Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package WP11
Deliverable D11.4

Roadmap RI Consumer Data Platform

Date delivered:
M35

Authors:
Barbara Koroušić Seljak, Peter Korošec, Tome Eftimov, JSI; Krijn Poppe, WUR; Indira Carr, USurrey; Javier de la Cueva, Abogado; Paul Finglas, QIB; Mark Roe, EuroFIR

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:	Roadmap RI Consumer Data Platform
Deliverable No.:	D11.4
Authors:	B. Koroušić Seljak, P. Korošec, T. Eftimov, IJS; Indira Carr, USurrey; Javier de la Cueva, Abogado; Paul Finglas, QIB; Mark Roe, EuroFIR
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:	30.08.2018
Due date of deliverable:	31.08.2018
Dissemination level:	PU
Status:	Final

Change history:		
Version	Notes	Date
001	Draft version	16.05.2018
002	Comments provided by Indira, Javier, Paul, Mark and Krijn considered	06.07.2018
final	Comments provided by Fred considered	10.07.2018





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

In this deliverable, a roadmap for developing the RICHFIELDS Data Platform (RDP) is described from the technical point of view. Using agile development of the platform, the roadmap should be a living document, with short timeframes and frequent adjustments to accommodate changing priorities and market opportunities. This version of the document serves as an initial document, describing principles of FAIR, Open Science and Data provenance from which best practices need to be considered by the RDP. To specify the roadmap, first the RDP architectural model is introduced with detailed description of each of the modules. Then a protocol for using the RDP is discussed from different perspectives, i.e., information systems providing or requesting data, and individuals supplying (donating) or searching for data. Finally, steps to be performed with respect to different tasks and stages of the RDP development (i.e., core offering, mature and growth) are specified. As data linkage and harmonization is the most complex task to be developed more or less from scratch, we briefly describe it to make clear the concepts of semantic annotation and ontology population. For interested readers, other details are provided in D11.2. We conclude by providing implications for the final RICHFIELDS roadmap, which considers beside technical aspects presented in this deliverable, also governance, business and legal/ethical aspects.

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

Table of Contents

1 Introduction 6

2 Review of requirements 9

2.1 FAIR principles 10

2.1.1 Findability 11

2.1.2 Accessibility 12

2.1.3 Interoperability 12

2.1.4 Reusability 12

2.2 Open Science principles 13

2.2.1 OpenAIRE 13

2.3 Data Provenance 15

2.3.1 Data provenance semantics 15

2.3.2 Data provenance management 22

3 Data architecture 22

3.1 The model of the RICHFIELDS Data Platform 22

3.1.1 The RICHFIELDS Data Platform 24

3.1.2 The RICHFIELDS input units 25

3.1.3 The RICHFIELDS output units 26

3.2 Protocol for accessing the RDP 26

3.3 Data harmonization 28

3.4 Roadmap for the development of the RDP 32

3.4.1 Launch the RDP as part of the Minimum Viable Product 32

3.4.2 FAIRness approval 35

4 Conclusions and implications 35

References 36

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

Abbreviations

API	Application Programming Interface
CO	Core offering
CRIS	Current Research Information Systems
DOI	Digital Object Identifier
G	Growth
HTTP	Hyper Text Transfer Protocol
IS	Information System
ISO	International Organization for Standardization
M	Maturity / Mandatory
MVP	Minimum Viable Product
REST	Representational State Transfer
RDF	Resource Description Framework
RDP	RICHFIELDS Data Platform
SOAP	Simple Object Access Protocol
UAT	User Acceptance Test
URI	Uniform Resource Identifier

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

1 Introduction

The RICHFIELDS Data Platform (RDP) aims to collect, link and harmonize, analyse, store, and deliver food- and nutrition-related data and information to various stakeholders (data users). Data available through the RDP is collected either by other information systems connected with the RDP or by individuals collaborating with RICHFIELDS (

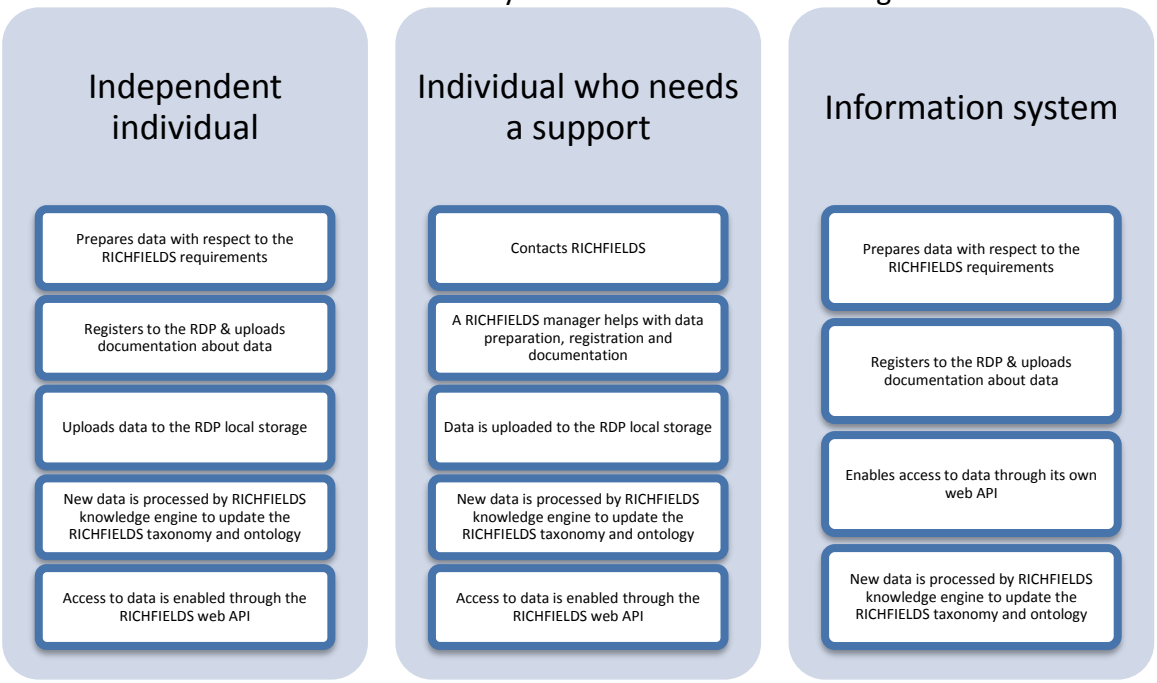


Figure 1), and may be of any type, i.e. structured, semi-structured or unstructured; small or big; open or linked open; raw or processed; aggregated or disaggregated; with or without microdata (i.e., personal data about an individual), etc.:

- Data provided by the individuals (i.e., researchers, business partners, consumers, etc.) is actually stored by the RDP and can be accessed via the web RICHFIELDS Application Programming Interface (API).
- Data collected by other information systems can only be accessed via web APIs defined by those information systems (this data won't be stored by the RDP).

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

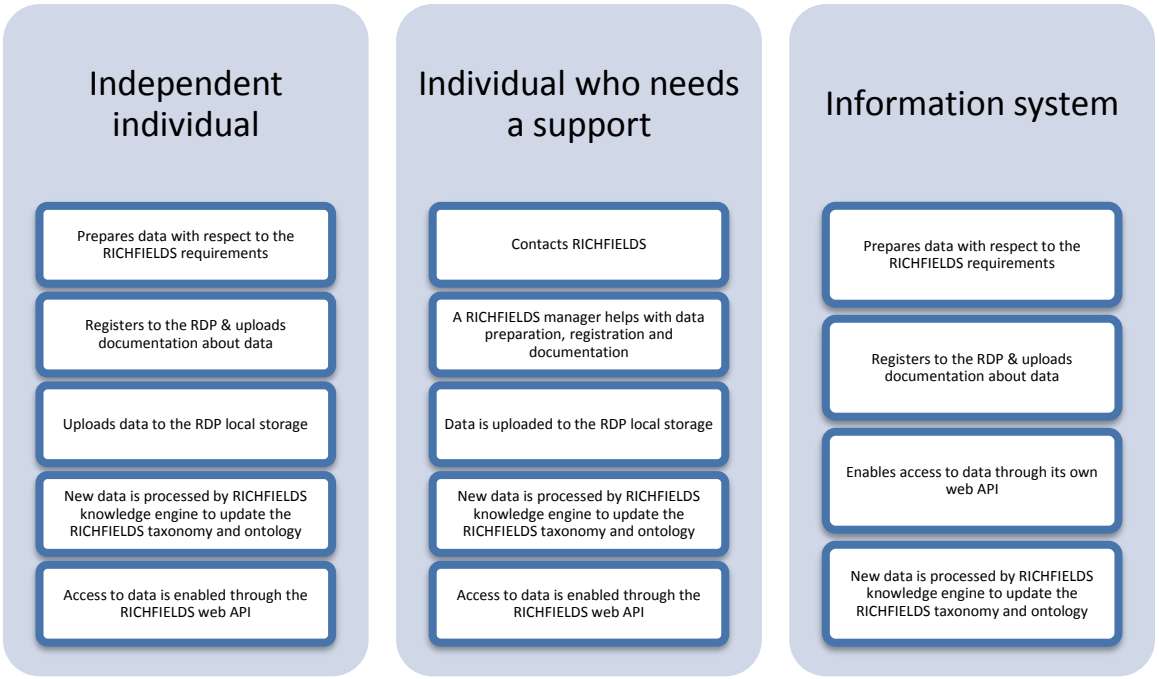


Figure 1. Data flows through the RDP (more details are provided in the following sections).

