ................. 20
5. Appendices ........................................................................... 23
   5.1 RICHFIELDS data sources and scope (work in progress)
   5.2 Desk research aid .............................................................
   5.3 Definition of purchase and consumption behaviours – previous work from the FP7 framework supported project CLYMBOL (health claims)
   5.4 Decision tree aid for in-/exclusion criteria to categorise tools collecting “big data” on purchase, preparation and consumption behaviours in phase 1 (Naomi Klepacz, WP 6)
   5.5 Interview, focus group and workshop aid
   5.6 Interview guide for WP8
   5.7 Further information (useful links and resources)
D4.1 contents

Aligned research infrastructure, research projects and events

European forums and inter-/national initiatives

Other resources

WP 11 needs internal note

WP 12 needs internal note

WP 13 needs internal note
1. Executive summary

This document outlines how knowledge generated in RICHFIELDS phases 1-2 will be collated to assist with the design of the RICHFIELDS food and health data platform in phase 3. The focus of this collation will be on, but not limited to, two key areas:

1) User needs: Identifying what RICHFIELDS is and for whom does it serve/ what need does it fill
2) Data architecture: Exploring how data will be categorised/organised in RICHFIELDS to ensure operability

2. Project objectives

The objective of RICHFIELDS is to design a world class research infrastructure on Food and Health Consumer Behaviour and Lifestyle, that will serve as an open access, distributed data-platform to collate, connect and collect, align and share innovative and existing data in order to enable researchers, policymakers and other stakeholders to develop, evaluate and implement effective food and health strategies both at the level of individuals and populations. This design will enable ESFRI, member states and other funding bodies to decide on the further preparation and implementation of the research infrastructure and will therefore also serve as building block for DISH-RI.

Fig. 1 RICHFIELDS concept of approach: phases 1-3 (RICHFIELDS DOA Nov 2015)
3. Work package 4 (methodology)

3.1 Objectives

- Facilitate the capture of knowledge and information generated in RICHFIELDS phases 1-2 (WP 5-10) for transfer to phase 3 (WP11-13) and the final RICHFIELDS design
- Provide insight into the constraints and challenges for the RICHFIELDS data platform (technical, conceptual and stakeholder/ user needs-related)
- Draft the outline of the RICHFIELDS data platform (D4.4).

3.2 Partners

- Surrey, UK (Kerry A Brown, Lada Timotijevic)
- DLO, NL (Karin Zimmermann, Dora Lakner, Kirsten)
- DIL, GE (Kerstin Lienemann initially, Sophie Hieke thereafter)
- WU, NL (Lan Ge, Pieter van’t Veer)
- IFR, UK (Mark Roe, Paul Finglas)
- All partner input.

4. Deliverable

4.1 Approach

This document was compiled following a series of initial scoping activities, including:

- Rapid literature review on European research infrastructures in food and health; European/inter-national data sharing, data classification and data quality initiatives (e.g., FAIR, ESFRI charters; Research Data Alliance).
- Consultation with external stakeholders (e.g., ENPADASI, DEDIPAC)
- RICHFIELDS phase 1-3 meetings and task protocols
- Work package 1-4 fortnightly meetings Jan-Jun 2016
- Consultation with phase 3 partners (WP 11-13) to identify design requirements
- Stakeholder platform meeting 2 Jun 2016 (D3.3)
4.2 User needs analysis

Rapid literature review and discussions with RICHFIELDS partners have identified a number of initial issues that will need to be explored during user needs analysis. These relate to …

Work package 4 will collate the information on user needs analysis generated via a series of activities within RICHFIELDS phases 1-2, including:

- Informal interviews with stakeholders at the RICHFIELDS stakeholder platform meeting 2\textsuperscript{nd} Jun 2016 (D3.3)
- Questionnaire survey distributed to existing research infrastructures on user groups and provision of food and health research services
- Focus group and workshop discussions with stakeholders in phases 1-2
- Questionnaire survey and follow-up interviews with stakeholders to identify user groups and user needs
- Workshops and stakeholder platform meetings throughout phases 1-3.

Refine and complete this section post-Milan partner/stakeholder meeting
4.3 Data architecture

Rapid literature review and discussions with RICHFIELDS partners have identified a number of initial issues that will need to be explored in relation to data architecture. These include, concepts and definitions; data quality; research question scope; access and re-use of data.

Information on data architecture will be collated in the following way:

- Inventory on WP 5-7
- Inventory on WP 10
- Focus group and workshop discussions with stakeholders in phases 1-2
- Workshops and stakeholder platform meetings throughout phases 1-3

Refine and complete this section post-Milan partner/stakeholder meeting

4.3.1 Concepts and definitions

In order to start developing a data type classification, it is important to provide preliminary definitions of some foundational constructs. The widely referenced Data-Information-Knowledge-Wisdom (DIKW) Hierarchy, otherwise known as the “Knowledge Pyramid” (Ackoff, 1985), is a central model of information and knowledge management, which specifies data, information, knowledge and wisdom in relation to each other. The model purports to describe the processes involved in the transformation of an entity at a lower level in the hierarchy (e.g. data) to an entity at a higher level in the hierarchy (e.g. information) (Rowley, 2007). Ackoff defines data as symbols that represent properties of objects, events and their environment and a product of observation. They are of no use until they are transformed into a form that is useable for a specific purpose. Information is contained in descriptions, answers to questions that begin with such words as who, what, when and how many, and is inferred from the data – therefore, the difference between data and information is functional, not structural. Knowledge is know-how, and is what makes possible the transformation of information into instructions. Wisdom/intelligence is an ability to exercise judgment in relation to the wealth of knowledge. Although this model is frequently contested, it is informative as it helps guide the thinking about the way in which we can connect different levels of generalisation from observation to inferences.

Fig 3: Knowledge Pyramid (Rowley, 2007)
This DIKW hierarchy has informed the broad understanding of data as: a) discrete, objective facts or observations, which are unorganized and unprocessed; b) items with no meaning or value because they are without context and interpretation; c) elementary and recorded descriptions of things, events, activities and transactions (Rowley, 2007). Data can vary not by the kind of information they provide, but the way in which they are recorded. They can be **structured** – relational or tabular, able to establish relationships between several variables, derived from apps/questionnaires; and **unstructured** - dynamic, complex, based on non-relational semantic structure, e.g., from the sensor output, device or system log, documents. This definition is reflected across disciplines. So for instance, computer science defines data types as:

"a particular kind of data item, as defined by the values it can take, the programming language used, or the operations that can be performed on it”

“The data type of a variable determines how the bits representing those values are stored in the computer's memory. When you declare a variable, you can also supply a data type for it. All variables have a data type that determines what kind of data they can store. By default, if you don't supply a data type, the variable is given the Variant data type. The Variant data type is like a chameleon - it can represent many different data types in different situations.” Microsoft

Within social science, data types are similarly defined in terms of the means of recording, which implies the degree of structuration:

“Different types of data are of relevance and may be covered by the portal a) Unit record data based on samples and questionnaires as data collection instruments, traditional survey data b) Register- or universe based aggregates, often collected through administrative procedures, stored in databases or as cubes. This may be referred as aggregate, regional or ecological data, these data are partly of a different methodology, partly collected for different types of units and partly of a relational character functioning to describe the contexts of individuals. c) Text-or qualitative data are of increasing value and abundance in a world where data are the direct outcome of administrative or other processes” CESSDA

“Big data” is often referred to as data generated on a vast scale; however big data are more accurately described by four elements, often referred to as the 4Vs: Volume, Velocity, Variety and Veracity (see figure 4).

“Every minute the world generates 1.7 million billion bytes of data, equivalent to 360,000 standard DVDs. More digitised data was created in the last two years than in the rest of human history. This trend and the mountains of data it produces is what we call "Big data”. The big data sector is growing at a rate of 40% a year. Handling big data requires increased technological capacity, new tools and new skills.” EC

“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.” IBM
These four v’s require consideration before “big data” can be translated into information. Data cleaning (or staging) is an essential part of extracting value out of the data and improving the quality of the data. In order to extract information, and therefore increase the value of the data, two issues must be considered: **measurement** and **inference** (Kreuter and Peng, 2014). Understanding the issue of measurement is important as it addresses whether data contain the right key variables and all covariates to answer the research question. Problems such as (unobserved) confounders, systematic measurement errors and the exact unit of analysis are important to clarify in order to enable inferences from the data (extract information from it). In addition, other questions about the measurement (and the data) need to also be assessable, such as sampling process, whether certain units appear multiple times, what do units present, etc. (Kreuter and Peng, 2014). When data linking is attempted, these questions need to be asked regardless of whether the data is structured (designed to collect specific type of information) or unstructured (big data), since the purpose of the data changes.

