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

Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package WP11
Deliverable D11.3

Standardisation requirements for RI Consumer Data Platform

An overview of standards in relation to the RI Consumer Data Platform

Date delivered:
M30

Authors:
Barbara Koroušić Seljak, JSI; Krijn Poppe, WUR;
RISE, AAU, GS1 and AALTO

Deliverable lead beneficiaries:
JSI

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

Project	
Project acronym:	RICHFIELDS
Project full title:	Research Infrastructure on Consumer Health and Food Intake for E-science with Linked Data Sharing
Grant agreement no.:	654280
Project start date:	01.10.2015
Document:	
Title:	Standardisation requirements for RI Consumer Data Platform - An overview of standards in relation to the RI Consumer Data Platform
Deliverable No.:	D11.3
Authors:	B. Koroušić Seljak, IJS; SP, AAU, GS1 and AALTO
Reviewer:	Karin Zimmermann – Project Coordinator Pieter van 't Veer – Scientific Coordinator Fred van Alphen – Project Advisory Board member
Start date:	1.10.2015
Delivery date:	19.03.2018
Due date of deliverable:	31.03.2018
Dissemination level:	PU
Status:	Final

Change history:		
Version	Notes	Date
001	Draft version	23.01.2018
002	Comments from Krijn Poppe considered	28.02.2018
003	Comments from Fred and Pieter considered	19.03.2018

Karin Zimmermann
Project Coordinator

Prof Pieter van 't Veer
Scientific Coordinator

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

Summary

This deliverable takes an inventory, discusses and assesses primary standardisation requirements for the functional and technical design of the RI Consumer Data Platform. Standards for collecting scientific data, business data and consumer data as well as standards for data linkage and harmonization are aligned with the user and RIs requirements (D11.1) and the semantic data model (D11.2). Existing European and global data standards for product codes (e.g. in UN/CEFACT or GS1), electronic health records, global positioning, physical exercise data etc. were taken in account. Special focus was given to work programmes of big data standardization consortia, like W3C Building the Web of Data67, ITU-T SG13 Future networks including cloud computing, mobile and next-generation networks, ISO/IEC JTC 1/SC 32 Data management & interchange, Research Data Alliance, OASIS Organization for the Advancement of Structured Information Standards, Transaction Processing Performance Council, etc. All this information will be needed for the communication between various mobile applications and the RI Consumer Data Platform.

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

Table of Contents

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

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

Abbrevations

API	Application Programming Interface
CSV	Comma Separated Value
GNSS	Global Navigation Satellite Systems
HTTP	Hypertext Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
REST	Representational state transfer
RI	Research Infrastructure
RIMS	RICHFIELDS Inventory Management System
SDK	Software Development Kit
URL	Uniform Resource Locator
UTF	Unicode Transformation Format
XML	Extensible Markup Language

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

1 Introduction

The main aim of the RICHFIELDS project is to design an **Open Architecture Data Platform** to collect, harmonize and share consumer, business and research data in order to provide the scientific research community with innovative data sets and the ability to generate new knowledge in the consumer food and health domain.

In D4.4, it was proposed that the data platform addresses the **determinants of consumer behaviour** relevant to food and health across three distinct instances of behaviour: purchase, preparation and consumption. By building on the determinants and intake ('DI') components of the proposed DISHRI (www.eurodish.eu), the design proposal arising from the RICHFIELDS project will be an important building block for subsequently constructing an ESFRI roadmap proposal for a pan European FNH-RI.

In order to develop the Core Offering Proposal into a detailed specification it was also proposed that initial consideration by Phase 3 is given to answering the following 4 questions as a priority:

1. **What data can be readily incorporated into the data platform** at the Minimum Viable Product (MVP) level **from an availability**/ethical **perspective?**
2. Are these data of sufficient value to the proposed primary users; If not **how will the additional data required be obtained?**
3. Is there a sufficient value offering for data providers to ensure access to the data required?
4. Which stakeholders are essential to form the MVP/MVE (Minimum Viable Ecosystem) ensuring appropriate levels of Governance and User engagement?

In the following subsections, we will address the first two questions from the technical perspective. We will identify currently available food and nutrition data sources to find out which data formats are relevant, and explore standards used to specify data collection in order to add sufficient value. Other criteria like data quality and data content are also important to meet customer needs, however these as well as the third and the fourth question are out of the scope of this deliverable and will be discussed in WP12 deliverables.

1.1 Currently available data

In D5.1, D6.1 and D7.1 an inventory of types of purchase, preparation and consumption data and data collection methodologies for consumer-generated food purchase data was created. Altogether, fifty-four mobile applications were identified for inclusion into the RICHFIELDS Inventory Management System (RIMS), an online management system created in response to these tasks.

In Table 1, we present an overview of formats in which data collected by the apps from the RIMS are provided. It can be seen that a small percentage (15%) of the apps provides an easy access to structured data via web services or SDK, while the remaining apps require natural language processing and data normalisation, which is described in D11.2 in details. The aim of this overview is not to select apps/data to be considered by the RICHFIELDS MVP, but to identify which data formats are available today and needs to be considered from both MVP's perspectives: standards and data semantics.

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

Table 1. An overview of data formats.

MOBILE APP	API	SDK	DATA FILE	GOOGLE DRIVE	GOOGLE ACCOUNT	DROPBOX	AIRDROP	EMAIL REPORT
Curd Collective	http / json							
Fitbit	http / json		Csv					
Untappd - Discover Beer	rest							
Calorie	rest / json		Excel					
Calorie Counter & Diet Tracker by MyFitnessPal	rest / json	X						
Carb & Fat Counter - Virtuagym Food	rest / json		Excel					
Foodspotting	rest / json							
UP by Jawbone - Track with UP Move,Ñ¢	rest / json	X						
UP Coffee	rest / json	X	X					
UP24,Ñ¢	rest / json	X						
S Health	X	X						
21 Day Tummy Tracker								X
Activ8rlives Health Monitoring and Food Diary App			pdf, excel					
Allergy Journal								X
and Carb Counter			Pdf					
and EDNOS			Pdf					
Binge Eating			Pdf					
Blood Sugar Control			Pdf					
BMI			Csv					
BMR			Csv					
Body Tracker - Body Fat Calculator			Csv					
Bowelle - The IBS tracker								X
Bulimia			Pdf					
Bulk Up! Protein Tracker - high protein diet counter to gain muscle & build strength			Pdf					X
Calorie Counter								X
Calorie Counter by FatSecret			csv, pdf					
Calorie Counter by YAZIO ,Äì Diet Tracker and Food Diary for Weight Loss			X					
Calorie/KJ Counter and Food Diary by MyNetDiary - for Diet and Weight Loss			excel, pdf					
CalorieSmart Calorie Counter			Excel					pdf
Cals & Macros FREE			Csv					
CarbsControl			Excel					X
CARROT Hunger - Talking Calorie Counter						X		
Cronometer			Csv					
Daily Water - Water Reminder and Counter						X		
Daily Water - Water Reminder and Counter								X

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

MOBILE APP	API	SDK	DATA FILE	GOOGLE DRIVE	GOOGLE ACCOUNT	DROPBOX	AIRDROP	EMAIL REPORT
Database and Calculator			Pdf				pdf	
Diabetes App Lite - blood sugar control								X
Diabetes Diary Glucose Tracker			Csv					
Diabetes Pedometer with Glucose & Food Diary			csv, pdf					
Diabetes Tracker with Blood Glucose/Carb Log by MyNetDiary			excel, pdf					
Diet and Fitness Tracker			Excel					X
Diet Assistant - Weight loss			Csv					
Diet Diary			Csv					
Diet Watchers Diary			Html			excel		
Drinkcontrol			Excel					
Food and Symptoms Diary			Pdf					
Food Diary			csv, excel					
Food Jotter								X
Glucose Buddy: Diabetes Log			Csv					
glucose tracker and carb counter								X
Health-Tracker			Csv					
HI - Health & Fitness Tracker				X				
Hydro Coach - drink water					X			
iFood Diary			Excel					X
Intolerance Food Diary			csv, pdf					
Keto - Low Carb Diet Tracking								X
Low Carb Diet Assistant			Csv					
Macronutrients			Csv					
Mijn Eetmeter			X					
My Daily Plate			X					
My Diet Diary Calorie Counter App			csv, pdf					
My Paleo Tracker - primal & low carb diet counter								pdf
MyPlate Calorie Tracker			X					
mySugr Diabetes Diary			Pdf					
mySymptoms Food & Symptom Tracker			Pdf					
Noom Healthy Weight Loss Coach		X						
Nutrition Tracker			Excel					X
OneTouch Reveal			csv, excel					
PercentEat food diary			Pdf					
Pic Healthy			Pdf					
PKU Diet Management			Pdf					
ProTracker Plus Watchers Nutrition and Exercise Value Tracker			Pdf				X	
Restaurants Calorie Tracker Free								html