The above-mentioned options for accessing data via web APIs are related to information systems connected with the RDP and not to individuals requesting data. Individuals supplying (donating) data will be able to contact RICHFIELDS managers to support them in storing their data by the RDP. Those individuals that are able to prepare data as requested by RICHFIELDS, will be able to automatically upload data to the RDP. Individuals interested in using data will also have a possibility of contacting RICHFIELDS managers to support them in acquiring requested data in an appropriate format. More detailed protocol for accessing the RDP is provided in the Section 3.2.

Instructions on how to use the RDP (by information systems) or reach RICHFIELDS managers (by individuals) will be clearly given on the public RICHFIELDS website, where also up-to-date data and knowledge catalogues and the information about research protocols and the RICHFIELDS taxonomy and ontology will be published.

It is also important to stress that RICHFIELDS will not only provide access to data but also to its related metadata and provenance data (for details please continue reading other Sections, e.g., Section 3.3 on Data harmonization).

More information about the general concept of the RDP is provided in D11.1 (User requirements' specification).

What is a difference between data, information and knowledge?

In general, *data* is a collection of facts, such as numbers, words, measurements, observations, descriptions of things (e.g., pictures and images, voices, aromas, etc.), carrying numeric or descriptive information. When data is processed, organized, structured or presented in a given context that makes it useful, it is called *information*. Both data and

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

information are used to build *knowledge*, which provides a framework for evaluating and incorporating new data and information. For example: 25°C, 24°C, 22°C, 28°C, 30°C is a list of data. If we add the location (e.g. London) and the dates (e.g. 1st, 2nd, 3rd, 4th and 5th July 2018) when data was obtained, then it is information. If the data was not from London but from the North Pole, then the same information could be a global warming evidence, that is to say knowledge.

What is meant by data collection, linkage, harmonization, analysis and delivery?
Data collection is the acquisition of data from digital data resources (e.g. electronic databases, spreadsheets, text documents, pdf documents).

Data linkage is a technique for connecting pieces of information that are related to the same entity, described by its metadata (Figure 2).

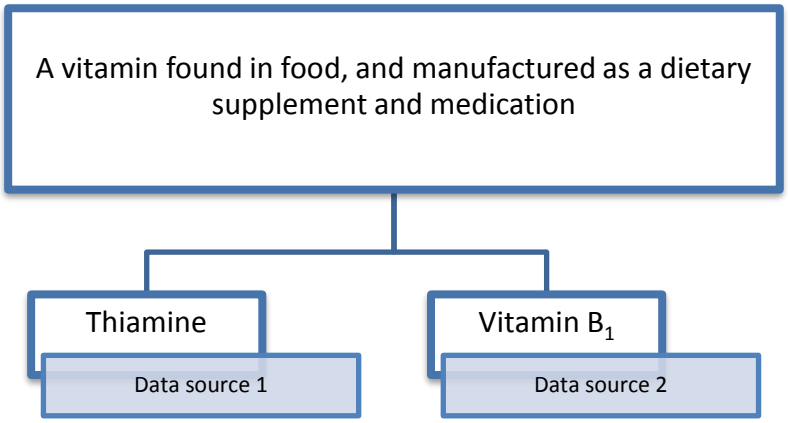


Figure 2. Data linkage.

Data harmonization means the improvement of data quality and its utilization through natural language processing, Semantic Web and data mining. More information about data harmonization and how it can be performed by specific state-of-the-art methods, developed within the RICHFIELDS project, is provided in D11.2.

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

Data delivery enables data sharing and exchange with other information systems.

What are API, web service and web API?
Starting with definitions, *API* stands for *Application Programming Interface* and, in general, presents a software intermediary that allows two applications to talk to each other. For example, when we use a mobile application, the application connects to the Internet and sends data to a web server. The web server then retrieves that data, interprets it, performs

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

the necessary actions and sends it back to our mobile device. The application then interprets that data and presents us with the information we wanted in a readable way.

Then, a *web service* commonly provides a web-based interface to a database server, utilized for example by another web server, or by a mobile application, that provides a user interface to the end user. There are two main protocols used for web services: Simple Object Access Protocol (SOAP) was developed as a messaging protocol for exchanging structured information (e.g., XML), while its younger alternative is Representational State Transfer (REST) that is based on HTTP (Hyper Text Transfer Protocol). Both facilitate information transfer, but REST is considered to be more flexible and dynamic than SOAP. REST allows a greater variety of data formats, is generally faster and uses less bandwidth. There is a difference between APIs and web services; while all web services are APIs, there are APIs that are not web services (do not require a network for their operation).

A *Web API* is an API for either a web server or a web browser. In this document, we use terms APIs and web services applying to APIs.

How to design the RICHFIELDS Data Platform?

Our goal is to develop the RICHFIELDS Data Platform using good design, which in general means, according to Dieter Rams (Rams, 2018), that

1. is innovative,
2. makes a product useful,
3. is aesthetic,
4. makes a product understandable (for its users),
5. is unobtrusive,
6. is honest,
7. is long-lasting (sustainable),
8. is thorough down to the last detail,
9. is environmentally friendly, and
10. involves as little design as possible.

In our case, the first two principles apply to the techniques for data linkage, harmonization and analysis (described in D11.2), while the third and the fourth principles relate to the RDP user interface that should not only be aesthetic but also advanced enough to be truly user-friendly. The ninth principle is maybe not so obvious but is relevant. RICHFIELDS should have green computing initiatives to maximize energy efficiency during the RDP lifetime as well as to reduce the environmental effect of its operations.

To achieve the above listed principles of good design, we specified a **product roadmap** as a high-level visual summary that maps out the vision and direction of the RDP offering over time. The roadmap presented in this deliverable is a guiding vision and strategic document as well as a plan for executing the strategy. It is an internal roadmap for engineers aligned with the internal roadmap for executives (defined in WP12) and the external roadmap for stakeholders (defined in WP13).

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

Using the *agile development* of the platform, the roadmap should be a living document, with short timeframes and frequent adjustments to accommodate changing priorities and market opportunities. This version of the document can serve as an initial document and should base on user epics and stories identified in D11.1. Otherwise, the agile development refers to any development process that is aligned with the concepts of the Agile Manifesto[1] (Agile, 2018). While agile development may develop with time, the RDP's development process should follow best approaches and practices at the time of actual development.

In Section 2, we provide a brief review of FAIR, Open Science and Data provenance principles to be followed by the RDP design. Section 3 discusses the RDP architectural model, a protocol for accessing the RDP and a roadmap for the agile development of the RDP. In Section 4, we conclude the deliverable providing some implication for the final RICHFIELDS roadmap document (D13.5).

2 Review of requirements

A review of requirements published in other RICHFIELDS deliverables was performed in order to identify principles of which good practices should be followed by the design:

1. FAIR
2. Open Science
3. Data Provenance

In the following subsections, each of the principles is discussed to identify their common and specific requirements to be considered by RICHFIELDS. In the first subsection on FAIR principles, basic definitions of data, metadata and provenance data are provided. While the FAIR principles put an emphasis on enhancing the ability of machines to automatically find and use scholarly digital research objects, Open Science aims to aggregate the research content from many different scholarly resources, which requires research data to be cleaned and enriched with metadata. Data provenance is the ability to record the derivation history of data and its place of origin.

In this roadmap, we suggest that good practices of these principles are followed as FAIR defines how data objects should be organized, Open Science gives directions on the definition of metadata, and Data provenance elaborates inputs, entities, systems, and processes (including transformations) that influence any piece of information in the RICHFIELDS data and knowledge repositories.

2.1 FAIR principles

1 The Manifesto was developed by a group of fourteen leading figures in the software industry, and reflects their experience of what approaches do and do not work for software development.