Irrespective of the size of the data, researchers must contend with the way in which data is generated, and this is partly what is involved in the process of “data classification”. Whereas with structured data, data-type considerations are primarily linked to the assumptions related to statistical models and the meta-data about the variables that are extrapolated from the raw data, unstructured data and the linked data presents a different set of considerations – those about confidentiality and privacy, among others. The reason for this is that, with the structured data, data collection is specifically designed with research questions in mind, which are translated into the concepts to be measured, which are in turn operationalised through a set of items developed to capture the relevant data. This process is documented, and it is this transparency that allows inferences about the quality of the data, and help understand the veracity and bias associated with the inferences made.
As “big data is generated as a by-product of human interaction with a computer or technology, its primary purpose is usually not addressing specific research question and engaging in inferential statistics. The data as such are a by-product of the “organic” human behaviour – the traces people leave in the context that enables capture of that behaviour. Thus, the two types of data vary by the purpose of data generation – which is answering research question in case of structured data, and addressing specific business-related or organisational aim or action, in the case of unstructured data. The main problem with the latter, therefore, is that measures often do not match the concepts necessary to answer the research question.

Fig. 5 Hashem 2015

This leads us to the conclusion that the first step in the process of developing a system of linked structured and unstructured data is to be focused on the following core issues:

- How is data generated? RICHFIELDS needs to classify data according to the source and the processes of data capture
- Purpose of data: Data type classification is dependent on the original purpose of data use, which in turn calls for clarity about the stakeholders/ users needs vis-a-vis the unstructured data
- Measurement: a problem with “big data” is that, “while they are case-rich, they are variable-poor” (Kreuter and Peng, p261), in other words, their value for capturing concepts that are necessary to address a research question, may be low. Clarity is needed about the extent to which the data captured is reflective of the research concepts, and can therefore be treated as a variable.
- Inference: a core problem of big data is that sampling is not defined a priori, and therefore the degree to which we can infer relationships is difficult to ascertain. The challenge is to assess under-/over-coverage (how representative is the sample), multiplication (a case appearing more than once within the sample) and the missing data, as well as privacy and confidentiality issues.
The Phase 1 team have developed a system for searching, recording and classifying the applications as the main source of consumer big data, which link up the research concepts of purchase, preparation and consumption (Level 1), the purpose of the data collected (Level 2) and the way data is generated (Level 3) – the data source (please see figure 4 for draft concept - completed version incomplete at time of deliverable submission).

Figure 6. (work in progress) phase 1 typology data collection tools related to purchase, preparation or consumption behaviour

4.3.2 Data quality

Assessments of data quality are a necessary stage enabling the extraction, cleaning and transfer of data into information. See below a selection of guidelines, from the many available, on which principles of data quality should be addressed in the process of data extraction/translation into information and/or knowledge. We review some of these below.

FAIR principles*ix

FAIR principles are designed to be flexible and general principles that allow data providers and data consumers (human or machine) to easily cite, find, access, inter-operate, and re-use “the vast quantities of information being generated by contemporary data-intensive science”. These principles were developed following a stakeholder meeting in 2014 with the Netherlands eScience Center and the Dutch Techcentre for the Life Sciences (DTL)*x FAIR principles are aligned with the Data Citation Principles that cover the purpose, function and attributes of citations. “These principles recognise the dual necessity of creating citation practices that are both human understandable and machine-actionable”. xiii
RICHFIELDS can align all data information systems with a flexible approach to FAIR principles, so that ultimately all the data that flows in and out of RICHFIELDS is:

- **Findable:**
  F1. (meta)data are assigned a globally unique and eternally persistent identifier
  F2. data are described with rich metadata
  F3. (meta)data are registered or indexed in a searchable resource
  F4. metadata specify the data identifier

- **Accessible:**
  A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  A2. the protocol is open, free, and universally implementable
  A3. the protocol allows for an authentication and authorization procedure, where necessary
  A4. metadata are accessible, even when the data are no longer available

- **Inter-operable:**
  I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
  I2. (meta)data use vocabularies that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data

- **Re-usable:**
  R1. meta(data) have a plurality of accurate and relevant attributes
  R2. (meta)data are released with a clear and accessible data usage license
  R3. (meta)data are associated with their provenance
  R4. (meta)data meet domain-relevant community standards

Research Data Alliance principles (RDA)\textsuperscript{\textregistered}

“The Research Data Alliance is an international organisation focused on the development of infrastructure and community activities aimed to reduce barriers to data sharing and exchange, and promote the acceleration of data driven innovation worldwide. With close to 4,000 members globally, RDA comprises individuals, organizations and policy makers representing multiple industries and disciplines, who are committed to building the social, organisational and technical infrastructure needed to reduce barriers to data sharing and exchange, and accelerating data driven innovation worldwide.”

The RDA Guiding Principles include:

- **Openness:** Membership is open to all interested individuals who subscribe to the RDA’s Guiding Principles. RDA community meetings and processes are open, and the deliverables of RDA Working Groups will be publicly disseminated.
- **Consensus:** The RDA moves forward by achieving consensus among its membership. RDA processes and procedures include appropriate mechanisms to resolve conflicts.
- **Balance:** The RDA seeks to promote balanced representation of its membership and stakeholder communities.
- **Harmonisation:** The RDA works to achieve harmonization across data standards, policies, technologies, infrastructure, and communities.
- Community-driven: The RDA is a public, community-driven body constituted of volunteer members and organizations, supported by the RDA Secretariat.
- Non-profit: The RDA does not promote, endorse, or sell commercial products, technologies, or services.

Overall, the assumption that “big data” and linked data are somehow superior to a standard, research driven data must be scrutinised with reference to the above principles, particularly from the point of view of various forms of validity e.g., how easy is it to assess the sample accessed (external validity claims); how much detail we can infer about the context of an observed behaviour (test validity claims); can we translate the inferences made from the data to the real life settings (ecological validity claims), how much we can infer about the concepts in question from the data (concept validity claims) and how much we can purport to be engaged in ethical science (ethics claims – not all accessible data are ethical) etc.

4.3.3 Research question scope

RICHFIELDS platform will be a unique open-access linked data platform that will provide an unprecedented opportunity to address the determinants of consumer behaviour relevant to food and health. The breadth of the domain (food and health) has been narrowed to three distinct instances of behaviour that are uniquely relevant and necessary components of food consumption: purchase, preparation and consumption. WP4 identified the most relevant initiatives currently engaged in developing the construct definitions and reviewing measurement approaches e.g., CLYMBOL (EC 2012-2016), DEDIPAC (2012-2016) ENPADASI and ELIXIR (both ongoing) etc. These are provided in detail within the Appendix, Section 7.3. Below we provide the top-level definitions of these concepts:

**Purchase** behaviour is defined as a process which goes beyond the act of purchase at the product shelf. Rather, it includes different factors which can influence the consumer before, during or after a purchase decision (cf. Solomon et al. 2013, p. 6). Levy, Weitz and Grewal (2014, p. 91) divide the purchasing process into 5 steps. It begins with the pre-purchase phase, which includes the recognition of a need (= motive), a more or less intensive information search determined by the current type of buying decision, and an evaluation of different options (cf. Howard and Sheth 1969, p. 25f). After the first phase, the purchase decision is made. Finally, the consumer evaluates his buying decision in the post-purchase phase (cf. Levy et al. 2014, p. 91). Below, we primarily focus on the pre-purchase, as well as the actual purchase phase, because the after-purchase phase is mainly covered by the study of consumption patterns in Task 3.4.

**Procurement**: Procurement has been added to account for the purchase related data flows that will be investigated in WP 8 and 10 that regard purchase behaviour between organisations (business to business or business to public organisations/government) rather than purchases for individual consumption. Procurement has been defined as the act of acquiring, buying goods, services or works from an external source, often via a tendering or bid process. It is favourable that the goods, services or works are appropriate and that they are procured at the best possible cost to meet the needs of the acquirer in terms of quality and quantity, time, and location xv.

Public procurement is the procurement of goods and services on behalf of government agencies and other public establishments. This type of procurement has been regulated and harmonized in the European Union by community law. The law is in place to promote open and fair competition and to prevent fraud, waste, corruption and/or local protectionism. xvii
**Preparation:** Final version currently unavailable at time of deliverable submission

**Consumption:** The Oxford English dictionary defines consumption as: “consume (verb) eat, drink or ingest”. There are three main areas of the food supply chain that provide opportunities to assess dietary consumption at either an individual or population level:

- Total population food available (domestic food production plus imports and minus exports) to provide estimates of population or average individual food/energy/nutrient intake per person.
- Household food available (food purchases, larder stocks, garden/allotment produce, gifts minus wastage) to provide estimates of household or average individual food/energy/nutrient intake.
- Individual food consumption (prospective or retrospective; actual intake, usual or habitual intake) to provide estimates of individual or average population food/energy/nutrient intake (Geissler and Powers, 2011; World Health Organization, 1996)

Consumption at an individual level is a complex process, which may be influenced by a number of physiological and psychological factors. It is important at the outset to decide which factors/indicators are to be included when evaluating methods that measure consumption.