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

MOBILE APP	API	SDK	DATA FILE	GOOGLE DRIVE	GOOGLE ACCOUNT	DROPBOX	AIRDROP	EMAIL REPORT
Rise Up + Recover: An Eating Disorder Monitoring and Management Tool for Anorexia			Pdf					
Simple Calorie Count			SQLite					
Simple Diet Diary			SQLite					
SITU Scale (Bluetooth)		X	Excel					
SITU Smart Food Nutriton Scale			X					
SparkPeople			Excel					
Sugar Sense - Diabetes App			Pdf					
Tap & Track -Calorie Counter (Diets & Exercises)			Csv					
The Monash University Low FODMAP Diet								pdf
TracknShare			X			X		X
Ultimate Food Value Diary Plus - Diet & Weight Tracker			Csv					
Weight Loss			Csv					
Weight Loss Diet & Calorie Calculator			X					
Weight Loss Tracker+ Food Diary			Excel					
Wijn								X

1.2 How to obtain additional value to currently available data?

Exploring the inventory of types of purchase, preparation and consumption data and data collection methodologies, we concluded that

- data are of different types (structured, unstructured, open, big, static, dynamic - real-time…), and
- data collection methodologies rely not only on different approaches but also on different standards.

Therefore, in Task 11.3 we focused on standards required for the efficient development of methodologies for data collection, linkage and harmonization. Section 2 provides criteria used to select standards relevant for data formats to be included in the MVP and beyond. In Section 3, an inventory of standards for data collection, linkage and harmonization is presented. Finally, Section 4 concludes the deliverable and gives implications for the final design of the Open Architecture Data Platform.

2 Criteria for the selection of standards

We identified standards for data collection and data linkage and harmonization. The first group of standards include standards for collecting scientific data, business data and consumer data. We consider standards established by the European networks of excellence, industrial data standardization consortia and the global social media network. The second group include standards established by the European and global data standardization consortia.

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

3 An inventory of standards

3.1 Standards for data collection

Data provided by the information systems, identified in the inventory of types of purchase, preparation and consumption data, complies with standards that are described in the following subsections. As these standards have not been aligned, advanced computer methodologies for data harmonization and linkage are required. In D11.2, few such methodologies and a food and consumer behaviour ontology as part of the RICHFIELDS data semantics model are presented.

3.1.1 Scientific data

In RICHFIELDS we have addressed food and nutrition scientific data, which are described by different standards for collection, indexing and classification. Let us mention few of them, such as CEN/TC 387 Food data and data structure standard (BS EN 16104:2012), LanguaL (http://www.langual.org), FoodEx2 (https://www.efsa.europa.eu/en/data/data-standardisation), ELIXIR, CORBEL and BBMRI-ERIC. Some of food and nutrition scientific data that can be accessed through the Quisper Server Platform, developed in the FP7 project QuaLiFY, has already been enriched with the power of semantics provided by the Quisper ontology (EFTIMOV, T, KOROUŠIĆ-SELJAK, B. *QOL - Quisper Ontology Learning using personalized dietary services*, The Jožef Stefan Institute Technical Report No. 11985, 2015.). More details are provided in D11.2.

CEN/TC 387 Food data and data structure standard

The current CEN standard is based on two initiatives, the EC 6th Framework EuroFIR Network of Excellence (http://eurofir.org) and Food and Beverage Extension to the GS1 GDSN Trade Item standard (https://www.gs1.org/gdsn/current-standard). The main aim of the standard is to provide a framework that facilitates and enables generation, compilation, dissemination and interchange of food[1] data that are comparable and unambiguous with respect to the identity of foods, the description of foods and food property measures including their quality. The standard is structured to be robust and flexible enough to incorporate future extensions with respect to various types of data.

The term food generally refers to substances intended for human consumption, normally with exceptions for e.g. medicines, and includes raw or processed food products and substances used in the manufacture. The exact definition, however, may vary depending on legislation and cultural differences. This standard can be used regardless of such variations. It uses food properties as a general term when describing food constituents such as nutrients, heavy metals, micro-organisms, but also when describing various physico-chemical properties of foods. However, this standard does not include all definitions that are required. For example, the set of food properties that can be used, such as contents of various nutrients and heavy metals, is not included in the standard. These and all other so called controlled vocabularies have to be agreed on within the community. An annex of the standard provides examples of required controlled vocabularies.

[1] By food we mean both food and drinks.

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

The exchange of food data among different parties requires an agreement on not only what data to exchange but also on the encoding of the data. This standard includes data encoding rules based on XML.
Detailed information about this standard is provided in: Food data - Structure and interchange format (EN 16104:2012). Published 31/01/2013, maintained by AW/275. Available online: doi:10.3403/30217587 or http://www.freestd.us/soft4/1640837.htm (accessed on 3rd February 2018).

LanguaL

stands for “LanguaaLimentaria" or “language of food". It is an automated method for describing, capturing and retrieving food information. In LanguaL, each food is described by a set of standard descriptors (indexing terms) chosen from the following facets of the nutritional or/and hygienic quality of the food: A (Product Type), B (Food Source), C (Part of Plant or Animal), E (Physical State, Shape or Form), and F to Z (additional descriptors for indexing the product information). The LanguaL facet terms are fully structured in a hierarchy, which enables displaying its thesaurus in a logical way. Each term may have several narrower terms giving the concept a more specific meaning. The hierarchy also possesses poly-hierarchical relationships, meaning that a term may be related to several broader terms representing the concept in a wider meaning.
Each food can be allocated to several food group classifications (Facet A Product type –Figure 1). At the moment, LanguaL includes 13 classification systems, including Codex Alimentarius (http://www.codexalimentarius.org), European Food Groups, EuroFIR Food Group Classification, GS1 Global Product Classification (http://www.gs1.org) and USDA standard reference. Over 75,000 foods and food products have already been indexed in various countries using the LanguaL system. Further details can be found in the LanguaL thesaurus (Møller, A.; Ireland, J. (2013). LanguaL™ 2012 – The LanguaL™ Thesaurus. EuroFIR Nexus Technical Report D1.17a. Danish Food Information.).

Figure 1. Example of the LanguaL description and classification.

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

FoodEx2

stands for “Food classification and description for Exposure assessment", version 2 and was developed by the European Food Safety Agency (EFSA), which gathers food consumption data for individuals in Europe and uses them for exposure assessment. Food consumption data collected from EU member states are stored in the Comprehensive Food Consumption Database, using a common food classification and coding system FoodEx2.
The current FoodEx2 version consists of a large number of individual food items aggregated into food groups and broader food categories in a hierarchical structure of parent-child relationships (Figure 2). Central to the system is a core list of more than 1100 food groups or individual food items that represents the minimum level of detail needed when coding or identifying a food collected in any domain for intake or exposure assessments. More detailed terms may exist below the core list and these are identified as the extended list (approx. 1509 terms). Apart from bearing a unique alphanumerical code, all terms in FoodEx2 are flagged with attributes defining their role (hierarchy, core list or extended list) and their state (e.g., raw commodity, ingredient, simple or composite food).