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

The FAIR guiding principles - Findability, Accessibility, Interoperability, and Reusability - put specific emphasis on **enhancing the ability of machines to automatically find and use scholary digital research objects, in addition to supporting its reuse by individuals** (Wilkinson, 2016).

Digital research objects are defined by concepts and data objects, which are in RICHFIELDS related to food and nutrition:

- A **concept** is any defined 'unit of thought' to which we refer in our digital formats.
- A **data object** is defined for the purpose of the FAIR principles as an identifiable data item with an *identifier*, *metadata, provenance data* and *data elements* (Figure 3).

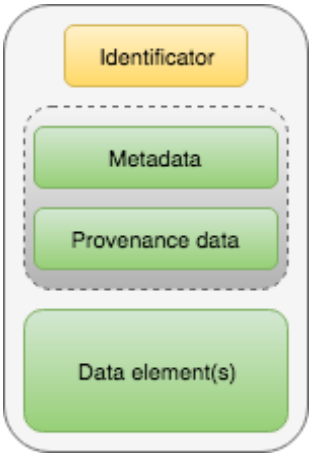


Figure 3. Data object

What is metadata?
Metadata is a set of data that describes and gives information about a data object.

What is provenance data?
Provenance data is another set of data that describes and gives information about entities, activities, and people involved in producing a data object, which can be used to form assessments about its quality, reliability or trustworthiness. An example of provenance data for data elements published in a document is a Digital Object Identifier or DOI. This is a persistent identifier used to uniquely identify objects, standardized by the International Organization for Standardization (ISO). DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports and data sets, and official publications though they also have been used to identify other types of information resources, such as commercial videos.

What are data elements?
Data elements are defined as the actual data that can be either raw data or aggregated data. There are two options to define a FAIR data object:

i. Multiple identifiable data elements can share the same metadata, provenance data and identifier, and form one FAIR data object (for instance a set of images).

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

ii. Individual identifiable data elements can be separately used, integrated, cited and distributed as new FAIR data objects with a new identifier, and carrying sufficient metadata from the original data elements to be traceable back to them (for instance one image from a larger set of images).

If data elements contain *microdata* obtained from some survey, they need to be maintained in such a way that the confidentiality of the information provided by survey participants is preserved. Microdata provided by information systems are under responsibility of those systems, while microdata provided by individuals and stored by the RDP need to be secured by the RDP. However, RICHFIELDS need to have arrangements with other providers that they meet the ethical standards that have been set down by RICHFIELDS. How RICHFIELDS should deal with microdata is described in D13.2.

How should the FAIR principles be implemented?
Although the FAIR principles are true for metadata/provenance data as well as for the actual, collected data elements in the data object, FAIR principles can be independently implemented for each of them.

Moreover, the FAIR principles apply not only to data objects in the conventional sense, but also to the algorithms, tools, and workflows that led to that data.

In the following subsections, each of the FAIR principles is described in more details. By metadata we mean both meta- and provenance data (as requested by R1.2). More information about the FAIR principles is provided by the GO FAIR Initiative (GO FAIR, 2018).

2.1.1 Findability

This FAIR facet relates to the findability of data objects at any point of time, which requires data objects to be uniquely and persistently identifiable:

F1. (Meta)data is assigned a globally unique and persistent identifier.
F2. Data is described with rich metadata (defined by R1 below).
F3. Metadata clearly and explicitly includes the identifier of the data it describes.
F4. (Meta)data is registered or indexed in a searchable resource.

Search engines usually index metadata content, supporting findability. However, the RDP should also explore ways of implementing a general search engine that supports searching of both, data with structured metadata as well as unstructured-metadata and metadata-free data.

2.1.2 Accessibility

This facet relates to the accessibility of data objects, meaning that data objects can be always obtained by machines and individuals:

A1. (Meta)data is retrievable by their identifier using a standardized communications protocol.
A1.1 The protocol is open, free, and universally implementable.

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

A1.2 The protocol allows for an authentication and authorization procedure, where necessary.
A2. Metadata is accessible, even when the data is no longer available.

2.1.3 Interoperability

This facet relates to the interoperability of data objects. It encourages the science and business communities to adopt standardised file formats, metadata, vocabularies and identifiers:

I1. (Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data uses vocabularies that follow FAIR principles.
I3. (Meta)data includes qualified references to other (meta)data.

Specification to be performed by RICHFIELDS:

- Consider Interoperability Service Quality Indicators aligned with FAIR Principles (https://identifiers.org) and ELIXIR Core Data Resource indicators.

2.1.4 Reusability

This facet relates to the reusability of data objects:

R1. (Meta)data is richly described with a plurality of accurate and relevant attributes.
R1.1. (Meta)data is released with a clear and accessible data usage license.
R1.2. (Meta)data is associated with detailed provenance.
R1.3. (Meta)data meets domain-relevant community standards.
R1.4. Lifecycle monitoring/logging of the (meta)data that can also be managed/monitored/covered in the Provenance

2.2 Open Science principles

Open Science is aimed to **aggregate the research content from many different scholarly resources, which requires research data to be cleaned and enriched with metadata**, in order to make research outcomes presented in the right research context.

2.2.1 OpenAIRE

OpenAIRE is an example of a European open-access infrastructure integrating and linking entities from a wide range of scholarly resources (openaire.eu). When a research content is harvested by OpenAIRE, it automatically complies with the EC's policies on Open Access.

OpenAIRE collects metadata from a variety of data sources: publication repositories, data archives and Current Research Information Systems (CRIS) across Europe and beyond. It has defined a series of Interoperability Guidelines to assist data providers expose their content in an interoperable format:

- **Publication repositories:** use of vocabularies, and how to provide links to funding and other research entities (e.g., datasets), and access rights.

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

- **Data archives:** based on the latest DataCite Metadata Schema, extended with information on links to funding and access rights. The metadata from data archives should be included in the OpenAIRE information space, and exposed when data are related to an open-access publication (e.g., a dataset cited by an article). OpenAIRE uses the OAI-PMH v2.0 protocol for harvesting dataset metadata (Elbaek and Nielsen, 2013).
- **CRIS metadata systems:** specify the interoperability layer between CRIS and the OpenAIRE infrastructure. The information interchange is based on the Common European Research Information Format[2] (CERIF) data model, the CERIF XML exchange format, and the OAI-PMH protocol. The guidelines are intended mainly for implementers and administrators of CRIS who plan to communicate research information to OpenAIRE.

In Table 1, metadata for data archives to be included in the OpenAIRE information space are presented (OpenAIRE Guidelines for Data Archives, 2018).

Table 1. The OpenAIRE metadata

Metadata	Description
Identifier[M]	The Identifier is a unique string that identifies a resource, e.g., ARK, DOI, Handle, PURL, URN, URL, etc. (occurrences: 1).
Creator[M]	The main researchers involved in producing the data, or the authors of the publication, in priority order (occurrences: 1-n).
Title[M]	A name or title by which a resource is known (occurrences: 1-n).
Publisher[M]	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role (occurrences: 1).
PublicationYear[M]	The year when the data was or will be made publicly available (occurrences: 1).

2 CERIF is a standard data model for research information and a recommendation by the European Union to its Member States (https://ercim-news.ercim.eu/en68/european-scene-qsupport-of-the-research-processq/cerif-the-common-european-research-information-format). The custody of CERIF has been entrusted by the European Union to euroCRIS (eurocris.org), an international not-for-profit organisation dedicated to the interoperability of CRIS.

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

Metadata	Description
Subject[R]	Subject, keyword, classification code, or key phrase describing the resource (occurrences: 0-n).
Contributor[MA/O]	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource (occurrences: 0-n).
Date[M]	Different dates relevant to the work (occurrences: 1-n).
Language[R]	The primary language of the resource (occurrences: 0-1).
ResourceType[R]	A description of the resource (occurrences: 0-1).
AlternateIdentifier[O]	An identifier or identifiers other than the primary Identifier applied to the resource being registered. This may be any alphanumeric string which is unique within its domain of issue. May be used for local identifiers. AlternateIdentifier should be used for another identifier of the same instance (same location, same file) (occurrences: 0-n).
RelatedIdentifier[MA]	Identifiers of related resources. These must be globally unique identifiers (occurrences: 0-n).
Size[O]	Unstructured size information about the resource (occurrences: 0-n).
Format[O]	Technical format of the resource (occurrences: 0-n).
Version[O]	The version number of the resource (occurrences: 0-1).
Rights[MA]	Any rights information for this resource (occurrences: 0-n).
Description[MA]	All additional information that does not fit in any of the other categories. May be used for technical information (occurrences: 0-n).
GeoLocation[O]	Spatial region or named place where the data was gathered or about which the data is focused (occurrences: 0-n).