**4.3.4 Access and re-use of the data**

An ongoing and continuous process of reflection is required about the ethical and policy implications of the RICHFIELDS platform, but also, the processes of work plan execution within the RICHFIELDS project. WP4 continues to liaise with Phase 3, specifically WP13 to include the issues of ethical relevance into the work of Phase 1 and Phase 2. In relation to the first point, the user needs elicitation through interviews and focus groups planned within Phase 1 and Phase 2 will deal specifically with the issue of ethics – privacy, confidentiality, risks and benefits of data sharing, as a means of informing access and re-use of data policy. In relation to the latter point, WP4 will ensure that workshops organised with key stakeholders throughout the life of the project will comply with the principles of Responsible Research and Innovation, defined as: “a transparent, interactive process by which societal actors and innovators become mutually responsive to each other, with a view to the ethical acceptability, sustainability, and societal desirability of the innovation process and its marketable products”. ([www.rri-tools.eu/about-rri](http://www.rri-tools.eu/about-rri)).
4.4 RICHFIELDS research activities

The following summarises the research activities designed in work packages 5-10 (phases 1-2) and how they will help to finalise the design of the RICHFIELDS platform in work packages 11-13 (phase 3).

4.4.1 Inventories (WP 5-7 and WP 10)\textsuperscript{xvii}

Inventories will be developed to provide an overview of the tools that create ‘big data’ related to purchase, preparation and consumption behaviours (WP 5-7) or generate data via laboratory services and facilities (WP 10). The starting point for the ‘big data’ inventories will be from the consumer perspective, regarding the “apps” available to purchase or download from platforms (e.g., itunes) that collect data either directly input by the consumer (e.g., participant height) or via a third party performing data analytics (e.g., downloads, page clicks etc.). The starting point for the laboratory and facilities inventory will be the EuroDISH mapping of food and health research infrastructures.

Inventories can align with the ELIXIR registry descriptions\textsuperscript{xviii} by providing at the minimum descriptive information on data sources, such as:

- The name of the data source, tool, laboratory or facility
- A URL for direct access or the ability to download
- A short and human-readable description
- Author list and list of publications describing the data source, tool, laboratory or facility (if available)
- Description of specific operations implemented and types of data processed or produced (in human- and machine-understandable forms (if appropriate and available))\textsuperscript{xix}

4.4.2 Interviews, focus groups and workshops\textsuperscript{xx}

Phase 1 and 2 will be conducting interviews, focus groups and workshops with RICHFIELDS stakeholders in addition to the stakeholder engagement activities being conducted at the project level (stakeholder platform meeting 1 & 2; stakeholder workshop 1-3).

Phase 1 focus groups will be focused on the user/stakeholder perspective by conducting focus groups with a) members of the general public and b) third parties with access to data analytics e.g., “app” developers. Phase 2 interviews/focus groups/workshops will be focused on the private/public non/research organisation perspective by engaging with those who have access to a) business to business data, b) business to government data, c) business to consumer data, d) aligned research infrastructure data and e) aligned laboratories/facilities. The qualitative information generated from the above will be particularly informative regarding stakeholder requirements/ user needs, best practices for data sharing and socio-legal aspects relevant to data quality, accessibility and re-use of data.

RICHFIELDS stakeholder groups and user needs have been suggested as a priori identified themes for designing and reporting qualitative results from interviews, focus groups and workshops. This can include reflection on the technical and conceptual issues related to:

- Terminology and data classification (e.g., taxonomies, semantics and their relationship)
- Research question scope (e.g., how are the necessary blocks or elements defined in RICHFIELDS to address determinants of dietary intake?)
- Data quality (e.g., criteria, data cleaning or metadata)
- Accessibility and re-use of data (e.g., ethics, legal, finance, intellectual property)
4.4.3 Case studies (or pilot actions)

Case studies across WP 5-10 will be selected from the inventories produced in RICHFIELDS, previous work produced by EuroDISH or based upon longstanding relationships between RICHFIELDS consortium members and commercial or research organisations. Case studies in WP 5-10 will allow a more detailed approach to focus on the technical components, interfaces and services necessary for RICHFIELDS to function.

Option to link phase 1 and 2 case studies (to be confirmed):

- Phase 1 and 2 can select a consumption “app” from the phase 1 inventory that matches the “app” to be investigated in WP 9 case study 4
- A map of connected “big data” can be created to view which “apps” or social media (photos of food, blogs, comments etc.) can be link with this one “app” (WP 5-7 and WP 9: preparation, purchase, other lifestyle behaviours etc.)
- The data trail can be followed to see where WP 8, WP 9, WP 10 aspects might interact with the “app” directly or the data generated by the “app” e.g., dietary assessment of consumption data.
- In addition, areas of particular concern can be investigated, such as resolving the challenges of - matching terminology and common classifications (e.g., resolving challenges of matching taxonomy between public and private institutions, social and natural sciences, front of house consumer or researcher facing and back of house ICT facing etc.) - identifying the necessary blocks or elements of RICHFIELDS - data quality - accessibility and re-use of data
4.4.3 Workshop 1

Date and location

Workshop 1 will be held for one day on the 27th Sep 2016 in Amsterdam

The main objective of the Workshop 1, as per DOA, is related to the specification and characterisation of the data that will be streamed into the Richfields Platform. Currently, both Phase 1 and Phase 2 of the project are addressing these issues. The questions being addressed within these two phases are:

- Sources of data: How is data generated (through apps and/or B2B/B2G)? What are the sources and the processes of data capture?
- Purpose of data: What is the original purpose of data and how is it currently being used?
- Types of data: How can we characterise the data? (e.g., self-reported vs. machine-generated)

However, in order to complete the picture of the data potential of the unstructured and structured data that would be feeding into the Platform, the two phases also need to address the following issues:

- Measurement: How to characterise the observations so as to enable inferences about the phenomenon being studied?
- Inferences: What are the limitations and possibilities for inferences from the analysis of the data, vis-à-vis the four constructs of purchase, procurement, preparation and consumption?

Measurement refers to characterising the observations in a way that enables inferences about the phenomenon being studied. The following issues ought to be addressed:

1. What is the core unit of analysis linked to the data? The unit of analysis is the major entity being analysed and clarity about it is essential in order to understand the limits to the inferences we can achieve from the data. In relation to the Richfields scope, the following units of analysis are possible:

   - An individual (captured through the measures of behaviour, cognitions, emotions)
   - A household (captured through the measures of e.g. energy consumption, purchase data, sensors)
   - An organisation (captured through the purchase/procurement data, legal/documentary data)
   - A geographical location (which may include a range of spatial data from the GPS-recorded movement data, through to epidemiological data on e.g. morbidity and mortality within a specific geographic location, waste production etc.);
   - Societal interactions (mainly captured through the internet interaction, e.g. social media discussion forums, recipe sharing, user ratings)
   - Other?
2. What is the measurement level – what are the attributes ascribed to the observations in order to make them “measurable”. The question for Richfields is how we can define the attributes and infer them from the data. This may be more easy in certain domains (e.g. frequency of intake), than in others (e.g. recipe sharing)

**Inference** is conclusions made based on the analysis of data. Clarity is needed about the extent to which the data captured is reflective of the research concepts, and can therefore be transformed into a variable. Questions to address include:

- To what extent can we infer about e.g. individual purchase behaviour from their online activity?
- What are the criteria that need to be to enable inferences from the data? This could be the issues of sampling, validity, ethics etc.

**Objectives of the Workshop 1**

WP4 will therefore propose that, based on the brief summary of the data-relevant issues above, the challenge for the Workshop 1 will be to discuss assessing the measurement/inferences scope of the data that will feed into the Richfields platform, vis-à-vis the core constructs of purchase, procurement, preparation and consumption.

Three issues should be explicitly addressed at the workshop, in order to provide the answer to the above overall objective:

1. **Measurement:**
   - What does the work completed by Phase 1 and Phase 2 tell us about the main units of analysis based on the data that will/may be streamed into Richfields?
   - What attributes or measurement values can be ascribed to data, in order to be able to draw inferences about it?

2. **Inferences:**
   - What inferences do the analyses at the given level allow, vis-à-vis the four constructs of purchase, procurement, preparation and consumption?
   - What are the criteria that need to be present in order to make inferences about the data vis-à-vis the four constructs (e.g. sample, validity, ethics)?

3. **Data linkages:**
   - How can we link the data from the same level of analysis (e.g. individual level of analysis) and the data from different levels (e.g. household and individual level of analysis)?

**Output of the Workshop 1**

The main aim for the workshop is to inform the deliverables on data quality and data type inventory (Phase 1) and feeding to the thinking about the linking of data (IC options, Phase 2).

The output from the workshop should be summarised in a framework to help assess the measurement and inference scope of the data that will be streamed into Richfields in order to provide a scientific case for the need for the data and the potential for data linkages.