Type	*Code*	*Food Group*	*flag*
H	A04AH	**Sheep meat food**	g
H	A04AJ	Sheep carcase	r
H	A04AK	Sheep fresh meat / fat tissue	r
C	A01RH	*Sheep fresh meat*	r
E	A01RJ	**Sheep (adult) fresh meat**	r
E	A01RK	**Lamb fresh meat**	r
E	A01SL	**Moufflon fresh meat**	r
E	A01VB	*Sheep, fresh fat tissue*	r
H	A04AL	Sheep, minced meat	d

Figure 2. Example of the FoodEx2 description and classification.

More information about FoodEx2:

- European Food Safety Authority. Guidance on the EU Menu methodology (online 3944). EFSA Journal 2014, 12, p. 12; doi:10.2903/j.efsa.2014.3944.)
- European Food Safety Authority (2015). The food classification and description system FoodEx 2 (revision2). EFSA supporting publication2015, 804, p. 90.

ELIXIR

is an intergovernmental organisation that brings together life science resources from across Europe (https://www.elixir-europe.org). These resources include databases, software tools, training materials, cloud storage and supercomputers. ELIXIR's activities are divided into five areas called platforms, which include Data platform, Tools platform, Interoperability platform, Compute platform and Training platform. Let us expose the Data platform and the Interoperability platform.

- The Data platform has developed a process to identify European data resources that are of fundamental importance to research in the life sciences and are committed to the long term preservation of data. These resources are called ELIXIR Core Data Resources. A list of active ELIXIR Core Data Resources is available at https://www.elixir-europe.org/platforms/data/core-data-resources. Key indicators,

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

which reflect the essence of the definition of an ELIXIR Core Data Resource, has been published in the paper (Durinx C, McEntyre J, Appel R et al. Identifying ELIXIR Core Data Resources [version 2; referees: 2 approved]. F1000Research 2017, 5(ELIXIR):2422 (doi: 10.12688/f1000research.9656.2));

- The Interoperability Platform, which is guided by the FAIR data principles, offers among other services also technical services which ensure that data is interoperable, that it can be accessed programmatically, and that it contains persistent identifiers. The technical services develop minimum information standards and vocabularies. An additional resource, maintained by ELIXIR UK in Oxford, is Biosharing (https://fairsharing.org/). This is a curated educational resource on inter-related data standards, databases, and data policies in life, environmental and biomedical sciences. The standards in FAIRsharing are manually curated from a variety of sources, including BioPortal (http://bioportal.bioontology.org/), MIBBI (https://fairsharing.org/collection/MIBBI) and the Equator Network (http://www.equator-network.org/about-us/).

CORBEL

is an initiative of biological and medical research infrastructures, which creates a platform for harmonised user access to biological and medical technologies, biological samples and data services required by cutting-edge biomedical research. CORBEL has established a Catalogue of Services, which is a tool to list the main services of European biological and medical sciences RIs that are working together in CORBEL towards offering shared services for life-science (http://www.corbel-project.eu/services.html).

BBMRI-ERIC

was set up to establish a pan-European distributed research infrastructure of biobanks and biomolecular resources in order to facilitate the access to resources as well as facilities and to support high quality biomolecular and medical research. One of the objectives of BBMRI-ERIC is to establish Standard Operating Procedures (SOPs) for all processes related to sample collection, processing, storage, retrieval and despatching. These SOPs will follow the procedures as specified in the WHO/IARC guidelines for biological resource centres for cancer research whenever feasible (http://ibb.iarc.fr/standards/index.php).

3.1.2 Business data

Standards for business data include: GS1, UN/CEFACT, electronic health records, global positioning etc. In D9.2 and D11.2 we present a methodology for enhancing GS1 data with semantics needed to link business data with scientific ones. In the case study presented there, the idea of Semantic Web, or attaching semantic metadata to documents, pointing to concepts in ontology was applied. Information can be exported as instances in the ontology, or text documents annotated with links to the ontology.

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

GS1

The GS1 Global Data Synchronisation Network® (GDSN®) is a network of interoperable data pools enabling collaborating users to securely synchronise master data based on GS1 standards (Figure 3). GDSN supports accurate, real-time data sharing and trade item updates among subscribed trading partners.

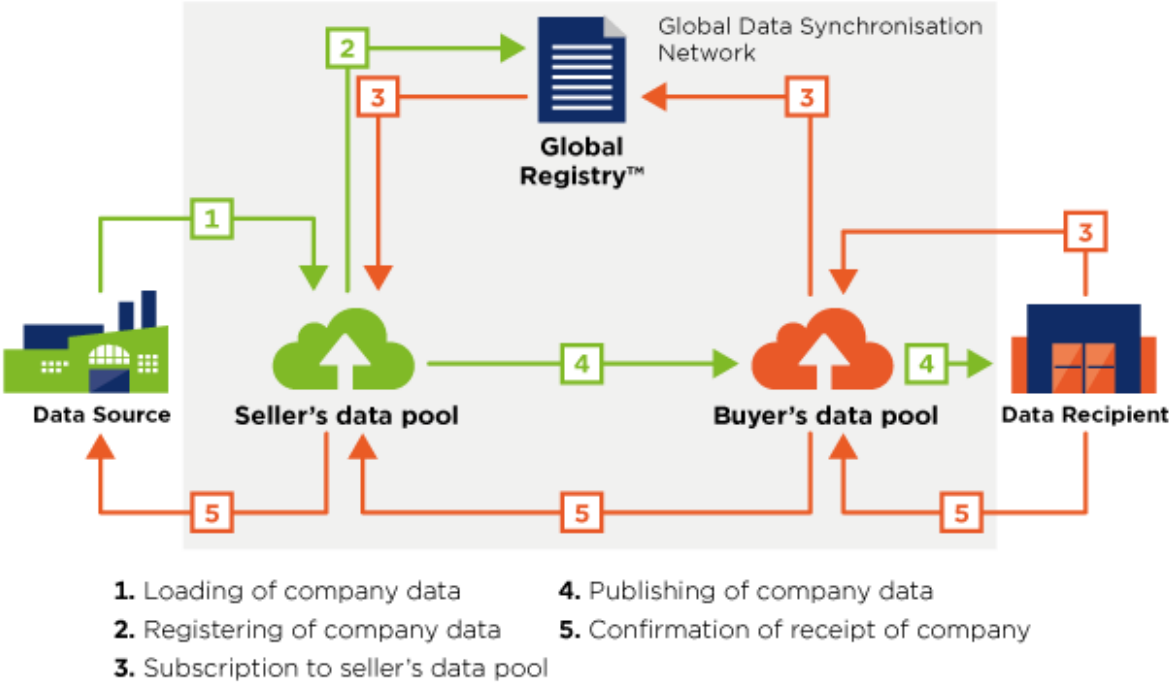

Figure 3. GS1 GDSN.

Currently available GDSN standards for nutrition and health are available at https://www.gs1.org/gdsn-standards. Let us expose few of them

- Product Image Specification (2017): that establishes rules for the storage of digital images associated to products and provides details on all aspects of digital imaging storage (https://www.gs1.org/sites/default/files/docs/gdsn/Product_Image_Specification.pdf);
- GS1 Fruit & Vegetable GTIN Assignment Implementation Guideline (2016): aimed for the fruit & vegetable industry, providing guidance on how to assign a GTIN (Global Trade Item Number) and when to assign a new GTIN (https://www.gs1.org/docs/freshfood/Fruit_and_Vegetable_GTIN%20Assignment_Guideline.pdf);
- Fruit & Vegetable Master Data Attribute Implementation Guide (2017): that provides support to companies seeking to electronically exchange fruit & vegetable product information in accordance with GS1 standards (https://www.gs1.org/docs/freshfood/Fruit_Vegetable__Master_Data_Attribute-ImpGuide.pdf);

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

- Fisheries & Aquaculture Master Data Attribute Implementation Guide (2016): which outlines which attributes should be used for fish items and recommends best practices for the use of these standards to exchange static fish data between suppliers and retailers.