Mandatory (M): the field must always be present in the metadata record. An empty element is not allowed.
Mandatory when applicable (MA): when the value of the field can be obtained, it must be present in the metadata record.
Recommended (R): the use of the field is recommended.
Optional (O): the property may be used to provide complementary information about the resource

2.3 Data Provenance

Data provenance (also referred to as *data lineage*) is the ability to **record the derivation history of data and its place of origin**. In the context of data provenance, provenance needs

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

to document the inputs, entities, systems, and processes (including transformations) that influence any piece of information in the RICHFIELDS data and knowledge repositories.

The generated evidence supports essential forensic activities such as data-dependency analysis, error/compromise detection and recovery, and auditing and compliance analysis, including the ability to detect advanced/persistent threats. Data provenance is necessary to guarantee the quality, usability and security of your data. Knowledge of data provenance is also relevant from the perspective of regulatory mechanisms to protect intellectual property.

Provenance assertions have a form of contextual metadata and can themselves become important records with their own provenance.

Data provenance is relevant for academic and research organizations, as well as for business establishments. In scientific fields, scientific publications are a common form of representing the provenance of research data and results. Increasingly, Digital Object Identifiers (DOIs, 2018) are used to cite data used in experiments and case studies so that the papers can relate the experimental or case study process and analysis – which form the data provenance – to the actual data used and produced.

2.3.1 Data provenance semantics

There exist several **data provenance vocabularies** that may be of RICHFIELDS interest because food- and nutrition-related data applies to the whole food chain:

1. Metadata for interoperability in geographic dataset search and access (ISO TC 211)
    - ISO 19115-1 Collection and Granule Metadata for geographic information (MD package)
    - ISO 19115-2 NOAA-NASA sponsored revision for imagery and gridded data (MI package)
    - ISO 19110 Geographic information for feature entity and attribute cataloguing
    - ISO 19119 Metadata for services
    - ISO 19157 How to describe, evaluate and report data quality
    - ANSI Content and Collection Management Standards and FGDC Digital Geospatial Metadata (ISO 19115 NAP) standards feed into ISO
2. Open Geospatial Consortium (OGC) Service Web Enablement (SWE)
    - Sensor Observation Service and WCS, WFS, WMS serving netCDF and HDF
    - sensorML (instrument specifications)
3. Consultive Committee on Space Data Systems (CCSDS)
    - Open Archive Information Systems (OAIS) Reference Framework (nomenclature)
    - Satellite data requirements (i.e., Level 0 data content), XML Formatted Data Unit
4. Library and Publishing applications (e.g., NISO, NARA, Library of Congress)
    - Dublin Core

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

- Metadata Exchange and Transfer Standard (https://www.w3.org/TR/ws-metadata-exchange/ (accessed on 12th July, 2018)).)
- PREMIS – preservation data (https://www.loc.gov/standards/premis/ (accessed on 12th July, 2018)).)

What is the Dublin Core Metadata Initiative (DCMI)?
This is an open organization engaged in the development of interoperable online metadata standards. It has specified the **DCMI Abstract Model** (http://dublincore.org/documents/abstract-model/ (accessed on 12th July, 2018)).) (Figure 4), which is built on the work undertaken by the World Wide Web Consortium (W3C) on the Resource Description Framework (RDF) (https://www.w3.org/RDF/ (accessed on 12th July, 2018)). This is related to data provenance data.

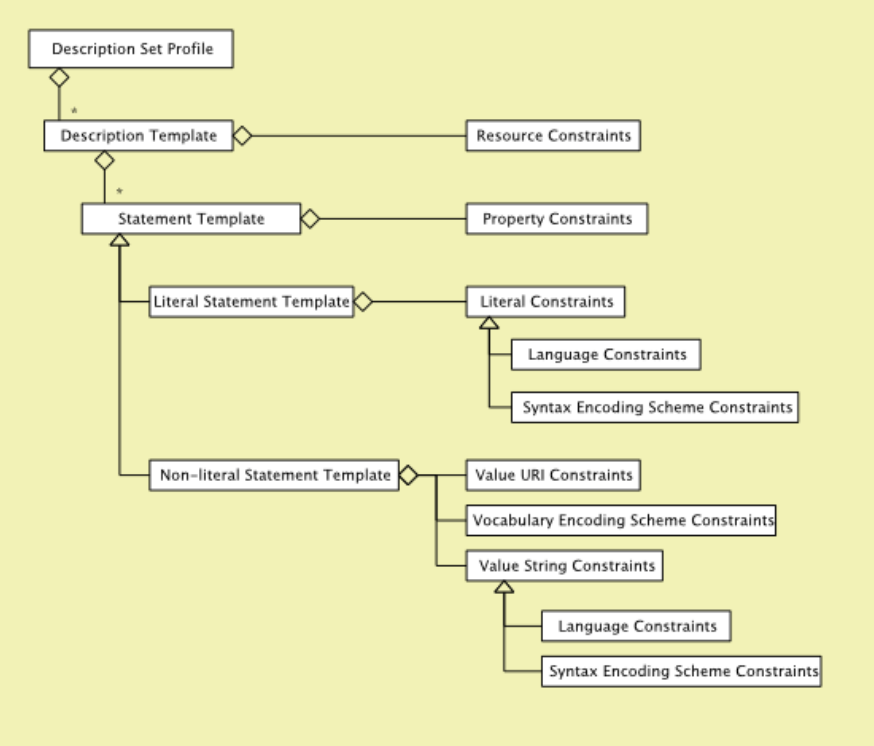


Figure 4. The DCMI Abstract Model

The DCMI abstract model of the resources described by metadata descriptions is as follows (http://dublincore.org/documents/abstract-model/):

- Each described resource is described using one or more property-value pairs.
- Each property-value pair is made up of one property and one value.
- Each value is a resource (i.e., the physical, digital or conceptual entity or literal) that is associated with a property when a property-value pair is used to describe a resource. Therefore, each value is either a literal value or a non-literal value, i.e., a physical, digital or conceptual entity.

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

- A literal is an entity, which uses a Unicode string as a lexical form, together with an optional language tag or datatype, to denote a resource (i.e., 'literal' as defined by RDF).

In the DCMI abstract model, each resource is described by a metadata description set, which may also contain a description about the description set itself (sometimes referred to as 'meta-metadata'). The abstract model of DC metadata description sets is as follows:

- A description set is a set of one or more descriptions, each of which describes a single resource.
- A description is made up of one or more statements (about one, and only one, resource) and zero or one described resource URI (an Uniform Resource Identifier that identifies the described resource).
- Each statement instantiates a property-value pair that is made up of a property URI (a URI that identifies a property) and a value surrogate.
- A value surrogate is either a literal value surrogate or a non-literal value surrogate. A literal value surrogate is a value surrogate for a literal value that is made up of exactly one value string. The value string is a literal, which encodes the literal value. A non-literal value surrogate is a value surrogate for a non-literal value that is made up of zero or one value URI (a URI that identifies the non-literal value associated with the property), zero or one vocabulary encoding scheme URI (a URI that identifies the vocabulary encoding scheme of which the non-literal value is a member), and zero or more value strings. Each value string is a literal, which represents the non-literal value. A value string is either a plain value string or a typed value string. A plain value string may have an associated value string language that is an ISO language tag (for example en-GB). Plain value strings are intended to be human-readable. A typed value string has an associated syntax encoding scheme URI that identifies a syntax encoding scheme.

The DC metadata element set is a vocabulary of fifteen properties for use in resource description:

Term Name (label): contributor	
URI:	http://purl.org/dc/elements/1.1/contributor (accessed on 12th July, 2018)
Definition:	An entity responsible for making contributions to the resource.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Term Name (label): coverage	
URI:	http://purl.org/dc/elements/1.1/coverage (accessed on 12th July, 2018)
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.

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

Comment:	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
References:	http://www.getty.edu/research/tools/vocabulary/tgn/index.html (accessed on 12th July, 2018)
Term Name (label): creator	
URI:	http://purl.org/dc/elements/1.1/creator (accessed on 12th July, 2018)
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Term Name (label): date	
URI:	http://purl.org/dc/elements/1.1/date (accessed on 12th July, 2018)
Definition:	A point or period of time associated with an event in the lifecycle of the resource.
Comment:	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
References:	http://www.w3.org/TR/NOTE-datetime (accessed on 12th July, 2018)
Term Name (label): description	
URI:	http://purl.org/dc/elements/1.1/description (accessed on 12th July, 2018)
Definition:	An account of the resource.
Comment:	Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
Term Name (label): format	
URI:	http://purl.org/dc/elements/1.1/format (accessed on 12th July, 2018)
Definition:	The file format, physical medium, or dimensions of the resource.
Comment:	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].