**Participants**

- Scientists from different disciplines (e.g. psychology, epidemiology; nutrition; business; sociology, geography, anthropology, economics, politics/policy);
- Other users of data for research purposes (e.g. food data analysts; sensory/consumer scientists from industry)
- Data linkage specialists
4.5 Issues to address throughout RICHFIELDS

Below is a summary of the issues considered necessary to discuss, explore and undertake throughout the duration of RICHFIELDS:

- Identify stakeholder groups and assess user needs from different perspectives in order to better adjudge the origin, purpose and use of unstructured and/or big data
- Identify cost-effective ways to assess the data captured via different sources, as specified within DOW
- Standardised data classification across phases 1-2
- Synthesise the results of phase 1 (D4.2) and phase 2 (D4.3)
- How to identify, align and amend existing food & health ontologies so suitable for RICHFIELDS?
- Build a data model showing the working components and data flows of RICHFIELDS
- What is meant by data quality as it pertains to the concepts of purchase, preparation and consumption?
- How do the concepts of privacy and confidentiality inform the assessments of data quality? In particular, how can we assess the claims of validity and ethics?
- How to identify and align existing standards that judge quality?
- How to ensure user feedback systems to provide user experience or alerts for quality issues?
- Clarity about RICHFIELDS relevant concepts, the measurement approaches, possibilities and limitations for modelling the causal pathways (determinants) to dietary intake or health
- How to differentiate between purchase and procurement data collected in phase 1 (consumer-provided) and phase 2 (business-provided) data.
- The meaning of concepts (in the context of DIKW hierarchy) for the users of data needs to also be clarified. This can be done through user requirements. This will be address through phases 1-2 interviews/focus groups.
- How should ethics be integrated into the considerations of data classification and quality criteria (phases 1-2) in particular the issues of privacy, confidentiality, accountability and risk?


3 https://www.google.co.uk/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8&q=what+is+data+type Accessed on 22nd of April 2016


12 Stakeholder meeting in 2014 with the Netherlands eScience Center and the Dutch Techcentre for the Life Sciences (DTL), last accessed Mar 2016. Available from: https://www.force11.org/node/6062/#Annex4


16 Acknowledgements to Haris Hondo, WP 8 for the phase 2 protocol definition of procurement

17 Acknowledgements to Marcus Mariner, Anouk Geelen, Naomi Klepacz, Monique Raats, Susanne Ekman, Anne Normann and Sophie Hieke for WP 5-7 and WP 10 inventory protocols

WP 5-7 and WP 10 protocols detail a greater array of information than is detailed here.

Acknowledgements to Kwabena Ofei, Bent Mikklesten and Haris Hondo for WP 8 protocol

Acknowledgements to Mark Roe and Paul Finglas for WP 9 protocol

Acknowledgements to Marcus Maringer for the protocol to map the “app” data flows
5. Appendices

5.1 RICHFIELDS data sources and scope (work in progress)

Scope centred upon dietary intake as a measurable outcome (RICHFIELDS unique aspect) and one that can link RIs, disciplines, research questions. Radiates out to include data, services and tools related to what influences dietary intake: those considered determinants (individual, social or environmental influences) and status or health effects of dietary intake can influence what one may eat.

<table>
<thead>
<tr>
<th>Data sources</th>
<th>“big data”</th>
<th>Business-generated</th>
<th>Research-generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determinants RIs (DEDIPAC</td>
<td>WP 8, Procurement behaviours</td>
<td>WP 10 food choice decision-making</td>
<td></td>
</tr>
<tr>
<td>determinants, CESSDA demographics etc.)</td>
<td>WP 5 + 10, Purchase behaviours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Why one eats what they eat</td>
<td></td>
<td>Lifestyle behaviours</td>
<td>WP 6 + 10, Preparation behaviours</td>
</tr>
<tr>
<td>WP 9 Intake RIs</td>
<td>WP 7-10 Dietary intake/ consumption measurement and assessment tools including food composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What one eats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Status and Health RIs (ENPADASI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nutri-genomics, clinical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interventions case study WP 9 etc.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect of what one eats</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig.9 Dietary intake considered the link between RICHFIELDS data sources and scope and those of other relevant food and health research infrastructures.
5.2 Desk research aid

Desk research will not be an exhaustive systematic review of literature and available information on data and data methodologies relevant to purchase, preparation and consumption behaviours. However, the search will provide a ‘snap shot’ of the types and sources of data and data collection methodologies available. WP 5-7 can tailor their desk research strategy so that it is appropriate to their research area. Suggestions for desk research sources, key word terms and filters (to prevent an unmanageable amount of or irrelevant hits) have been detailed below (table x). Further data collection sources can be obtained from using expertise within the RICHFIELDS consortium, manual searches of sourced desk research reference lists and iteratively via focus group or workshop findings. The working definitions (section x) can be used in the desk research search terms or to exclude literature findings. All search terms, search sources and dates of searches to be transparently recorded and included in the deliverable report. A template has been provided for recording both desk research, focus group and workshop findings - please adapt this as necessary to aid the data collection process (Appendix x).

Table 1. Desk research search criteria suggestions

<table>
<thead>
<tr>
<th>Desk research</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desk research sources</td>
<td>Academic literature databases e.g., PubMed, Web of Science, IT specific databases?</td>
</tr>
<tr>
<td></td>
<td>Internet search engines e.g., Google</td>
</tr>
<tr>
<td></td>
<td>Libraries</td>
</tr>
<tr>
<td></td>
<td>Specialist IT organisations e.g., Governmental organisations e.g., European Commission</td>
</tr>
<tr>
<td></td>
<td>Consumer organisations e.g., itunes</td>
</tr>
<tr>
<td></td>
<td>Other?</td>
</tr>
<tr>
<td>Key word terms</td>
<td>Purchase, preparation, consumption, intake, “food choice”, eat*, behav*, food, diet, nutrition …</td>
</tr>
<tr>
<td></td>
<td>Data, facts, figures, statistics, details, particulars, specifics, features, information, evidence, intelligence, material, background, input, statement, report, file, documentation, archive …</td>
</tr>
<tr>
<td></td>
<td>Application, programme, software, mobile, “smart phone” …</td>
</tr>
<tr>
<td></td>
<td>Other?</td>
</tr>
<tr>
<td>Filters</td>
<td>Five/ ten years from publication</td>
</tr>
<tr>
<td></td>
<td>First 10 pages or 100 hits of search</td>
</tr>
<tr>
<td></td>
<td>NOT: …</td>
</tr>
<tr>
<td></td>
<td>Other?</td>
</tr>
</tbody>
</table>
Section 2: Definition of the term purchasing

In the past, purchasing was mostly defined as the actual decision at the point-of-sale, say putting an item into a shopping cart. Today, most researchers define purchasing behaviour as a process which goes beyond the act of purchase at the product shelf. Rather, it includes different factors which can influence the consumer before, during or after a purchase decision (cf. Solomon et al. 2013, p. 6).

Levy, Weitz and Grewal (2014, p. 91) divide the purchasing process into 5 steps. It begins with the pre-purchase phase, which includes the recognition of a need (= motive), a more or less intensive information search determined by the current type of buying decision, and an evaluation of different options (cf. Howard and Sheth 1969, p. 25f). After the first phase, the purchase decision is made. Finally, the consumer evaluates his buying decision in the post-purchase phase (cf. Levy et al. 2014, p. 91). Below, we primarily focus on the pre-purchase, as well as the actual purchase phase, because the after-purchase phase is mainly covered by the study of consumption patterns in Task 3.4.

The different types of buying decisions affect the processes and duration of a purchase decision, because different decision styles require different levels of information search effort (cf. Levy et al. 2014, p. 102, Schiffman and Kanuk 2007, p. 526). Four basic types of buying decisions can be identified in the literature: impulsive decision-making, extensive decision-making, which takes up the most time and effort, limited and habitual decision-making, which needs the least time and minimal or no effort (cf. Levy et al. 2014, p. 102f, Schiffman and Kanuk 2007, p. 526f). Extensive decision making is most cognitively controlled and characterized by a lack of “cognitive anchors” (cf. Kroeb-Riel and Gröppel-Klein 2013, p. 470). When the decision patterns have not yet been established, consumers are more willing to apply the extensive problem solving strategy. Hereby, a lack of experience leads to an increase in information acquisition and processing. As a result, the duration of the decision-making increases. In summary, the major characteristics of extensive decision making are: a high information demand, a long duration of decision-making and a necessity to acquire evaluation criteria as well as to reduce purchase risks.

The limited decision-making strategy requires few cognitive activities (cf. Kroeb-Riel and Gröppel-Klein 2013, p. 471f). Consumers use limited problem solving “when they have had already some prior experience with the product or service and their risk is moderate” (Levy et al. 2014, p. 102). In other words, decisions of this type are made deliberately and based on personal knowledge and experience. Thereby, the acquisition of information is less important, whereas the information search is still necessary. Limited problem solving uses mainly internal information resources, which means that the consumer recalls the required information from memory (cf. Kroeb-Riel and Gröppel-Klein 2013, p. 472). For this reason, limited problem solving requires a moderate amount of time and effort (cf. Levy et al. 2014, p. 102).