UN/CEFACT

stands for the United Nations Centre for Trade Facilitation and Electronic Business, which is an intergovernmental body of the United Nations Economic Commission for Europe (UNECE). UN/CEFACT focusses on two main areas of activity to make international trade processes more efficient and streamlined:

- *Trade facilitation* involves the simplification of trade procedures (or the elimination of unnecessary procedures). This includes work to standardize and harmonize the core information used in trade documents, to ease the flow of information between parties by relying on appropriate information and communication technology, and to promote simplified payment systems to foster transparency, accountability and cost-effectiveness;
- *Electronic business* focuses on harmonizing, standardizing and automating the exchange of information that controls the flow of goods along the international supply chain.

UN/CEFACT has produced few tens of trade facilitation recommendations and a range of electronic business standards, which are used throughout the world by both governments and the private sector. They reflect best practices in trade procedures and data and documentary requirements. The International Organization for Standardization (ISO) has adopted many of them as international standards. Some of the more well known UN/CEFACT electronic business standards are (http://www.unece.org/cefact/recommendations/rec_index.html):

- Recommendation 1: United Nations Layout Key (UNLK) for Trade Documents. This provides an international basis for the standardized layout of documents used in international trade.
- Recommendation 16: UN/LOCODE Code for Trade and Transport Locations provides an alphabetic code for seaports, airports, inland freight terminals and other customs clearance sites.
- Recommendation 25 and the UN/EDIFACT Standard represent a set of internationally agreed standards, directories, and guidelines for the electronic interchange of structured data, between independent computerized information systems. UN/EDIFACT is the international standard for Electronic Data Interchange and is used throughout the commercial and administrative world.
- Recommendation 33 on Single Windows proposes that governments establish a Single Window facility that allows parties involved in trade and transport to lodge

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

standardized information and documents with a single entry point to fulfil all import, export and transit-related regulatory requirements.

A family of Supply Chain "Cross-Industry" messages are exchanged globally between trading partners covering the majority of business-to-business (B2B) electronic exchanges from order to payment. One of the key documents within this family is the Cross Industry Invoice (CII) which functions primarily as a request for payment, used as a key document for Value Added Tax (VAT) declaration and reclamation, for statistics declarations and to support export and import declarations in international trade.
The eDAPLOS message describes the data crop sheet exchanged between farmers and their partners. This message has allowed users to harmonize the definitions of technical data, develop consensual data dictionaries which can be used as a basis for all the steps of traceability and create a standardized Crop Data Sheet message. DAPLOS, which is based on the UN/CEFACT Core Component Library, has been adopted by 25 000 farmers and regional agriculture chambers in Europe.
CITES (the Convention on International Trade in Endangered Species of Wild Fauna and Flora) has developed a version of their declaration using the Core Component Library of UN/CEFACT and has generated an XML message according to the specifications of UN/CEFACT. CITES is an international agreement between governments. Its aim is to ensure that international trade in specimens of wild animals and plants does not threaten their survival. The CITES declarations are used in customs clearance procedures in all countries around the globe.
In 2017, 123 electronic data forms (XML schemas) were developed by UN/CEFACT (http://www.unece.org/cefact/xml_schemas/index).

Electronic health records (EHR)

The EU-funded project EHR4CR (http://www.ehr4cr.eu/) was aimed to develop a robust and scalable platform that can utilise de-identified data from hospital EHR systems, in full compliance with the ethical, regulatory and data protection policies and requirements of each participating country. In this project, robust and acceptable technical and procedural approaches that should be taken to ensure privacy protection and compliance with European and national/regional regulations on data protection were developed.
The European Institute for Innovation through Health Data (i~HD) (http://www.i-hd.eu) is a non-profit organisation, arising in part out of the EHR4CR project, to develop and promote best practices in the governance, quality, semantic interoperability and uses of health data, including its reuse for research. This organisation governs the operational platform InSite (http://www.insiteplatform.com), which enables trustworthy re-use of EHR data for research for industry, hospitals and academia.

Global positioning

At the moment, there are four global navigation satellite systems (GNSS) in operational or close to operational stage available:

- GPS (US GNSS - Global Positioning System, fully operational from 1995)
- GLONASS (Russian GNSS - Globalnaya navigatsionnaya sputnikovaya sistema, fully operational from 2011)
- Galileo (Eropean GNSS, full operation expected in 2020)

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

- BeiDou-2 (Chinese GNSS - formerly known as COMPASS, full operation expected in 2020)

During the development of these GNSSs and other regional satellite systems in the last 50 years several standards and protocols were adapted or intentionally developed to enable communication of the GNSS satellite receivers with other devices. Data protocol dictates what information can be conveyed from one party to another during communications. Incorrect choice of protocols may result in lack of information needed. It also affects the quality of the communications, i.e. whether it can be carried properly within the required timeframe. Most of GNSS receiver manufacturers have developed and maintained their own data formats. For example, Leica LB2, Trimble RT17 and RT27, Ashtech MBEN and PBEN, and Topcon TPS. These are mostly encoded binary formats and are efficient in terms of bandwidth, but often require the use of mobile receiver software/hardware supplied by the relevant manufacturer as the reference receiver.
Since individual GPS manufacturers have their own proprietary formats for storing GPS measurements, it can be difficult to combine data from different receivers. A similar problem is encountered when interfacing various devices, including the GPS system. To overcome these limitations, a number of research groups have developed standard formats for various user needs. Here we present the most widely used standard formats, namely, RINEX, NMEA 0183, NGS-SP3, GPX, KML, and RTCM SC-104.

RINEX

Receiver Independent Exchange (RINEX) format was introduced by the Astronomical Institute of the University of Berne for exchanging GNSS data with a standard file format. However, it is not applicable for real time data transmission. Most professional grade GNSS receivers support RINEX by providing utilities to convert from their native binary format to RINEX. Receivers typically do not store data natively in RINEX. One of the reasons is because its ASCII-based nature requires large storage size. Most GNSS processing software support RINEX as this is a convenient method to accommodate data from other brands of receivers. It would take a. lot of resources to provide support for other manufacturer's data formats and keeping up with the changes in those formats.
Because it is designed as a format for file exchange, RINEX is not suitable at all to be used as real-time transmission protocol. Its ASCII-based nature requires a lot of bandwidth and processing power and it does not have any mechanism for integrity check to ensure that data was not corrupted during transmission.
The RINEX format covers three different ASCII files: observation files, navigation files, and meteorological files. Each file contains a header session and a data session. The header is placed at the beginning of the file and contains information about the station that collects the data and global information applicable to the entire file. Each observation file and meteorological file contains data from one site and one session, while the navigation RINEX file contains the navigation message broadcasted by the satellites. As such, receivers monitoring the same satellite will receive the same navigation messages. The observation file contains in its header information that describes the file's contents such as the station name, antenna information, the approximate station coordinates, number and types of observation, observation interval in seconds, time of first observation record, and other information. The

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

navigation file contains epoch and satellite clock information. The meteorological file contains time-tagged information such as the temperature (in degrees Celsius), the barometric pressure (in millibars), and the humidity (in percent) at the observation site.
Currently the most commonly used RINEX versions is RINEX 2.11 which allows to store pseudorange carrier-phase and Doppler measurements from GPS, GLONASS and also augmentation systems like EGNOS and WAAS in the same file. With the development of new global navigation satellite systems like Galileo and BEIDOU, it became clear that a new standard, which will full integrate all the satellite system, is needed. With this purpose RINEX Version 3 has been developed and it is under revision.