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

References:	http://www.iana.org/assignments/media-types/ (accessed on 12th July, 2018)
Term Name (label): identifier	
URI:	http://purl.org/dc/elements/1.1/identifier (accessed on 12th July, 2018)
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Term Name (label): language	
URI:	http://purl.org/dc/elements/1.1/language (accessed on 12th July, 2018)
Definition:	A language of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].
References:	http://www.ietf.org/rfc/rfc4646.txt (accessed on 12th July, 2018)
Term Name (label): publisher	
URI:	http://purl.org/dc/elements/1.1/publisher (accessed on 12th July, 2018)
Definition:	An entity responsible for making the resource available.
Comment:	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Term Name (label): relation	
URI:	http://purl.org/dc/elements/1.1/relation (accessed on 12th July, 2018)
Definition:	A related resource.
Comment:	Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term Name (label): rights	
URI:	http://purl.org/dc/elements/1.1/rights (accessed on 12th July, 2018)
Definition:	Information about rights held in and over the resource.
Comment:	Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.
Term Name (label): source	

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

URI:	http://purl.org/dc/elements/1.1/source (accessed on 12th July, 2018)
Definition:	A related resource from which the described resource is derived.
Comment:	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term Name (label): subject	
URI:	http://purl.org/dc/elements/1.1/subject (accessed on 12th July, 2018)
Definition:	The topic of the resource.
Comment:	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary.
Term Name (label): title	
URI:	http://purl.org/dc/elements/1.1/title (accessed on 12th July, 2018)
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.
Term Name (label): type	
URI:	http://purl.org/dc/elements/1.1/type (accessed on 12th July, 2018)
Definition:	The nature or genre of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
References:	http://dublincore.org/documents/dcmi-type-vocabulary/ (accessed on 12th July, 2018)

The formal DCMI abstract model semantics can be defined by reference to the RDF and RDF Schema semantics (https://www.w3.org/TR/rdf11-mt/ (accessed on 12th July, 2018)).

Which data provenance ontologies are of the RICHFIELDS interest?

- PROV-O (https://www.w3.org/TR/prov-o/ (accessed on 12th July, 2018)) is a domain-independent and general-purpose W3C provenance ontology that allows and encourages extensions to cover more specific needs. In particular, to track authoring and versioning information of web resources, PROV-O provides a basic methodology but not any specific classes and properties for identifying or distinguishing between

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

the various roles assumed by agents manipulating digital artefacts, such as author, contributor, and curator.

- PAV ontology: provenance, authoring and versioning (http://purl.org/pav/ (accessed on 12th July, 2018)) is a lightweight ontology for capturing 'just enough' descriptions essential for tracking the provenance, authoring and versioning of web resources (https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-4-37 (accessed on 12th July, 2018)). PAV distinguishes between contributors, authors and curators of content and creators of representations in addition to the provenance of originating resources that have been accessed, transformed and consumed. PAV extends the W3C PROV-O ontology to support broader interoperability.
- P-PLAN (http://www.opmw.org/model/p-plan17092013/ (accessed on 12th July, 2018)) is an extension of the PROV-O ontology created to represent the plans that guided the execution of scientific processes. P-PLAN describes how the plans are composed and their correspondence to provenance records that describe the execution itself.
- The Bibliographic ontology (http://purl.org/ontology/bibo/ (accessed on 12th July, 2018)).) provides main concepts and properties for describing citations and bibliographic references (i.e., quotes, books, articles, etc) on the Semantic Web.
- The DataCite ontology (http://www.sparontologies.net/ontologies/datacite (accessed on 12th July, 2018)) associates identifiers to a bibliographic entity (e.g., a dataset, a person, an article) specifying their exact nature by means of the object property. In addition, it is also possible, through the object property, to link an entity to another item representing an entity description of a particular type. In this way, it is possible to associate written documents (e.g., journal articles) as descriptions of datasets.

2.3.2 Data provenance management

Provenance information is focused on two concepts: the ancestral data and the transformations performed to produce that data. It is convenient to conceptualize the provenance information in the form of a graph, with nodes representing the piece of information and edges representing the transformation processes.

There exist several graph-based approaches to solve the data provenance challenges (like open source Noe4j - https://neo4j.com, Ontotext GraphDB – RDF Triplestore - http://www.linkedlifedata.com/about) and graph visualization solutions (like Linkurious).

3 Data architecture

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

The FAIR Facets - Findability, Accessibility, Interoperability, and Reusability - are related, but technically independent from each other and **may be implemented in any combination**. In this section, we will define the main characteristics, norms, and practices that RICHFIELDS data resources, tools, and infrastructures should exhibit in order to be considered 'FAIR' as well as follow the principles of Open Science and Data provenance.

3.1 The model of the RICHFIELDS Data Platform

In Figure 5 below, the main model of the RICHFIELDS Data Platform (RDP) is provided.

What is a computer model?
In general, a computer model defines important *features of events* that may happen in the system. Features that can change - or vary - are known as variables. In RICHFIELDS, the following events (ordered by the development stages) may happen in the RDP:

- **Core offering:**
  - An information system connects with the RDP, in order to provide or use data objects.
  - An information system requires access to one or more specific data sets.
  - An individual explores the RICHFIELDS data catalogues published by the RDP to find one or more data sets of his (her) interest.
  - An information system's manager or a RICHFIELDS manager explores statistics about web services (s)he is responsible for.
- **Mature:**
  - An information system requires access to one or more specific data sets in a harmonized way.
  - An information system requires access to the RICHFIELDS taxonomy and ontology.
  - An individual explores the RICHFIELDS knowledge catalogues published by the RDP to find one or more data sets of his (her) interest.
  - An individual contacts a RICHFIELDS manager, in order to supply (donate) data.
  - An individual contacts a RICHFIELDS manager, in order to get access to one or more data sets.
- **Growth:**
  - An individual explores the information about RICHFIELDS taxonomy and ontology published by the RDP to find one or more data sets of his (her) interest.
  - An individual contacts a RICHFIELDS manager, in order to get access to one or more data sets in a harmonized way.

Next, the computer model needs to specify *rules* that control these features and their relationships. In RICHFIELDS, these rules need to be in accordance with business and legal/ethical rules specified in WP12 and WP13.

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

The computer model also needs algorithms and data. *Algorithms* are sets of instructions, which tell the computer how to make decisions and when to do calculations. These algorithms are specified in D11.2. *Data* are facts and statistics about something – in our case data includes data objects as defined in Section 2.1.