In contrast, habitual decision-making simplifies the decision behaviour noticeably (cf. Kroeb-Riel and Gröppel-Klein 2013, p. 485). Consumers make decisions based on their experience, which are
hardly cognition-based, but more reactive and automatic. With constant repetitions, associations between specific behaviours (e.g. purchase of a specific product) and internal (e.g. hunger) or external stimuli (e.g. a brand name) in specific situations (e.g. in a supermarket) are formed. After learning this association, the perception of the context stimuli triggers the appropriate behaviour. In other words: “I’ll buy the same thing I bought last time from the store” (Levy et al. 2014, p. 103). In this way, the habit assists fast and low-risk purchases (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 486), which undoubtedly take place in the food segment.

Impulsive buying is a stimulus-driven decision behaviour, which is normally accompanied by emotions (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 490; Rook 1987, p. 191; Youn and Faber 2000). It is mostly unplanned and barely cognitively controlled. The consumer chooses the product without a second thought, simply because he likes it (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 491). Before consumers make a purchase decision, they notice their unsatisfied need. Need recognition occurs, for example, when a product is no longer available in one’s household, when somebody is no longer satisfied with a product or when new needs arise. This need recognition triggers the purchasing process (cf. Levy et al. 2014, p. 92). Then, consumers continue with searching for “information about retailers, channels or products to help them satisfy that need” (Levy et al. 2014, p. 93). Finally, the pre-purchase phase ends with the evaluation of alternatives (cf. Levy et al. 2014, p. 91). During decision making, the consumer can be affected by different internal influences such as lifestyle, culture, social class and emotions, as well as the level of involvement or his attitudes (cf. Jansson-Boyd 2010, p. 131). One must not forget the external influence of retailers, of the family and other reference groups (cf. Levy et al. 2014, p. 91, 104f). These examples are just some of many influencing aspects. It is important to add that the intensive pre-purchase phase primarily characterizes extensive decision making. For other types of buying decisions, this phase is substantially shorter, so that the duration of a decision decreases. But before a purchase decision can be made, the products have to attract the consumer’s attention. The attention, caused by a stimulus, raises the individual awareness and processing of this stimulus, whereas the awareness and processing of another stimuli decreases (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 366f). Attention guarantees conscious awareness and efficient processing (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 367). Thus, it is an important element at the beginning of the purchase process.

Besides attention, former attitudes of consumers have a great influence on purchase behaviour. Attitudes, which are a main element in the purchase decision process, are defined as “a psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour” (cf. Eagly and Chaiken 2007, p. 598). Attitudes consist of affective and cognitive elements, which influence the purchase decision processes and the actual buying behaviour (cf. Kroeber-Riel and Gröppel-Klein 2013, p. 242). Especially the information search and evaluation of alternatives depend on consumer involvement and attitudes. After the purchase, factors like satisfaction play an important role for future rebuying intentions (cf. Levy et al. 2014, p. 101). The purchase decision has to be understood as a complex process, which contains different stages and psychological influencing factors. In conclusion, if this process should be measured, it cannot solely contain the purchase decision or intention, but also other related constructs, like attention, attitudes or involvement. Furthermore, purchase decision types have to be taken into consideration.
2.1 Types of consumption and co-variables

2.1.1 What is consumption?

The Oxford English dictionary defines consumption as: “consume (verb) eat, drink or ingest”. There are three main areas of the food supply chain that provide opportunities to assess dietary consumption at either an individual or population level:

- Total population food available (domestic food production plus imports and minus exports) to provide estimates of population or average individual food/energy/nutrient intake per person.
- Household food available (food purchases, larder stocks, garden/allotment produce, gifts minus wastage) to provide estimates of household or average individual food/energy/nutrient intake.
- Individual food consumption (prospective or retrospective) to provide estimates of individual or average population food/energy/nutrient intake (Geissler and Powers, 2011; World Health Organization, 1996)

Consumption at an individual level is a complex process, which may be influenced by a number of physiological and psychological factors. It is important at the outset to decide which factors/indicators are to be included when evaluating methods that measure consumption.

2.1.1 Pre-consumption indicators

Methods measuring intention or expectation to consume have been included in the review of methods. Although intention and expectation do not measure consumption behaviour, they have been considered to be related to consumption and could act as possible outcome variables to indicate future consumption behaviour. Methods that measure purchase behaviour could also be indicators of consumption i.e. dietary assessment can be conducted from household purchase data to provide approximate nutrient intake. These methods will be referred to in the inventory but are dealt with more fully elsewhere (Deliverable, D3.3).

2.1.2 Consumption behaviour, such as observed and self-reported consumption

Methods that measure observed (e.g., weight of food before and after consumption) and self-reported consumption (e.g., dietary record) are included in the review.

2.1.3 Post-consumption indicators

Methods that indicate consumption via nutrition intake or status of bio-chemical markers (e.g., doubly-labelled water) or anthropometric measures (e.g., height/weight) have been included within this review. These methods do not measure consumption behaviour but can be indicative of nutrient intake and status of food energy/energy balance, fat (e.g., cholesterol), protein, minerals and vitamins. Hence, a bio-chemical marker or anthropometric measure could act as an outcome variable to indicate consumption via changes in status.
2.1.4 Additional factors associated with consumption

There are a number of other factors which may influence consumption e.g. sensory properties of food, physiological state (hunger, satiety, health status) but methods that measure these factors have been excluded from the current review. Although these factors are recognised as being related to consumption or providing a degree of consumption information, they were not deemed sufficient to act as a proxy measure for consumption. In addition, measures of understanding were excluded from the review. Factors included and excluded are listed in Table 2.1.

Table 2.1 Consumption-related factors included/excluded in evaluation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Example</th>
<th>Included</th>
<th>Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-consumption indicator</td>
<td>Intention or expectation to consume and purchase data</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Consumption behaviour</td>
<td>Intake</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Post-consumption indicator</td>
<td>Biomarker of nutrient intake or status/anthropometric measures</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Quantity of consumption</td>
<td>Amount</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Quality of consumption</td>
<td>Dietary assessment</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Consumer unit of interest</td>
<td>Individual consumption</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Household consumption</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Population consumption</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Consumption unit of interest</td>
<td>Single nutrient/substance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Single food/beverage</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Single food/beverage group</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combination of &gt;1 nutrient/substance or individual foods/beverages or food/beverage groups</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall diet including dietary habits/patterns/dietary diversity</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Self-reported</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Reference period (minutes, hours, days, weeks etc.)</td>
<td>Retrospective</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>On-going/prospective</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Sensory properties of consumption</td>
<td>Liking, taste, smell, flavour, texture</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Type of diet</td>
<td>Vegetarian, gluten-free</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Disordered eating</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Interest in eating</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Relationship between dietary intake and health outcomes</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Dietary restraint</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Physiological state</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Liking</td>
<td>Hunger/satiety/physical activity</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Preference</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Familiarity</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Consumption setting</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Source of food procurement</td>
<td>Eaten inside/outside the home</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Pre-made/home-made</td>
<td>Purchased, own production/hunting, borrowed/bartered/exchanged/gift from friends or family, food aid</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Cooking/preparation method</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
5.4 Decision tree aid for in-/exclusion criteria to categorise tools collecting “big data” on purchase, preparation and consumption behaviours in phase 1 (Naomi Klepacz, WP 6)
5.5 Interview, focus group and workshop aid

Interviews, focus groups and workshops can be conducted concurrently with desk research and form part of case studies to allow iterative data collection. Specific protocols to be designed by WP leaders and co-ordinated by phase leaders. Things to consider:

- Purpose of interviews, focus groups and workshops
- Number of interviews, focus groups and workshops
- Ethics applications prior to recruitment and data collection
- Recruitment
- Procedures
- Interview, focus groups and workshop schedules
- Recording of data (audio, video, field notes)
- Data analysis
- Data interpretation
- Data presentation
- Language

Example questions related to user needs:

- Who do you believe are the future users of this technology or the RICHFIELDS research infrastructure?
- What experience do you have of using other research infrastructures or similar technology?
- What are your current data activity needs?
- What data systems do you currently access?
- How prepared would you be to use this technology/RICHFIELDS

Table 2 focus group procedure - tips

<table>
<thead>
<tr>
<th>Focus group</th>
<th>Protocol details</th>
</tr>
</thead>
</table>
| Before the focus group | Recruitment  
Arrange date, location and time  
Focus group details/directions  
Organise equipment required (tape recorder and spare batteries), room, pens and paper, flip charts and pens, tables, seating, refreshments etc.  
Check through focus group questions and prompts |
| Conducting the focus group | Outline of study, introductions, aim of focus group, confidentiality/right to withdraw etc.  
Turn on recording equipment (ensure permission granted for any recordings)  
Follow focus group schedule – try to keep on point as much as possible and return conversation to key issues that are to be discussed  
Take notes where appropriate |
| After the focus group | Immediately jot down interesting points and notes  
Listen to recordings  
Use template provided for desk research and focus group findings to be recorded – adapt as necessary  
Write e-mail or letter of thanks to participant and confirm any feedback or follow-up contact (re-contact by phase 3 partners or workshop participations?) if that was agreed |
| Other? | Other? |
5.6 Interview guide for WP8

Introduction
Welcome and thanks for agreeing to be an interviewee.