RINEX-Like Standards

After the presentation of RINEX format, several RINEX-like formats have been defined, mainly used by the International GNSS Service (IGS):

- IONEX: Exchange format for ionosphere models determined by processing data of a GNSS tracking network.
- ANTEX: Exchange format for phase center variations of geodetic GNSS antennae

Exchange format for satellite and receiver clock offsets determined by processing data of a GNSS tracking network.
The latest version of RINEX format is 3.03 from August 2015.
More information is available at:

- ftp://igs.org/pub/data/format/rinex303.pdf
- http://www.navipedia.net/index.php/Interfaces_and_Protocols

NMEA 0183

NMEA is an acronym for the National Marine Electronics Association, formed in 1957, before the invention of GPS, the first GNSS. Today NMEA is a standard data format supported by all GNSS receiver manufacturers, much like ASCII is the standard for digital computer characters in the computer world.
GPS receiver communication is defined within this specification. Most computer programs that provide real time position information understand and expect data to be in NMEA format. This data includes the complete PVT (position, velocity, time) solution computed by the GPS receiver. The idea of NMEA is to send a line of data called a sentence that is totally self-contained and independent from other sentences. There are standard sentences for each device category and there is also the ability to define proprietary sentences for use by the individual company. All of the standard sentences have a two letter prefix that defines the device that uses that sentence type (for GPS receivers the prefix is GP.) which is followed by a three letter sequence that defines the sentence contents. In addition NMEA permits hardware manufactures to define their own proprietary sentences for whatever purpose they see fit. All proprietary sentences begin with the letter P and are followed with 3 letters that identifies the manufacturer controlling that sentence. For example a Garmin sentence would start with PGRM and Magellan would begin with PMGN.

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

Each sentence begins with a '$' and ends with a carriage return/line feed sequence and can be no longer than 80 characters of visible text (plus the line terminators). The data is contained within this single line with data items separated by commas. The data itself is just ascii text and may extend over multiple sentences in certain specialized instances but is normally fully contained in one variable length sentence. The data may vary in the amount of precision contained in the message. For example, time might be indicated to decimal parts of a second or location may be show with 3 or even 4 digits after the decimal point. Programs that read the data should only use the commas to determine the field boundaries and not depend on column positions. There is a provision for a checksum at the end of each sentence which may or may not be checked by the unit that reads the data. The checksum field consists of a '*' and two hex digits representing an 8 bit exclusive OR of all characters between, but not including, the '$' and '*'. A checksum is required on some sentences.
What makes NMEA a bit confusing is that there are quite a few “NMEA” messages, not just one. So, just like there are all kinds of GPS receivers with different capabilities, there are many different types of NMEA messages with different capabilities. Furthermore, NMEA data can be transmitted via different types of communications interfaces such as RS-232, USB, Bluetooth, Wi-Fi, UHF, and many others.
More information is available at:

- https://www.nmea.org/
- http://www.tronico.fi/OH6NT/docs/NMEA0183.pdf

NGS-SP3

To facilitate exchanging precise orbital data, the U.S. National Geodetic Survey developed the SP3 format, which later became the international standard. The SP3 is an acronym for Standard Product #3, which was originally introduced as SP1 in 1985. The SP3 file is an ASCII file that contains information about the precise orbital data and the associated satellite clock corrections. The line length of the SP3 files is restricted to 60 columns (characters). All times are referred to the GPS time system in the SP3 data standards.
A precise ephemeris file in the SP3 format consists of two sections: a header and data. The header section is a 22-line section. The first line starts with the version symbols (#a) and contains information such as the Gregorian date and time of day of the first epoch of the orbit, and the number of epochs in the ephemeris file. Line 2 starts with the symbols (##) and shows the GPS week number, the seconds of the week, the epoch interval, and the modified Julian day. Lines 3-7 start with the symbol (+) and show the total number of satellites (on line 3) as well as list the satellites by their respective identifiers (PRN number). Lines 8-12 start with the symbols (++) and show the accuracy exponents for the satellites shown on lines 3-7. The meaning of the accuracy exponent (ae) is explained as follows: the standard deviation of the orbital error for a particular satellite = 2ae mm. Lines 13-19 of the SP3 header are reserved for future modification, while lines 19-22 are used freely for comments.
More information is available at: https://www.ngs.noaa.gov/orbits/SP3_format.html

GPS Exchange Format

GPX is a light-weight XML data format for the interchange of GPS data (waypoints, routes, and tracks) between applications and Web services on the Internet. It is an XML schema designed

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

as a common GPS data format for software applications. It can be used to describe waypoints, tracks, and routes. The format is open and can be used without the need to pay license fees. Location data (and optionally elevation, time, and other information) is stored in tags and can be interchanged between GPS devices and software. Common software applications for the data include viewing tracks projected onto various map sources, annotating maps, and geotagging photographs based on the time they were taken.
The essential data types contained in GPX files are:

- wptType is an individual waypoint among a collection of points with no sequential relationship. It consists of the WGS 84 GPS coordinates of a point and possibly other descriptive information.
- rteType is a route, an ordered list of route point (waypoints representing a series of significant turn or stage points) leading to a destination.
- trkType is a track, made of at least one segment containing waypoints, that is, an ordered list of points describing a path. A Track Segment holds a list of Track Points which are logically connected in order. To represent a single GPS track where GPS reception was lost, or the GPS receiver was turned off, start a new Track Segment for each continuous span of track data.

Conceptually, tracks are a record of where a person has been and routes are suggestions about where they might go in the future. For example, each point in a track may have a timestamp (because someone recorded where and when they were there), but the points in a route are unlikely to have timestamps (other than estimated trip duration) because route is a suggestion which might never have been traveled.
The minimum properties for a GPX file are latitude and longitude for every single point. All other elements are optional. Some vendors use extensions to the GPX format for recording street address, phone number, business category, air temperature, depth of water, and other parameters.
More information is available at: http://www.topografix.com/gpx/1/1/

KML

Keyhole Markup Language (KML) is a file format used to display geographic data in an Earth browser such as Google Earth. KML uses a tag-based structure with nested elements and attributes and is based on the XML standard. All tags are case-sensitive and must appear exactly as they are listed in the KML Reference. The Reference indicates which tags are optional. Within a given element, tags must appear in the order shown in the Reference.
The KML file specifies a set of features (place marks, images, polygons, 3D models, textual descriptions, etc.) for display in Here Maps, Google Earth, Maps and Mobile, or any other geospatial software implementing the KML encoding. Each place always has a longitude and a latitude. Other data can make the view more specific, such as tilt, heading, altitude, which together define a "camera view" along with a timestamp or timespan. KML shares some of the same structural grammar as GML (Geography Markup Language). Some KML information cannot be viewed in Google Maps or Mobile.
KML files are very often distributed in KMZ files, which are zipped KML files with a .kmz extension. These must be legacy (ZIP 2.0) compression compatible (i.e. stored or deflate

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

method), otherwise the .kmz file might not uncompress in all geobrowsers. The contents of a KMZ file are a single root KML document (notionally "doc.kml") and optionally any overlays, images, icons, and COLLADA 3D models referenced in the KML including network-linked KML files. The root KML document by convention is a file named "doc.kml" at the root directory level, which is the file loaded upon opening. By convention the root KML document is at root level and referenced files are in subdirectories (e.g. images for overlay images).
For its reference system, KML uses 3D geographic coordinates: longitude, latitude and altitude, in that order, with negative values for west, south and below mean sea level if the altitude data is available. The longitude, latitude components (decimal degrees) are as defined by the World Geodetic System of 1984 (WGS84). The vertical component (altitude) is measured in meters from the WGS84 EGM96 Geoid vertical datum. If altitude is omitted from a coordinate string, e.g. (-77.03647, 38.89763) then the default value of 0 (approximately sea level) is assumed for the altitude component, i.e. (-77.03647, 38.89763, 0).
More information is available at: https://developers.google.com/kml/documentation/

Differential GNSS data exchange standards

Differential GNSS (DGNSS) is a kind of GNSS augmentation system based on an enhancement to primary GNSS constellation information by the use of a network of ground-based reference stations which enable the broadcasting of differential information to the user to improve the accuracy of his position – the integrity is not assured. There are several DGNSS techniques, such as the classical DGNSS (or DGPS), the Real-Time Kinematics (RTK) and the Wide Area RTK (WARTK).
The internationally accepted data transmission standards for DGNSS are defined by Radio Technical Commission for Maritime Services (RTCM), particularly by its Special Committee SC-104.