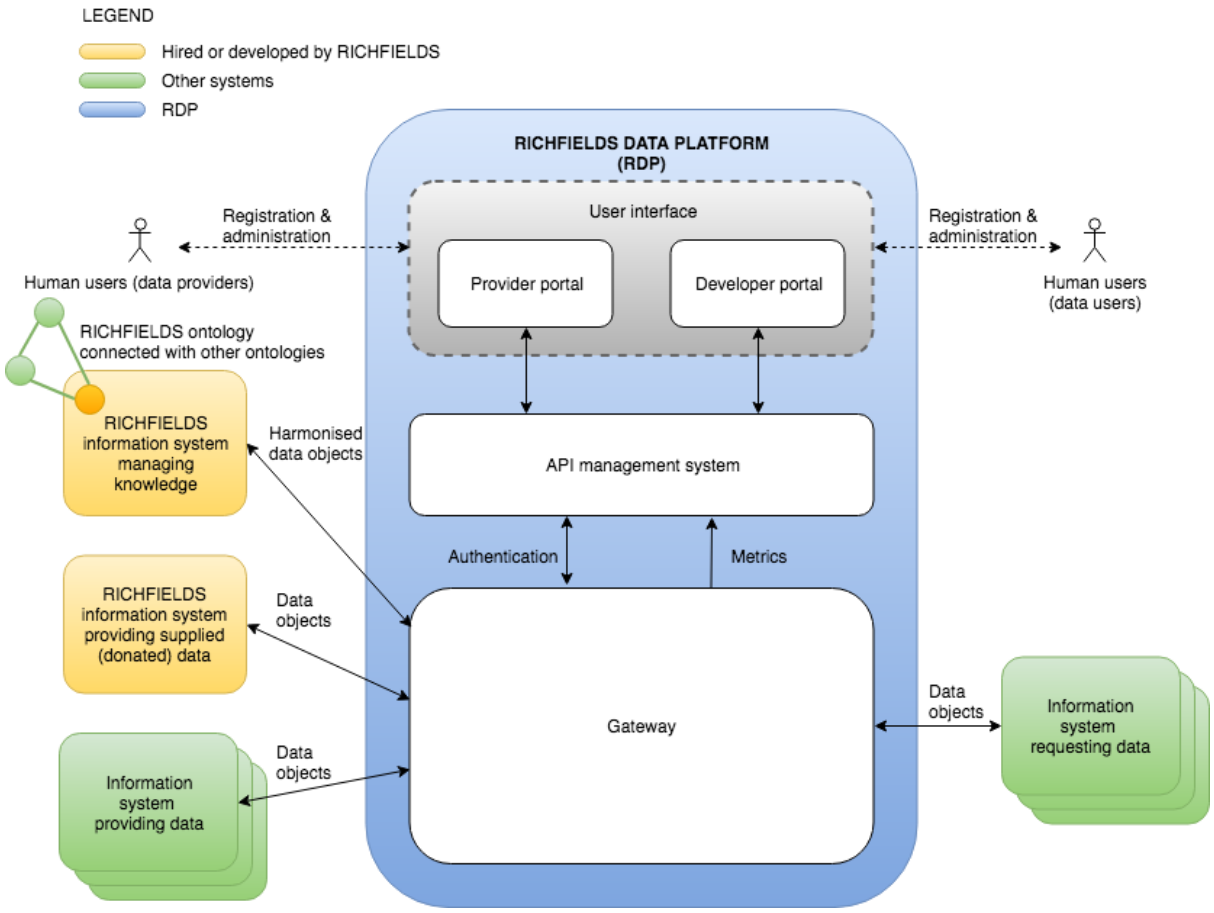


Figure 5. The RICHFIELDS Data Platform

3.1.1 The RICHFIELDS Data Platform

The main entrance point to the RICHFIELDS Data Platform is a **user interface**, where up-to-date public information about RICHFIELDS data and knowledge repositories, research protocols, taxonomy, and ontology is provided. The user interface is visually integrated with the public front-end website and provides access to portals for data providers and developers. *Data providers* are managers of information systems that provide data to the RICHFIELDS data platform via web services. *Developers* are managers of information systems that acquire data provided by the RICHFIELDS data platform. Other users, such as *researchers* that donate data (raw or aggregated) to be stored by the RICHFIELDS data platform, will not be able to directly access the RICHFIELDS Data Platform. They will contact

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

RICHFIELDS managers who have a role of 'data providers' supporting researchers and other donators of data.

The **RICHFIELDS provider portal** is the entrance point for data providers, where each data provider can manage his (her) web services, documentation, and web service users, as well as explore statistical data about the usage of his (her) web services.

The **RICHFIELDS developer portal** is the entrance point for developers. It allows developers to create an account and apply for one or more web services. Developers can also see their own statistics and read documentation about applied web services.

The RICHFIELDS developer and provider portals are connected with an **API management system**, which is the central part of the RICHFIELDS Data Platform. This system contains all administrative functionality that is needed to manage APIs. It needs to deliver scalable solutions to package, distribute, control, and monetize APIs. Built-in features need to help developers build more successful applications, including their access control, rate limits, payment gateway integration, and developer experience tools.

The **RICHFIELDS gateway** is a separate application server that actually transfers data and knowledge to the actual web service (and the other way around). The API management system does not ever 'see' data elements passing the gateway server, but only metadata (who makes which call) and associated metrics.

3.1.2 The RICHFIELDS input units

In Figure 5, **information systems providing data,** which are depicted on the left side of the figure, presents systems providing food- and nutrition-related data to the RICHFIELDS Data Platform via web services. These can be either:

- scientific cloud, such as European Open Science Cloud, Zenodo, FigShare, etc.,
- server platform, such as Quisper, EuroFIR, GS1 GDSN, EFSA, MetroFood, CORBEL, etc., or
- application server, such as RICHFIELDS*, PRECIOUS, FitBit, Twitter, Facebook, etc., or
- specific application server analysing big data, e.g., RICHFIELDS, or
- food lab, or
- IoT platform, or
- national 'RICHFIELDS' node.

*RICHFIELDS may develop its own mobile app to collect consumer data generated by various apps.

As each of these information systems may use specific web services, RICHFIELDS will need to support the implementation of all of them.

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

Data providers will also need to provide metadata and provenance data as defined by RICHFIELDS. This information should be included through the web service providing data elements.

The **RICHFIELDS information system providing data supplied by individuals** is aimed to store data supplied (donated) by researchers who cannot provide data via web services. Those data, once stored by RICHFIELDS data managers in this information system, will be treated in the same way as other information systems providing data. RICHFIELDS will define its own web services for handling this data. The specification of these web services could also be offered to data providers that have not yet developed their own web services for delivering data.

The **RICHFIELDS information system managing knowledge** is a system that collects and stores metadata and provenance data about all datasets provided through the RICHFIELDS Data Platform. Moreover, it performs the semi-automatic creation of the RICHFIELDS ontology and the corresponding harmonization web service. When connecting a new web service to the RDP, the RICHFIELDS taxonomy and ontology will adapt to new concepts, terms, and relations between them, so that the harmonization of web services can be performed. More details about data harmonization with examples is provided in Section 3.3.

3.1.3 The RICHFIELDS output units

In Figure 5, **information systems using data**, which are depicted on the right side of the figure, presents systems accessing data via provided web services. If they decide to access selected data objects via web services defined by the data providers, they will need to implement all of them. Another possibility is to access selected dataset via the RICHFIELDS harmonization web service.

3.2 Protocol for accessing the RDP

1. Information system (IS)
    a. As data provider
        1. The IS manager checks the conditions on the public RICHFIELDS website (requirements for providing API, which metadata and provenance data needs to be provided, documentation, business model, etc.).
        2. If needed, the IS manager can contact the RICHFIELDS manager whose contact can be found on the RICHFIELDS website.
        3. The IS manager registers in the RDP and agrees to the RICHFIELDS conditions (by signing a contract and/or accepting RICHFIELDS terms and conditions, which can be done via the RICHFIELDS website or manually).

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

4. The IS manager enters the provider portal and registers their APIs. For each API, (s)he setups monetization policy, provides required metadata and provenance data as well as the documentation.
5. Through the developer portal, the IS manager can track statistical and monetization data (i.e., earned money) about the usage of his (her) web services.
6. IS manager can deregister while he is signed in and no obligations towards the RDP exist. With the deregistration, all user data is deleted (registration information, created APIs, etc.) and all depended services (e.g., harmonization) are updated.

b. As data user

1. The IS manager checks the conditions on the public RICHFIELDS website (requirements for using APIs, which metadata and provenance data are available for each API, documentation, business model, etc.).
2. If needed, the IS manager can contact the RICHFIELDS manager whose contact can be found on the RICHFIELDS website.
3. The IS manager registers in the RDP and agrees on the RICHFIELDS conditions (by signing a contract and/or accepting RICHFIELDS terms and conditions, which can be done via the RICHFIELDS website or manually).
4. The IS manager enters the developer portal, registers their application, and requests for selected APIs. For each API, (s)he also checks metadata and provenance data as well as the provided documentation. (S)he can also opt for the RICHFIELDS harmonization API, which enables a harmonized access to requested data.
5. Through the developer portal, the IS manager can track statistical and monetization data (i.e., spent money) about the usage of his (her) applications.
6. IS manager can deregister while he is signed in and no obligations towards the RDP and web service providers exist. With the deregistration, all user data is deleted (registration information, created application profiles, etc.).

2. Individual

a. As data provider

1. The individual checks the conditions on the public RICHFIELDS website (existing data and knowledge catalogues, requirements for

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

providing data, which metadata and provenance data needs to be provided for each dataset, documentation, business model, etc.).

2. The individual contacts the RICHFIELDS manager whose contact can be found on the RICHFIELDS website.
3. The individual temporarily registers in the RDP and agrees on the RICHFIELDS conditions (by signing a contract and/or accepting RICHFIELDS terms and conditions, which can be done via the RICHFIELDS website or manually).
4. The individual enters the provider portal, uploads the dataset, provides metadata and provenance data as well as the requested documentation.
5. The individual relinquishes the data to the usage of RDP, either free of charge or with one-time money compensation. If any kind of monetization is required/demanded by the individual, (s)he should contact some IS provider manager to support him (her) with IS support (effectively becoming an IS provider, so such rules applies).
6. The uploaded data is then managed by the RICHFIELDS manager within the RICHFIELDS IS.
7. The individual can deregister while he is signed in, at any time. With it registration information is deleted and relinquished data stays in RDP as part of RICHFIELDS IS.

b. As data user

1. The same principles apply as with IS data user. In case a help is needed with regard to connecting to selected web services, it is expected that support is given by RICHFIELDS manager (e.g., basic support clients could be provided by IS provider for each of the web services) or (s)he can contact any of the IS users for support.