Purpose of this interview
We are conducting this interview as part of a large Richfield Project. The WP 8 aims at investigating best practices cases of extracting purchasing and procurement behaviour from existing data sources and using new technologies and devices of the farm to fork supply chain: retail and food service and to be utilized by the RI Consumer Data Platform.

The purpose of this interview is to elicit stakeholders view on

   (1) what constitutes best practices for businesses to collect food purchasing/procurement data,
   (2) how they can these data be used
   (3) how these data are structured.
   (4) how IC technologies can be used to capture food purchasing/procurement data.
   (5) if such data can be of value to research community
   (6) in that case what the privacy issues/IPR and/or ethical constraints might be

Interview questions

Section A: Best practices of capturing food purchasing/procurement data
For what purpose does your organization collect food purchasing/procurement data?

How are these data structured?

Would you explain why is it structured as it is now?

What do you consider to be the content of the data captured?
   - Would you elaborate on the relevance of the data content?

What advantages do the food purchasing/procurement data collections have for the industry?
   - Can you give some examples of how the data have been used?

What would you consider to be the significant challenge(s) of capturing the food purchasing/procurement data?
   - How do you think such challenge(s) can be overcome?

What is your opinion about using the food purchasing/procurement data to address societal challenge, “eg. Prevention of diet-related diseases”?

Could you think of cases where your data could be shared with the research community eg Richfield project?
- What benefits do you think it could bring to the stakeholders providing the data?

What would you consider as privacy restrictions (e.g., ownership issues) when using the food purchasing/procurement data for research purposes?

- Can you elaborate on any measures your organization have in place that protect the privacy of the captured data?

Could you suggest “ethical” topics to be discussed in Richfield project?

- Why do you consider it to be important?

Section B: IC technologies used to capture data

What key IC technology does your organization use to extract information on food purchasing/procurement data?

- Could you give a brief account of the software/hardware dimensions of the technology?

- Besides this software, does your company use other software packages?

- Why do you use it?

- Why it is not considered to be the key IC technology?

What criteria are used to select the tool or technology currently in use?

How do you evaluate the effectiveness of your IC technology or tools used for data collection?

What role do social media play in the gathering of purchase/procurement data?

- Can you think of examples where social media has been used successfully?

What do you consider to be the significant challenge so the capturing of the data using the kind of tools or technology available to you?

How do you think the use of IC technology can be developed further in the future to overcome the challenges?
5.7 Further information (useful links and resources)

Aligned research infrastructure, research projects and events

- JPI-HDHL knowledge hub ENPADASI
  http://www.enpadasi.eu/

- JPI-HDHL knowledge hub DEDIPAC

- ELIXIR
  - Pilot actions https://www.elixir-europe.org/about/elixir-pilot-actions-2015
  - Practical aspects of data curation and metadata annotation of experimental datasets for submission and sharing https://www.ebi.ac.uk/biosamples/

- Metadata
  REFERENCE
  CREATOR
  LANGUAGE
  CLASSIFICATION
  KEYWORDS
  DESCRIPTION
  DATE

- CESSDA
  - Data catalogue http://cessda.net/eng/Data-Catalogue
  - Big data Europe http://www.big-data-europe.eu/

- DASISH - Data Service Infrastructure for the Social Sciences and Humanities

- STATegra and the COST Action SeqAhead (big data).

  “…requirement for developing scalable infrastructures able to manage these quantities of data while making it available for efficient access and indexing…The main summary of BIGdata analysis is that even minor changes or low-level associations may be uncovered by the use of (very) large numbers of data points; therefore it remains to be seen how big data concepts will further reshape data integration in the life sciences.”


Appendices

- **Open science**
  “OpenScienceLink is an EU-funded project which will introduce and pilot a holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information. OpenScienceLink will pilot a range of novel services that could alleviate the lack of structured data journals and associated data models, the weaknesses of the review process, the poor linking of scientific information, as well as the limitations of current research evaluation metrics and indicators. Five pilot services will be integrated and piloted in particular:
  (a) Data journals development based on semantically-enabled research dynamics detection,
  (b) a novel open, semantically-assisted peer review process,
  (c) A services for detection and analysis of research trends,
  (d) services for dynamic researchers’ collaboration based on non-declared, semantically-inferred relationships, and,
  (e) a set of scientific field-aware, productivity- and impact-oriented enhanced research evaluation services.” [http://opensciencelink.eu/](http://opensciencelink.eu/)

- **DAPHNE**
  “There is certain consensus on the bibliography that for complex systems like Daphne and assuming that there are enough resources, the Architecture descriptions should not be done in a monolithic way, with a single mixture of formal and informal methods, just as it has been done traditionally. Instead it is far more adequate to use a distributed approach, with different tools describing each of the heterogeneous aspects of the system and the problem it is trying to solve. To enforce this, the ISO 42010 provides topics like views and viewpoints thus giving the Architect the possibility of looking at the system from different perspectives and so enriching the overall description of it.” [http://www.daphne-fp7.eu/sites/default/files/D2.6%20Daphne%20System%20Architecture%20first%20design_0.pdf](http://www.daphne-fp7.eu/sites/default/files/D2.6%20Daphne%20System%20Architecture%20first%20design_0.pdf)

**European forums and inter-/national initiatives**

- **The European e-Infrastructure Forum (EFF)**
  “…is a forum for the discussion of principles and practices to create synergies for distributed Infrastructures. The initial membership included GÉANT, TERENA (research networking), EGEE, EGI (grid computing), DEISA and PRACE (high-performance computing).” [http://e-irg.eu/documents/10920/238805/e-irg-blue_paper_on_data_management_v_final.pdf](http://e-irg.eu/documents/10920/238805/e-irg-blue_paper_on_data_management_v_final.pdf)

- **ePIC The European Persistent Identifier Consortium**
  “The European Persistent Identifier Consortium provides a Service for the European research community. ePIC was founded in 2009 by a consortium of European partners in order to provide PID services for the European Research Community, based on the handle system (TM, [http://www.handle.net/](http://www.handle.net/)), for the allocation and resolution of persistent identifiers. The consortium signed a Memorandum of Understanding aiming to provide long term reliability for the PID services” [http://www.pidconsortium.eu/](http://www.pidconsortium.eu/)

- **EC RICH**
  “To support the preparation of proposals, RICH, the network of National Contact Points for Research Infrastructures, in cooperation with the European Commission’s Research and Innovation Directorate-General, organized an Information Day on Wednesday 28 October 2015 in Brussels (The Cinquantenaire Museum, Parc du Cinquantenaire 10). The aim of the day was to inform potential

- EU regulation No 1291/2013 - public and private partnership definition
  - “public-private partnership means a partnership where private sector partners, the Union and, where appropriate, other partners, such as public sector bodies, commit to jointly support the development and implementation of a research and innovation programme or activities”
  - “Research infrastructures mean facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. Where relevant, they may be used beyond research, for example for education or public services. They include major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or scientific data; e-infrastructures such as data and computing systems and communication networks; and any other infrastructure of a unique nature essential to achieving excellence in research and innovation. Such infrastructures may be 'single-sited', ‘virtual' or 'distributed' REGULATION (EU) No 1291/2013 [http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/fp/h2020-eu-establact_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/fp/h2020-eu-establact_en.pdf)

- OECD (quality and data management)
  - Openness, Flexibility, Transparency, Legal conformity, Intellectual property protection, Formal responsibility, Professionalism, Interoperability, Quality, Security, Efficiency, Accountability, Sustainability
  - “Access arrangements are defined as the regulatory, policy and procedural framework established by research institutions, research funding agencies and other partners involved, to determine the conditions of access to and use of research data.”
  - Quality:
    “Although all areas of research can benefit from improved data quality, some require much more stringent standards than others. For this reason alone, universal data quality standards are not practical.”
    “Data access arrangements should describe good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding quality and authenticity.”
    “The origin of sources should be documented and specified in a verifiable way. Such documentation should be readily available to all who intend to use the data and incorporated into the metadata accompanying the data sets. Developing such metadata is important for enabling scientists to understand the exact implications of the data sets.”
    “Whenever possible, access to data sets should be linked with access to the original research materials, and copied data sets should be linked with originals, as this facilitates validation of the data and identification of errors within data sets.”
    “Research institutions and professional associations should develop appropriate practices with respect to the citations of data and the recording of citations in indexes, as these are important indicators of data quality.”