RTCM SC-104

The Radio Technical Commission for Maritime Services SC-104 has introduced formats and protocols that are now accepted as international standards. The data protocol has evolved over many years by incorporating new message types. On the other hand, the Networked Transport of RTCM via Internet Protocol (NTRIP) and RT-IGS protocols were developed as network transport protocols to deliver GNSS data via the internet.
RTCM is a standard that defines the data structure for differential correction information for a variety of applications. It has become an industry standard for communication of correction information. RTCM is unreadable with a terminal program as it is a binary data protocol. All GNSS receivers support RTCM v2.x messages for DGNSS positioning. However, it does not support RTCM v2.x messages for RTK positioning. RTCM v3.x messages are suitable for RTK positioning. The error correction data sent by this differential GNSS protocol is quite heavy; it requires at least 19.2kbps of bandwidth for data transfer in RTK (see below) mode. Here, a radio frequency of UHF or higher is required to achieve this data rate. RTCM's standard supports very high accuracy navigation and positioning through a broadcast from a reference station to mobile receivers. These messages contain information such as the pseudorange correction (PRC) for each satellite in view of the reference receiver, the rate of change of the pseudo-range corrections (RRC), and the reference station coordinates.
More information is available at: http://www.rtcm.org/

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

Real Time Kinematics

Real Time Kinematics (RTK) is a differential GNSS technique originated in the mid-1990s that provides high performance positioning in the vicinity of a base station. From an architectural point of view, RTK consists of a base station, one or several rover users, and a communication channel with which the base broadcasts information to the users at real time.
The technique is based on the following high-level principles:

- In the neighbourhood of a clean-sky location, the main errors in the GNSS signal processing are constant, and hence they cancel out when differential processing is used. This includes the error in the satellite clock bias, the satellite orbital error, the ionospheric delay and the tropospheric delay.
- The noise of carrier measurements is much smaller than the one of the pseudo-code measurements. However, the processing of carrier measurements is subject to the so-called carrier phase ambiguity, an unknown integer number of times the carrier wave length, that needs to be fixed in order to rebuild full range measurements from carrier ones.
- The phase ambiguity can be fixed for dual-frequency differential measurements for two close receivers.
- The base station broadcasts its well-known location together with the code and carrier measurements at frequencies L1 and L2 for all in-view satellites. With this information, the rover equipment is able to fix the phase ambiguities and determine its location relative to the base with high precision. By adding up the location of the base, the rover is positioned in a global coordinate framework.
- The RTK technique can be used for distances of up to 10 or 20 kilometres, yielding accuracies of a few centimetres in the rover position. RTK is extensively used in surveying applications.

The main limitations of RTK are as follows:

- Limited range with respect to the base location.
- The need of a communication channel for real time applications.
- Some convergence time is needed to fix the phase ambiguities. This time depends on the processing algorithm and the distance between base and rover, and ranges from a few seconds to a few minutes.
- In order to avoid re-initialization of the processing, the rover has to track the GNSS signals continuously. This makes the RTK not suitable for urban applications.

Recently, different approaches have been followed to improve the limitation regarding the range of the base station, namely Network RTK and Wide Area Real Time Kinematics (WARTK). Network RTK is based on the provision of corrections from a network of base stations in such a way that the phase measurements are provided with consistent ambiguities; this has the

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

advantage that the rover can switch from one base station to another without the need of re-initializing the ambiguity fixing filters.
The standards applying to RTK systems are the same of classical DGNSS systems, i.e. the ones of data transmission standards defined by the Special Committee 104 on DGNSS of the Radio Technical Commission for Maritime Services (RTCM).
More information is available at:

- https://www.e-education.psu.edu/geog862/node/1838
- https://en.wikipedia.org/wiki/Real_Time_Kinematic

CMR+

The CMR (Compact Measurement Record) protocol was developed by a receiver manufacturer Trimble Navigation, and then made public. Since then, other manufacturers such as Leica, Ashtech, NovAtel and Topcon have included support for CMR in their receivers. CMR provided a more bandwidth-efficient altemative to RTCM Version 2 for GPS RTK users. CMR+ is a slightly improved version of CMR that has a less peaked throughput. Data streams in CMR/CMR+ format are available from CORS networks in Victoria (GPSnet), New South Wales (SydNET) and Queensland (SunPOZ). Network-RTK software such as T1imble GPSnet, Leica Spider and UNS'V SIMRSN can also produce streams in CMR/CMR+ format.
More information is available at:
http://gps.0xdc.ru/static/sirf/doc/SirfStar/gpsd.berlios.de/vendor-docs/trimble/cmr.pdf

GNSS on Smartphones

Every smartphone includes support for at least one GNSS provider, some of them even for all four of them (GNSS data protocols: choice and implementation, IGNSS symposium 2006, Australia, http://rap.uca.es/web_RAP/documentacion/2006-GNSS_data_protocol_choice_and_implementation.pdf). Current GNSS chips in smartphones provide a location accuracy within 5-meters of the actual location. The new generation of chips will be able to increase that to 30 cm (http://www.navipedia.net/index.php/Interfaces_and_Protocols). Along with the improved accuracy, the new GPS chip should also perform better in cities and other urban locations where there are a lot of tall buildings and other concrete structures.
The smartphone operating systems provide the necessary high-level functions to access the GNSS data. Several examples and manuals are available on the net helping the developers to code their apps.

Android

Most devices manufactured in 2016 or later and shipped with Android 7.0 or higher provide raw GNSS data. Depending on the device, raw GNSS measurements can include all or some of the following data (https://developer.android.com/guide/topics/sensors/gnss.html):

- Pseudorange and pseudorange rate.
- Navigation messages.
- Accumulated delta range or carrier.
- Hardware clock.

More often the application needs only location information from the integrated GNSS module. It is worth pointing out that location information contains much more than just

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

latitude and longitude values. It can also have values for the bearing, altitude and velocity of the device (http://gpsworld.com/what-exactly-is-gps-nmea-data/).The functions needed to programmatically obtain location and other GNSS-connected information are a part of several android classes, such as (http://www.gpsinformation.org/dale/nmea.htm): LocationProvider, LocationManager, Location, GpsStatus, GpsSatellite, GnssStatus, GnssNavigationMessage, GnssMeasurementsEvent, GnssMeasurement, GnssClock, Geocoder etc. The proper use of this classes and detailed instructions with examples are available on the net (https://developer.android.com/guide/topics/location/strategies.html; https://www.androidauthority.com/get-use-location-data-android-app-625012/).

IOS

In IOS applications get location data through the classes of the Core Location framework. This framework provides several services that can be used to get and monitor the device's current location (https://developer.apple.com/library/content/documentation/UserExperience/Conceptual/LocationAwarenessPG/CoreLocation/CoreLocation.html):

- The standard location service offers a highly configurable way to get the current location and track changes.
- Region monitoring lets you monitor boundary crossings for defined geographical regions and Bluetooth low-energy beacon regions. (Beacon region monitoring is available in iOS only.)
- The significant-change location service provides a way to get the current location and be notified when significant changes occur, but it's critical to use it correctly to avoid using too much power.

The Core Location framework allows to locate the current position of the device and use that information in user's app. The framework reports the device's location to the code and, depending on how the service is configured, also provides periodic updates as it receives new or improved data.

Two services can give the user's current location:

- The *standard location service* is a configurable, general-purpose solution for getting location data and tracking location changes for the specified level of accuracy.
- The *significant-change location service* delivers updates only when there has been a significant change in the device's location, such as 500 meters or more.

Gathering location data is a power-intensive operation. For most apps, it's usually sufficient to establish an initial position fix and then acquire updates only periodically after that. Regardless of the importance of location data in the app, one should choose the appropriate location service and use it wisely to avoid draining a device's battery.

Windows Phone

All Windows Phone devices are required to ship with integrated GNSS receiver. As with the Android and IOS the function to obtain the location information are available within the operating system. For example, in C# one need to reference System.Device assembly and declare *System.Device.Location* before taking advantage of location service which must be turned on. Next, one needs to set Desired Accuracy and provide MovementThreshold. The

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

actual position data is obtained by subscribing to PositionChanged event (Lee H, Chuvyrov, E. Beginning Windows Phone App Development, Apress, 2012).
There are other online tutorials and programming examples available how to how to get and handle GNSS coordinates under Windows Phone (http://windowsapptutorials.com/windows-phone/location/get-gps-coordinates-in-windows-phone-8-1-app/; http://www.thewindowsclub.com/gps-location-api-calling-web-services-windows-phone-apps-development-tutorial-part-25).