3.3 Data harmonization

As methods for data linkage and harmonization are already described in D11.2, we provide here only a brief explanation about the semantic interoperability concept.

Semantic interoperability involves two basic steps:

1. An ontology, that describes the domain of interest, needs to be selected or newly developed.
2. Collected data needs to be enriched with tags from the ontology, which are machine-processable data pieces (metadata).

In D11.2, the RICHFIELDS ontology based on the RICHFIELDS taxonomy is already presented. We have suggested to create it in a dynamic way, performing semi-automatic adaptation

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

whenever a new data set indexed and classified on its own way is connected with the RDP. Semi-automatic means that a human expert is required to check and confirm machine-generated relations between terms and concepts used in the data set with terms and concepts from other RICHFIELDS data sets.

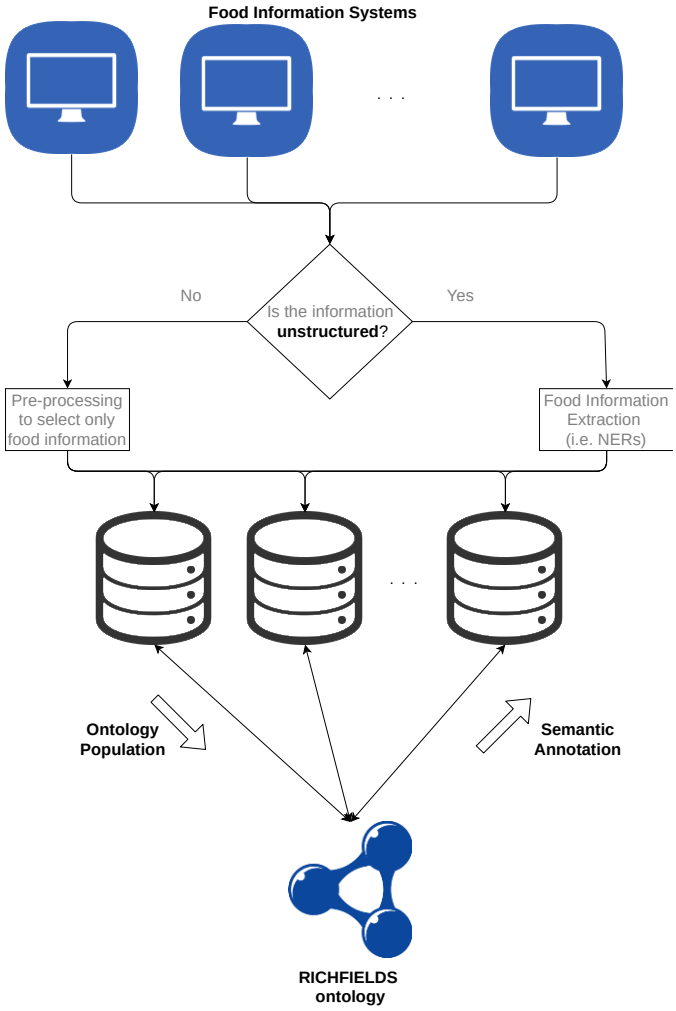


Figure 6. The RICHFIELDS semantic interoperability.

The second basic step (data enrichment with ontology tags) requires the following actions, which are also depicted in Figure 6:

- Apply *data pre-processing* with regard to the data type to obtain structured data. In case of structured or semi-structured data, specific rules based on regular expressions could be applied. If the data is unstructured, an information extraction method (e.g., named-entity recognition - NER) could be applied to extract and structure the relevant information.
- Link the structured data to concepts that exist in the ontology using a *similarity measure*. The data linkage is performed either by semantic annotation or by ontology population. First, a threshold value for the similarity measure is defined

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

(e.g., 0.125). Then, if the value of the similarity measure is greater or equal than the selected threshold, a *semantic annotation* is performed. It means that the matching is a good one, i.e., a concept that was searched for in the ontology and that has already existed, and should be annotated with the metadata, so the data set is enriched by including tags from the ontology. Otherwise, if the value of the similarity measure is lower than the selected threshold, the data cannot be linked to the ontology because such a concept for which the ontology is searched for does not exist yet, and first *ontology population* needs to be performed, i.e., an instance for such a concept needs be added to the ontology.

Selecting a good information extraction method is not an easy task. Methods that are good for a specific domain can be inefficient for other domains. In D11.2, we describe the drNER method (Eftimov et al, 2017a) that is a combination of a terminological-driven NER and rule-based NER. The difference with purely terminological-driven NERs is that we do not only use dictionaries with concepts and synonyms (as terminological resources), but we allow for reuse of some corpus-based NERs that exist for some entities. If corpus-based NERs exist for some entities that we are interested in, we use them to annotate text data and then to see if some tokens have labels that correspond to entities of interest. We also combine corpus-based NERs that exist for some entities in which we are interested, following the idea of ensemble learning in order to achieve better performance than the performance obtained from any corpus-based NER alone. The difference with the rule-based NERs is that we do not use rules associated with the characteristics of the entities. This is because having rules for each of the entities we are interested in requires too much time and effort to produce them. We only use a small number of Boolean algebra rules that are not related to the characteristics of the entities, but help us define the phrases that are the entities mentions. Evaluation of the method showed that the method gives promising results in the domain of the RICHFIELDS interest (Food Intake).

Finding a good similarity measure is another challenge since this is also dependent on the domain of interest. In case of unstructured data presented as text, text normalisation methods that involve text similarity measures can be applied. This is valid also for numerical data, which are described with metadata that is textual data. In D11.2, we describe the a text normalisation method, named StandFood, which can be used for standardization of foods according to FoodEx2 that is a comprehensive food classification and description system for exposure assessment introduced by EFSA (Eftimov et al., 2017b). StandFood is a semi-automatic system for classifying and describing food concepts, and consists of three parts: the first classifies the food concept into one of four FoodEx2 categories (i.e., raw, derivatives, simple composite, and aggregated composite) using ensemble of classifiers, the second describes the food concept using the FoodEx2 code using natural language processing approach, and the third combines the results from the first and the second part to improve classification by defining post-processing rules. As a similarity measure, we use the description part of the StandFood. A preprocessed data set containing food concepts is linked to the RICHFIELDS ontology so that each food concept from the data set is linked to

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

food individuals from the RICHFIELDS ontology using a POS-tagging probability weighted method.

What is the flow of (meta)data in the RDP?

When an information system connects with the RDP, in order to provide an access to a data set, its manager uses the provider portal to i) describe the corresponding web service, ii) upload the requested documentation about the web service, and iii) set up the monetisation rules. Each data set needs to include beside data element(s) also metadata and provenance data as defined by RICHFIELDS. Provenance data and some metadata needs to be provided directly by the information system, either automatically as part of the web service or manually by the manager, who can enter this information using the provider portal. Other metadata needed for data harmonization are generated automatically by the RICHFIELDS framework for semantic interoperability (as described in the above section). While data objects provided by information systems are managed (stored) by the information systems themselves, metadata and provenance data would be managed (stored) also by the RICHFIELDS information system managing knowledge (Figure 5) because they are needed for harmonization.

Some real-life (practical) examples

Whenever a new data set is included in the RICHFIELDS data repository, data objects from the data set will be acquired by the RICHFIELDS information system providing knowledge to be described by the RICHFIELDS ontology tags (metadata). In Figure 7, an example of the food instance 'Nectarine' acquired from the PRECIOUS data set and described by the RICHFIELDS_tag '#A01GN' is presented.

```
{
        _id: ObjectId(**),
        key: FOOD_INTAKE,
        from: ISODate(2017-09-01T21:54:39.463Z),
        to: ISODate(2017-09-01T21:54:39.464Z),
        value: [{
                food_amount: 80,
                food_name: Nectarina,
                food_type: 5,
                RICHFIELDS_tag:http://www.semanticweb.org/
tome/ontologies/2018/2/Richfields#A01GN"
        }],
        id: 1470088479463,
        user: ObjectId(ffffffffffffffffffffffff),
        posted: ISODate(2017-09-02T09:06:33.484Z),
        __v: 0
}
```

Figure 7. An example of annotated food instance from the PRECIOUS data set.