- UK data archive - quality [http://www.data-archive.ac.uk/create-manage/format/quality](http://www.data-archive.ac.uk/create-manage/format/quality)

- DG Connect DCAT (catalogue and definition of data set)
  - “The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing public sector datasets in Europe. Its basic use case is to enable a cross-data portal search for data sets and make public sector data better
searchable across borders and sectors. This can be achieved by the exchange of descriptions of data sets among data portals. The specification of the DCAT-AP was a joint initiative of DG CONNECT, the EU Publications Office and the ISA Programme. The specification was elaborated by a multi-disciplinary Working Group with representatives from 16 European Member States, some European Institutions and the US.”

- Dataset in DCAT is defined as a "collection of data, published or curated by a single agent, and available for access or download in one or more formats”. A dataset does not have to be available as a downloadable file. https://www.w3.org/TR/vocab-dcat/#vocabulary-overview”

https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11


“Researchers, public and private research organisations, universities and funding organisations must observe and promote the principles of integrity in scientific and scholarly research. These principles include: honesty in communication; reliability in performing research; objectivity; impartiality and independence; openness and accessibility; duty of care; fairness in providing references and giving credit; and responsibility for the scientists and researchers of the future.”

- ALLEA - data practices

“Data practices, including data management and storage, placing data at the disposal of colleagues who want to replicate the findings, adequate preservation of original data.

- Unicode - international software standards

http://www.unicode.org/

“The Unicode Consortium is a non-profit corporation devoted to developing, maintaining, and promoting software internationalization standards and data, particularly the Unicode Standard, which specifies the representation of text in all modern software products and standards. The Unicode Consortium actively develops standards in the area of internationalization including defining the behavior and relationships between Unicode characters. The Consortium works closely with W3C and ISO—in particular with ISO/IEC/JTC 1/SC2/WG2, which is responsible for maintaining ISO/IEC 10646, the International Standard synchronized with the Unicode Standard. The publications of the Unicode Consortium include The Unicode Standard, with its Annexes and Character Database, Unicode Technical Standards and Reports, Unicode Technical Notes and the Unicode Locales project, the Common Locale Data Repository.

Members of the Consortium include major computer corporations, software producers, database vendors, government ministries, research institutions, international agencies, various user groups, and interested individuals. A white paper outlining the overall value of a Unicode membership to an organization is available separately.”

- European Commission charter access to research infrastructure documentation

http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=charter_access_ri

- EC - big data definition
“A phenomenon resulting from 3 factors, commonly referred to as the '3 Vs' (Volume, Variety and Velocity), describing the fact that traditional data-handling and analysing technologies are increasingly challenged by the growth in:
- volume of data collected and stored,
- the variety of these data in terms of structure and formats (Excel spreadsheet data vs videos uploaded to Youtube)
- velocity (more and more data being generated due to Web 2.0 possibilities).


- EC - cloud
  - Apr 2016 European open science cloud https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

- ESFRI - research infrastructure definition
  “Research Infrastructures are facilities, resources or services of a unique nature that have been identified by European research communities to conduct top-level activities in all fields. This definition of Research Infrastructures, including the associated human resources, covers major equipment or sets of instruments, in addition to knowledge-containing resources such as collections, archives and data banks. Research Infrastructures may be “single-sited”, distributed”, or “virtual” (the service being provided electronically). They often require structured information systems related to data management, enabling information and communication. These include technology based infrastructures such as Grid, computing, software and middleware.” (ESFRI 2011)

- 2012 e-IRG (data management, CERIF, cloud, grid, intellectual property definitions)
  - CERIF (The Common European Research Information Format)
  - “Cloud computing (or simply ‘Cloud’) is an on-demand service offering a large pool of easily usable and accessible virtualised resources (such as hardware, development platforms and/or services) in a pay-per-use model. Clouds are usually offered commercially and currently use proprietary interfaces”
  - “Grid is a system that federates, shares and coordinates distributed resources from different organisations that are not subject to centralised control, using open, general-purpose and in some cases standard protocols and interfaces to deliver non-trivial qualities of service. Grid computing is used by VOs.”
  - “Intellectual Property Rights refer to the controlled right of use of created items, so that the creator benefits from that use. Intellectual property is broken down into several types, each of which apply to different created items; copyright, designs, patents, trademarks, protection from passing off and protection of confidential information” http://e-irg.eu/documents/10920/238805/e-irg-blue_paper_on_data_management_v_final.pdf

- Euraxess research in motion
  “EURAXESS - Researchers in Motion is a unique pan-European initiative providing access to a complete range of information and support services to researchers wishing to pursue their research careers in Europe or stay connected to it.” http://ec.europa.eu/euraxess/

- H2020 data management
**FAIR (principles and data object and metadata definitions)**

- Principles “Level 1: Each Data Object has a persistent identifier and intrinsic FAIR metadata (in essence 'static')
- Level 2: Each Data Object has 'user defined' (and updated) metadata to give rich provenance in FAIR format of the data, what happened to it, what it has been used for, can be used for etc., which could also be seen as rich FAIR annotations
- Level 3. The Data Elements themselves in the Data Objects are 'technically' also FAIR, but not fully Open Access and not Reusable without restrictions (for instance Patient data or Proprietary data).
- Level 4: The metadata as well as the data elements themselves are fully FAIR and completely public, under well-defined license. (Non-licensed data considered 'public' by their owner will still be excluded from integration projects by for instance Pharmaceutical companies).”

- “Data object: An Identifiable Data Item with Data elements + Metadata + an Identifier. We propose the term 'Data Object' to refer to the combination of data elements + their metadata + a unique identifier. These objects are arbitrarily complex and may appear in many forms and syntaxes.”
- “The metadata of a Data Object should be sufficiently rich that a machine or a human user, upon discovery, can make an informed choice about whether or not it is appropriate to use that Data Object in the context of their analysis. Metadata contained within the Data Object should inform the consumer about the license of the data elements; this metadata should be machine-readable to facilitate automated data harvesting while maintaining proper attribution. The Metadata contained within the Data Object should inform about any access-control policy, such that consumers can determine which components of the data they are allowed to access. The Metadata within the Data Object should inform about the authentication protocol leading to access, if applicable.”

**Other resources**

- Taxo tips - resources to help with taxonomies and controlled vocabularies

- Search tools listings (updated until 2012) - Tools for Taxonomies, Browsable Directories, and Classifying Documents into Categories

- GS1 - data quality

- Data Quality and Data Quality Assurance Policy, Oxford University - quality
  “In March 2007, the Audit Commission published a framework to support improvement in data quality in the public sector. This framework includes six key characteristics of good quality data, to which the University has added one further characteristic, and which may be summarised as follows:

  **ACCURACY**
  - Data should provide a clear representation of the activity/interaction
  - Data should be in sufficient detail
  - Data should be captured once only as close to the point of activity as possible

  **VALIDITY**
  - Data should be recorded and used in accordance with agreed requirements, rules and definitions to ensure integrity and consistency

  **RELIABILITY**
- Data collection processes must be clearly defined and stable to ensure consistency over time, so that data accurately and reliably reflects any changes in performance

**TIMELINESS**
- Data should be collected and recorded as quickly as possible after the event or activity
- Data should remain available for the intended use within a reasonable or agreed time period

**RELEVANCE**
- Data should be relevant for the purposes for which it is used
- Data requirements should be clearly specified and regularly reviewed to reflect any change in needs
- The amount of data collected should be proportionate to the value gained from it

**COMPLETENESS**
- Data should be complete
- Data should not contain redundant records

**COMPLIANCE**
- Data must comply with regulations on data protection and data security

http://www.admin.ox.ac.uk/pras/aboutus/data_quality/#d.en.41674

- IBM - Analytics Technology Platform
  “Turn faster insight into faster action with unmatched analytics capabilities on a foundation of data you can trust. Adapt easily to today's hybrid data--structured and unstructured, at rest and in motion, on premises, on cloud and in mixed environments. Energize your business with rapid innovation enabled by an open platform built on Spark.”
  http://www.ibm.com/analytics/us/en/technology/?S_TACT=M161089W&iiio=panalytics&cmp=m1610&ct=m161089w&cr=newsroom&cm=h&csot=-&ccy=-&cpb=-&cd=-

- Open stack object storage (SWIFT)
  “The OpenStack Object Store project, known as Swift, offers cloud storage software so that you can store and retrieve lots of data with a simple API. It's built for scale and optimized for durability, availability, and concurrency across the entire data set. Swift is ideal for storing unstructured data that can grow without bound.”
  https://wiki.openstack.org/wiki/Swift#OpenStack_Object_Storage_.28.22Swift.22.29

- Elastic stack
  “The Elastic Stack — that's Elasticsearch, Logstash, Kibana, and Beats — are open source projects that help you take data from any source, any format and search, analyze, and visualize it in real time. Products like Shield (security), Watcher (alerting), and Marvel (monitoring) extend what's possible with the Stack. And you can deploy it all as a service or on premise using Elastic Cloud.”
  https://www.elastic.co/products
WP 11 needs internal note

Project RICHFIELDS – Internal note
Barbara Koroušic-Seljak – February 25th 2016

RICHFIELDS ontology
In RICHFIELDS, data will be collected in order to
1. Reuse or mix data
2. Reuse or mix knowledge extracted from data
The extracted knowledge will be encoded so that the structure of information will be able to be shared among people or information systems (software agents). The outcome will be a knowledge base, i.e., an ontology and a set of individual instances of classes.