3.1.3 Consumer data

Consumer data is mainly collected by mobile apps that may comply with social network standards, such as Twitter, Facebook, Instagram, Pinterest, Snapchat etc., or self-created standards. Social data contains information about the socio-psychological determinants of diet and lifestyle.
Another important source of massive data is the Internet of Things (IoT). Addressing interoperability challenges among IoT devices is still unfolding. Many industry coalitions (such as the Industrial Internet Consortium (http://www.iiconsortium.org), ZigBee Alliance (http://www.zigbee.org), and AllSeen Alliance (https://openconnectivity.org/developer/reference-implementation/alljoyn), among many others) have emerged alongside traditional Standards Developing Organizations (like IETF (http://www.ietf.org), ITU (https://www.itu.int), and IEEE (https://iot.ieee.org)) to increase efforts to work on standards and protocols related to IoT. In recent years, several ontology-based approaches for IoT data processing have been developed. Let us mention Agri-IoT as an example of a semantic framework for IoT-enabled smart farming applications (A. Kamilaris A, Gao F, Prenafeta-Boldu FX and Ali MI, Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications, 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, 2016, pp. 442-447; doi:10.1109/WF-IoT.2016.7845467).

Social network data standards

In the following subsections, we will present few standards that are applied by social media, such as Twitter, Instagram and Facebook, to represent and share data. Once data is collected from social media, ontologies are used to enrich social information with the power of semantics and enable the integration and normalisation of data. Let us mention two ontologies (https://link.springer.com/chapter/10.1007/978-981-10-6620-7_25):

- The Friend of A Friend (FOAF) ontology (http://www.foaf-project.org/)
- Socially Interconnected Online Communities (http://sioc-project.org/)

More about the techniques that are needed to extract knowledge from social data is described in D11.2.

Twitter

Data collected by Twitter includes tweets, tweet IDs, Twitter end user profile information, periscope broadcasts (i.e. live or on-demand video streams), broadcast IDs, geo-locations, etc. All this data is made available through Twitter APIs, which are well documented at the Twitter's developer site located at https://developer.twitter.com. The standard (free) Twitter APIs consist of a REST API and a Streaming API. The Streaming API provides low-latency access to tweets. Additionally, there are some families of APIs (such as the Ads API) which require

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

applications to be whitelisted in order to make use of them. With the exception of the Streaming API, the Twitter API endpoints attempt to conform to the design principles of Representational State Transfer (REST). Twitter APIs use the JSON data format for responses (and in some cases, for requests).
Some API methods take optional or requisite parameters:

- Parameter values should be converted to UTF-8 and URL encoded.
- The page parameter begins at 1, not 0.

Where noted, some API methods will return different results based on HTTP headers sent by the client. Where the same behavior can be controlled by both a parameter and an HTTP header, the parameter will take precedence.
Twitter uses Snowflake as a service used to generate unique IDs for objects within Twitter (tweets, direct Messages, users, collections, lists etc.). These IDs are unique 64-bit unsigned integers, which are based on time, instead of being sequential. The full ID is composed of a timestamp, a worker number, and a sequence number. When consuming the API using JSON, it is important to always use the field id_str instead of id. This is due to the way Javascript and other languages that consume JSON evaluate large integers. If you come across a scenario where it doesn't appear that id and id_str match, it's due to your environment having already parsed the id integer, munging the number in the process. More information on how Twitter generates its ids is provided at https://developer.twitter.com/en/docs/basics/twitter-ids.
As tweets distinguish with the noisy nature of text, machine learning approaches such as Named Entity Recognition (NER) need to be applied in order to extract information from data. This challenging task is described in D11.2 in more details.

Instagram

Data collected by Instagram mainly include media objects, comments and meta-data (like hashtags), but recently Instagram has already supported a text-based 'Type' feature for Stories. This data is accessible through the Instagram API Platform, which will be deprecated beginning in July 2018. There exists also the Instagram (actually Facebook) Graph API that allows to programmatically access Instagram Business Accounts in order to more easily manage media objects, view comments and meta-data, and get insights and metrics with custom built apps.
Another major restriction is the quality of available data. For example, you cannot search for a keyword, only for a hashtag. This limits the ability for threat discovery through text. Instagram recently removed their photo maps view from their mobile application. This means that geo-location information has been moved from open-source access.
More information is available at: https://www.instagram.com/developer/

Facebook

The Facebook API functions have been split among the following APIs: Facebook Ads, Facebook Atlas, Facebook Graph, and Facebook Marketing. These APIs allow applications to use the social connections and profile information to make applications more involving, and to publish activities to the news feed and profile pages of Facebook, subject to individual users privacy settings. With the API, users can add social context to their applications by utilizing

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

profile, friend, page, group, photo, and event data. The API uses RESTful protocol and responses are in JSON format.
The Facebook Graph API is the primary way to get data out of, as well as put data into, Facebook's platform. It is named after the idea of a 'social graph' - a representation of the information on Facebook composed of:

- nodes - basically "things" such as a User, a Photo, a Page, a Comment. Each node has a unique ID, which is used to access it via the Graph API.;
- edges - the connections between those "things", such as a Page's Photos, or a Photo's Comments;
- fields - info about those "things", such as a person's birthday, or the name of a Page.

The Graph API is HTTP-based, so it works with any language that has an HTTP library, such as cURL and urllib. It means that the Graph API can be used directly in a browser.
More information is available at: https://developers.facebook.com/docs/.

IoT standards for data delivery

IoT data is created by a smart device or a sensor in three stages. The first stage is the initial creation of data, which takes place on the device, and then the data is sent over the Internet. The second stage is how the central system collects and organizes that data. The final stage is the ongoing use of that data for the future. The most common standard protocols used for the delivery of data are HTTP, MQTT (http://mqtt.org) and CoAP (http://coap.technology):

- HTTP provides a method for providing data back and forth between devices and central systems. Originally developed for the client-server computing model, today it supports everyday web browsing through to more specialist services around IoT devices too. While it meets the functionality requirements for sending data, HTTP includes a lot more data around the message in its headers. When working in low bandwidth conditions, this can make HTTP less suitable;
- MQTT (Message Queuing Telemetry Transport) was developed as a protocol for machine-to-machine and IoT deployments. It is based on a publish/subscribe model for delivering messages out from the device back to a central system that acts as a broker, where they can then be delivered back out to all of the other systems that will consume them. New devices or services can simply connect to the broker as they need messages. MQTT is lighter than HTTP in terms of message size, so it is more useful for implementations where bandwidth is a potential issue. However, it does not include encryption as standard so this has to be considered separately;
- CoAP (Constrained Application Protocol) is another standard developed for low-power, low-bandwidth environments. Rather than being designed for a broker system like MQTT, CoAP is more aimed at one-to-one connections. It is designed to meet the requirements of REST design by providing a way to interface with HTTP, but still meet the demands of low-power devices and environments.

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

Depending on the IoT device, the network and power consumption restraints, data can be sent in real time, or in batches at any time. However, the real value is derived from the order in which data points are created. This time-series data has to be accurate for IoT applications. If the order of data is not completely aligned and accurate, then it points to potentially different results when analyzed. Each write has to be taken as it is received from the device itself and put into the database. For more traditional relational database technologies this can be a limiting factor as it is possible for write-requests to go beyond what the database was built for. NoSQL platforms (like MongoDB, Apache Cassandra, etc.) provide a better fit for their requirements.

In Self-created standards

In RICHFIELDS, we performed a case study on physical exercise data collected by a mobile app PRECIOUS that was developed in the FP7 project PREventive Care Infrastructure based On Ubiquitous Sensing (http://www.theprecioussproject.eu). The app is aimed to enhance preventive health and wellbeing care with advanced, transparent sensing and scientifically developed feedback structures. In PRECIOUS, a Semantic Vocabulary Creation Cycle (CSVCC) was used for collecting parameter definitions. In D11.2 we present a methodology for enhancing PRECIOUS data with semantics needed to link consumer data with scientific and business ones.