What is an ontology?
Ontology is a formal explicit description of
• Concepts in a domain: classes
• Subclasses, which represent concepts more specific than their superclasses
• Properties of each concept describing features and attributes of the concept: slots, roles or properties
• Restrictions on slots: facets or role restrictions
Example:
1. Classes Bread (subclasses Wheat, Mixed, Gluten-free) Bakery
2. Two Slots of Bread: Producer Taste
3. Instance of Bread: Krjavelj
4. Slot Producer Bakery Grosuplje
5. Slot Taste sour
We say that the bread Krjavelj is produced by Bakery Grosuplje and has got a sour taste; remark that the producer is a bakery (that is why the class bakery was introduced)

Methodological steps in ontology development
I. Determine the domain and scope of the ontology
II. Consider reusing existing, core ontologies (e.g., QualiFY/Quisper, ENPADASI etc.)
III. Enumerate important terms in the ontology (create a taxonomy – we could reuse QuaLiFY/Quisper taxonomy)
IV. Define the classes and the class hierarchy
V. Define the properties of classes - slots
VI. Define the facets of the slots
VII. Create instances
To determine the domain and scope, we will need to find answers to the following questions:
1. What is(are) the RICHFIELDS domain(s)?
2. What kind of knowledge do we want to extract from RICHFIELDS data? Can we predict which kind of decision making will be taken upon the collected and linked RICHFIELDS data?
   • Knowledge about eating habits / patterns
   • Knowledge about consumption habits / patterns
   • Knowledge about purchasing habits / patterns
   • Other?
3. Can we define for which aims this knowledge will be used? Who will be the users/stakeholders?
   • Research aims:
     - Development of new health policies
     - Development of new methods, tools for?
     - Other?
   • Business aims:
Classification of RICHFIELDS data
In RICHFIELDS, we will need to bring together heterogeneous data, interacting with multiple information sources. Those data will be collected from different data sources, and will have different information content as well different data formats and gathering modalities. There exist many data classifications. From the WP11 perspective, an important classification is based on the type of data sources. Data sources define data type, data alignment, technology for knowledge extraction, etc.

Data imported from relational databases, worksheets, pdf, excel, word documents
- How to align different data classifications/indexing?
- Are we going to deal with this kind of data in RICHFIELDS?

Data extracted from existing content/sites
By using HTML scraping and natural language processing techniques

(Linked) Open data
- While Open data refers to data freely available without restrictions, Linked data is refereeing to machine-readable data and semantically linked. Data can be open but not linked or linked but not open, however if data is open and linked it then becomes Linked open data.
- http://lod-cloud.net (ontologies need to be aligned)

Big data
Big data may be classified as
- Human-sourced data (social networks, broadcasting, web blogs etc.)
- Machine-generated data (Internet of Things, e.g. data from sensors, records from electronic devices etc.)
- Process-mediated data (e.g. health records, business transactions etc.).
Big data may be described by different sets of parameters:
1. Volume, variety, velocity, veracity (noise, abnormality in data)
2. Cardinality, continuity, and complexity
In any case, big data concern large-volume, growing datasets that are complex and have multiple sources. Considering all these types of big data, methodology for knowledge extraction (data analysis – statistical methods, mathematical models, machine learning approaches, data visualization) will need to be defined. It is important to know in advance who will be the data stakeholders as data analysis/visualization needs to be adapted to the needs of the users.
In WP 12 different alternatives for the RICHFIELDS infrastructure’s Business model Concepts will be proposed in terms of Value proposition (service offered), Supply chain configuration (means to deliver services to users) and Revenue model (remuneration mechanism for the platform).

The picture below represents the information inputs required by the four WP12 tasks to the other project WPs:

- Type of data collected and available through the platform, in order to define kinds of services to the users
- Data sources and envisioned data collection and delivery mechanisms (apps, etc…), in order to design suitable incentives for data providers and data users, in order to enable the data exchange through the platform
- Users segments and needs specifications in order to shape offerings to targeted users:
  1. User, beneficiary, stakeholder specification: who would be using, benefiting from, having an interest in this sort of platform,
  2. Needs characterisation – Elicitation of needs/ requirements of different users and stakeholders
  3. Decision-making model: tasks that each user/stakeholder group currently undertake and also decision-making models of these groups
WP 13 needs internal note

Project I – Internal note
Marc-Jeroen Bogaardt – February 2nd 2016

Needs of WP13 ‘Governance, Ethics and Final Design’ for methodological support from WP4

In this note I present a number of questions and issues that occur when reading closely the text in the DOA concerning the work of WP13. These questions and issues need to be answered in order to further elaborate and determine the work of WP13. The idea is that WP4, with its function of providing methodological support, could help with answering the questions.

**Task 13.1 is about data ownership, privacy and IPR.**

1) What exactly do we mean with ‘best practices’ in other Ris? How do we select best practices in other Ris? How many do we select? How do we know what is ‘best’? How are we going to judge that? And are there any reports describing those practices? Do we use the work done in EuroDISH? To what extent is the work done in EuroDISH suitable enough?

2) According to the text in the DOA we are must design ‘rules and policies for the RI on how to deal with data ownership and privacy issues.

   a. We should start by defining **data ownership** with regard to the three types of generated data.
   
   b. What kind of privacy issues are we dealing with?
   
   c. What is the difference between ‘rules’ and ‘policies’?
   
   d. And rules for whom?
   
   e. Enforcement: perhaps we should also consider the issue of how do we accomplish that those rules are actually enforced?

3) Because we are dealing with different kinds of owners of data, is there a difference in practice between the ‘ownership rights’ of an individual and the rights of an enterprise or public organisation?

4) How do we define ‘open data’? To what extent is our RI dealing with open data?

5) What definition of ‘big data’ do we use?

6) How do we define ‘open access’? What is the difference between open data and open access?

7) A policy concerning the access of the data for researchers. But what about a policy concerning the access of the data for others?

8) Definition of **intellectual property rights** with respect to the results of the researcher that used the data provided by the RI?

**Task 13.2 Ethics**

1) Which issues will be raised concerning big data of consumers? What do we mean with big data of consumers food related behaviour?

2) We have to design a code of ethics for big data analytics that does not alienate consumers by disregarding or overlooking their privacy concerns.

   a. How do we define ethics?
   
   b. We are not only designing rules and policies but apparently also a code. What is the relation between these three? How should we position them?
c. What is the purpose of the code of ethics? For whom is the code?

**Task 13.3 Governance structure**

1) Governance structure is not the same as governance. So first we must agree on the definition of governance of a RI. That should help to determine the elements of governance to be included in the design of the RI. And the governance is focused on the situation in which the RI is in full operation.

2) How do we define a business model of a RI? What is included and excluded?

3) **Public-private partnership**? We will design a RI that collaborate with both public as private organisations: universities and research institutes funded by the (national) government, and business such as international operating food retailers, and food and beverage companies, but also marketing research companies and app providers (firms who provide apps for consumers). What does this imply for the selection and examination of the best practises (see above)? Because based on the work and experience in EuroDISH there are not much existing Ris in Europe which are public-private collaboration.

**Task 13.4 Final design**

1) The main question is: what will be or should be the main elements of the final design? What should be described in the final design?

2) The text of the DOA mentions ‘data platform’ and ‘eco-system of app developers’. How do we see the position of the data platform with respect to the RI? In my opinion the data platform is just a (large and important!) part of the RI. In other words, the RI is more than just the data platform. But where exactly is the boundary?

3) What do we define as an eco-system of app developers? First of all I would prefer the term app providers. And we only app providers and no other actors?

**Task 13.5 Roadmap and recommendations**

1) We will draw a roadmap that should outline ‘the steps’ to be taken in order to actually build the RI and make it fully operation, based on the final design.

2) What do we mean with ‘recommendations’? To my opinion recommendations are actions to be performed by others. We refer to ESFRI and others. Recommendations to ESFRI, which is a strategic forum and not a funding organisation, will have another character than recommendations to e.g. the national governments.

3) The DOA state that the roadmap is focused on reaching the ERIC status of the RI. This means that the legal structure of the RI needs to become an ERIC. In that case, only member states, associated countries and intergovernmental organisations can be members of the RI. However, a State may decide to be represented by one (or more) public entity or private entity with a public-service mission. So we should be aware that the goal of becoming an ERIC has many implications particularly for the governance and business model of the RI.

I realize that a lot of the questions and issues mentioned above are the responsibility of the partners working in WP13. But I hope that the partners of WP4 could help clarifying some of the issues, for example the underlined terms.

Marc-Jeroen Bogaardt

Wageningen, February 2nd, 2016.