3.2 Standards for data linkage and harmonization

Exploring standards for data linkage and harmonization, special focus was given to work programmes of big data standardization consortia, like W3C Building the Web of Data, ITU-T SG13 Future networks including cloud computing, mobile and next-generation networks, ISO/IEC JTC 1/SC 32 Data management & interchange, RDA Europe - Research Data Alliance, OASIS Organization for the Advancement of Structured Information Standards, and TPC.

W3C Data Activity - Building the Web of Data

The W3C Data Activity (https://www.w3.org/2013/data/) works to overcome a diversity of “big” and “small” data to facilitate potentially Web-scale data integration and processing. It does this by providing standard data exchange formats, models, tools, and guidance.
In December 2017, the W3C Data Activity group published a report about the W3C study of practices and tooling for Web data standardisation (https://www.w3.org/2017/12/odi-study/). The report starts with an introduction to the Web of data and W3C’s standardisation activities. This is followed by a section on the challenges for measuring the popularity of standards, the need to support the communities that develop and use them, and how to gather feedback that can be used to improve standards and identify gaps where new standards are needed. The report closes with a look at the potential of multidisciplinary approaches including AI, Computational Linguistics and Cognitive Science to transform the process of creating standards, and to evolve the Semantic Web into the Cognitive Web. In this report it is interesting to read that the difficulty of manually creating complex ontologies can in principle be avoided through the use of machine learning algorithms that are applied to a training corpus. In D11.2 we have already presented methodology for semi-automatic creation of a food and consumer behaviour ontology based on natural language processing.

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

ITU-T SG13 Future networks including cloud computing, mobile and next-generation networks

ITU-T is the United Nations specialized agency for information and communication technologies (ICTs), which develops technical standards to ensure networks and technologies seamlessly interconnect (https://www.itu.int). ITU-T also deals with the e-Health standardization, which has resulted in Resolution 78 - Information and communication technology applications and standards for improved access to e-Health.
In 2017, a roadmap of ITU-T e-Health standardization was published at https://www.itu.int/en/ITU-T/studygroups/2017-2020/16/Pages/rm/ehealth.aspx. This roadmap defines a framework of underlying standards and criteria required to ensure the interoperability of devices and data used for personal connected health. It also contains design guidelines that further clarify the underlying standards or specifications by reducing options or by adding a missing feature to improve interoperability. These guidelines focus on interfaces for personal health devices, services and healthcare information systems.

ISO/IEC JTC 1/SC 32 Data management & interchange

SC 32 works on standards for data management within and among local and distributed information systems environments. It provides enabling technologies to promote harmonization of data management facilities across sector-specific areas. Specifically, SC 32 standards include (https://www.iso.org/committee/45342/x/catalogue/p/0/u/1/w/0/d/0):

- reference models and frameworks for the coordination of existing and emerging standards;
- definition of data domains, data types, and data structures, and their associated semantics;
- languages, services, and protocols for persistent storage, concurrent access, concurrent update, and interchange of data;
- methods, languages, services, and protocols to structure, organize, and register metadata and other information resources associated with sharing and interoperability, including electronic commerce.

RDA Europe - Research Data Alliance

is the European plug-in to RDA that is an international member-based organisation focused on the development of infrastructure and community activities to reduce the social and technical barriers to data sharing and re-use and to promote the acceleration of data driven innovation and discovery worldwide. A list of RDA endorsed recommendations is accessible at https://www.rd-alliance.org/recommendations-and-outputs/all-recommendations-and-outputs. Let us mention few of most relevant ones:

- Scalable Dynamic-data Citation methodology, supporting accurate citation of data subjected to change, for the efficient processing of data and linking from publications;
- Data Description Registry Interoperability Model, providing researchers with a mechanism to connect datasets in various data repositories based on various models such as co-authorship, joint funding, grants, etc.;

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

- Data Type Model and Registry, ensuring data producers classify their data sets in standard data types, allowing data users to automatically identify instruments to process and visualise the data;
- Workflows for Research Data Publishing: Models and Key Components, assisting research communities in understanding options for data publishing workflows and increases awareness of emerging standards and best practices.

OASIS Organization for the Advancement of Structured Information Standards

is a non-profit consortium that drives the development, convergence and adoption of open standards for the global information society. It promotes industry consensus and produces worldwide standards for security, Internet of Things, cloud computing, energy, content technologies, emergency management, and other areas. OASIS open standards offer the potential to lower cost, stimulate innovation, grow global markets, and protect the right of free choice of technology. OASIS Committee Specifications are listed at https://www.oasis-open.org/standards#oasiscommiteespecs.

TPC

stands for the Transaction Processing Performance Council that is a non-profit corporation founded to define transaction processing and database benchmarks and disseminate objective, verifiable TPC performance data to the industry (http://www.tpc.org/default.asp). A list of active TPC benchmarks is listed at http://www.tpc.org/tpc_documents_current_versions/current_specifications.asp. Let us mention TPCx-HS Big Data Benchmark developed to provide an objective measure of hardware, operating system and commercial Apache Hadoop File System API compatible software distributions. TPCx-HS stresses both the hardware and software stacks including the execution engine (MapReduce or Spark) and Hadoop Filesystem API compatible layers. This workload can be used to assess a broad range of system topologies and implementation of Hadoop clusters.

4 Conclusions and implications

In this deliverable, we identified currently available food and nutrition data sources and its formats, which need to be considered by the RICHFIELDS data semantics (please see D11.2). As data is defined by standards used to describe and classify scientific, business and consumer data, an overview of well-established standards for data collection and data linkage and harmonization was done. The first group of standards presented in the deliverable included standards for collecting scientific data, business data and consumer data. We considered standards established by the European networks of excellence, industrial data standardization consortia and the global social media network. The second group included standards for data linkage and harmonization which were established by the European and global data standardization consortia.

We can conclude that the presented standards cover a broad range of data relevant for RICHFIELDS. Most probably these standards will not be changed in short term as existing information systems already rely on them and their adaptation to any new standard would

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

require a lot of efforts and costs. Therefore, at least for the RICHFIELDS MVP, there is still a strong need for the further development of ontologies used to enrich information from collected data with the power of semantics, which is needed to enable the integration and normalisation of data. More details about data semantics is provided in D11.2, and we strongly suggest that methodologies for semantic enrichment of data is considered in the final design of the RICHFIELDS platform (to be discussed more throughly in D11.4 on the RICHFIELDS roadmap).

The overview of standards partly answers the first question stated in the Introduction „What data can be readily incorporated into the data platform at the Minimum Viable Product (MVP) level from an availability/ethical perspective?“. It addresses different types of data identified as ready to be incorporated at the MVP level with a focus on data formats and not data quality nor data quality. The second question „Are these data of sufficient value to the proposed primary users; If not how will the additional data required be obtained?“ can also be partly answered. Using standards for data linkage and harmonization, data of insufficient value can be linked and harmonized with other data.

It would also be helpful to keep information about relevant standards, ontologies and methodologies in the RIMS or other similar information system. In the near future, most probably, additional standards, ontologies and methodologies will develop and it is recommended to keep track of the development. Food covers a broad area of interests, including also fields like payment/banking, time management, etc. that have not been included in D11.3. There might also be a need for the development of a new standard, which would cover an integrative field of food, health and determinants, such as dietary advising that includes health and sustainability considerations. However, standardization requires extensive work to cover all dimensions. For this reason, we suggest that at the MVP level, focus is given on the development of the RICHFIELDS ontology and linkage and harmonisation methods to make sure that both distributed data or locally stored data have one meaning.

Another problem is, that despite all the buzz about the unprecedented volumes of data that humanity generates every day, the fact remains that databases (such as historical records, paper files, and many other forms of non-digital data recording) remain in an un-digitized form, and thus untapped regarding usage. To extract knowledge from this data ecosystem advanced computer techniques, such as text mining, deep learning etc., will be required.

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