This example is obtained from a case-study performed in D9.2, where the proposed RICHFIELDS framework for semantic interoperability was investigated. In this case, we decided to describe each food intake data element by EFSA FoodEx2 descriptors (using the StandFood method described above). In this way, all food intake data elements will be described (harmonized) using the same indexing and classification system (FoodEx2). The same approach can be taken for any other system for indexing and classification.

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

Another example is given on unstructured data, where the following text extracted from a web site (https://www.thekitchn.com/whats-the-difference-between-peaches-and-nectarines-221699) was annotated by RICHFIELDS ontology tags:
"The main physical difference is that peaches (Richfields#A01GM) have a fuzzy coating, whereas nectarines (Richfields#A01GN) are smooth and do not have this coating. They are almost identical genetically, but there is a gene variant between the two."
Any unstructured data (e.g., twits, scientific articles, etc.) can be annotated in the same way.

Similarly, ontology tags are provided for data provenance as the RICHFIELDS ontology can be linked with a data provenance ontology as well.

The examples above are given for the Food intake domain, but the approach is independent from the domain and can be used also for other domains (e.g., Determinants). The only requirement is that the most appropriate indexing and classification system is selected (or newly defined by RICHFIELDS) to describe the domain's concepts and entities. When an information system requests data in a harmonized way, the RICHFIELDS harmonization web service will provide i) data objects enriched not only with metadata and provenance data originally provided by the data provider but also with RICHFIELDS ontology tags, and ii) information about the RICHFIELDS semantic model (ontology).

3.4 Roadmap for the development of the RDP

In the following subsections, we present a roadmap for launching the RDP as part of the RICHFIELDS Minimum Viable Product (MVP) and endorsing it as a FAIRport. The roadmap is defined by considering the user needs identified and described in D11.1 through user epics and stories. The final roadmap for launching the RDP also from business, legal and ethical, and governance aspects is presented in the final WP13 deliverable.

3.4.1 Launch the RDP as part of the Minimum Viable Product

We propose the following basic steps to be performed:

1. Work organisation.
2. Definition of standards, protocols, and polices.
3. Design and modelling.
4. Technical specifications.
5. Implementation of the RICHFIELDS MVP data platform.
6. Testing, validation and launching of the RICHFIELDS MVP data platform.

Although the steps are listed in a waterfall style, they are aimed to be performed in the agile way. For each step, we defined stage(s) in which the step needs to be established. In any stage, all the steps (activities) already established in previous stages need to be improved. The abbreviation CO means Core offering, G stands for Growth, and M means Maturity.

Work organization (restricted to the technical part. Other parts are covered by final deliverables of WP12 and WP13):

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

- (CO) Recruit a team of RDP developers and split it into four groups of i) managers (team manager, scrum master, functional manager, application manager, data manager, release manager, etc.), ii) designers (solution architect, data scientist, designer, HCI expert, etc.), iii) developers and iv) testers/validators (internal and external). Assign responsibilities.
- (CO) Agree on working and communication protocols and environments. Consider the agile development concepts and practices (e.g., SCRUM).
- (CO) Define a risk management plan.

Definition of standards, protocols and polices:

- (CO) Define a **standard for the description of RICHFIELDS data objects** (in machine readable format) to enable attribute-based search, linkage, harmonization, analysis and delivery – from the content and the format views. Consider established metadata / data provenance schemata. This step needs to be performed in close collaboration with the RICHFIELDS Legal/Ethics board.
- (CO) Define a format of the API documentation (for administrators, developers and providers).
- (CO) Define a usability testing plan (User Acceptance Test or UAT).
- (CO) Define security measures.
- (CO) Define quality measures. Consider Interoperability Service Quality Indicators aligned with FAIR Principles (https://identifiers.org) and ELIXIR Core Data Resource indicators.
- (CO) Define a validation plan and success factors (considering user epics and stories tests).
- (CO) Define **monetisation policies**. Define appropriate licensing of each data provider's web service. This step needs to be performed in close collaboration with the RICHFIELDS Business board.

Design and modelling:

- (CO) Design a flexible data model of RICHFIELDS data objects – to specify all the data object's elements (data elements – most probably to be stored in a relational database, and metadata/provenance data – to be stored in NoSQL, graph database, etc.).
- (CO) Design a database management system for managing donated data in the RICHFIELDS information system providing donated data.
- (CO) Design a knowledge base management system for managing metadata, provenance data, and the RICHFIELDS taxonomy and ontology. Consider the FAIR principles and outcomes of other projects (e.g., FREYA).
- (CO) Define **concepts and entities about determinants, food intake, status and health** to be considered by the RICHFIELDS harmonization web service. Consider

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

WP9 deliverables. This step needs to be performed in close collaboration with the RICHFIELDS Scientific board.

- (G) Identify **ontologies (including those for data provenance) to be linked** with the RICHFIELDS ontology and harmonization web service. This step needs to be performed in close collaboration with the RICHFIELDS Legal/Ethics board.
- (CO) Design the RICHFIELDS data platform, specifying hardware, software and communication/network to implement microservices (relational databases, NoSQL/graph databases, event streaming, container orchestration).
- (CO) Design the user interface as part of the RICHFIELDS website, and the provider and developer portals, considering human-computer interactions (HCI) aspects. In the first stages (CO, G) the user interface can be designed as a simple interface, while later (M) it should be an advanced interface.
- (CO) Design the RICHFIELDS website.
- (CO) Design the RICHFIELDS mobile app.

Technical specifications:

- (CO) Specify a priority list of microservices (functions as services) to be developed in each spring (agile methodology). Consider comments collected during the RICHFIELDS Stakeholder meetings. Also take into account services provided by other data infrastructures at the European level (e.g. ELIXIR, EUDAT, etc.).
- (CO) Define API specifications for each web service selected for the RDF MVP and event streams. Consider also the RICHFIELDS API for accessing donated data.
- (CO) Define the RICHFIELDS API for accessing metadata and provenance data as well as to link this information with corresponding data elements.
- (M) Define the **RICHFIELDS API for accessing data in a harmonized way**.

Implementation of the RICHFIELDS MVP data platform:

- (CO) Agree on storage/cloud/knowledge engine/API management system vendors.
- (CO) Implement the RICHFIELDS website.
- (CO) Implement the RICHFIELDS mobile app.
- (CO) Implement the RICHFIELDS MVP data platform (user interface with provider and developer portals, API management system, gateway).
- (CO) Implement APIs for each web service.
- (CO) Implement the RICHFIELDS API for accessing metadata and provenance data as well as to link this information with corresponding data elements.
- (G) Implement the RICHFIELDS API for accessing data in a harmonized way.
- (CO) Implement quality measures.
- (CO) Implement security measures.

Testing, validation and launching of the RICHFIELDS MVP data platform:

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

- (CO) Verification (databases, functionality – use case tests).
- (CO) Usability testing.
- (CO) Security audit.
- (CO) Performance testing and improvements.
- (CO) Beta user testing.
- (CO) External RICHFIELDS website testing.
- (CO) External RICHFIELDS mobile app testing.
- (CO) External RDP MVP validation.
- (CO) Launch the RICHFIELDS website, mobile app and MVP data platform.
- Repeat all the steps using agile methodology.

3.4.2 FAIRness approval

As FAIR is not a trademark, repositories of data objects are usually endorsed as FAIRports by authorities, such as ELIXIR nodes/the Hub, NIH or SciELO. We propose that these authorities are defined as trusted parties for RICHFIELDS as well. The specific steps shall be:

1. Define a selected authority for each semantic category of concepts referred to in RICHFIELDS data objects.
2. Define minimal criteria to qualify data objects as FAIR.
3. Review individual data FAIRports against these established criteria.
4. Give a FAIR [***Trusted party***] stamp of approval to the RICHFIELDS data platform.
5. Publish in Open Repositories (preferably FAIR themselves) what can be expected from the RICHFIELDS data platform in their index and with their quality stamp.

4 Conclusions and implications

In this deliverable, we presented the architectural model of the RDP. One of the most important features of the RDP is that it can provide an easy access to linked and harmonized data on food and nutrition, collected in different formats and from various data sources. Therefore, the deliverable also reports about the FAIR, Open Science and Data provenance policies that define metadata and provenance data as well as approaches for their collection and management. All data flows through the RDP need to be enriched with metadata and provenance data not only to enable data linkage and harmonization, but also to preserve authenticity and provenance of data. We conclude by the roadmap for developing the RDP from the technical point of view. Specific steps and the phase (Core offer, Growth or Maturity), in which they need to be initiated, are identified.

We suggest to design the RDP in such a way that all data objects are technically handled outside the RDP, so that its maintenance can be made as easy, efficient and cost effective as possible. In this way, even more complex tasks as taxonomy and ontology learning, metadata and provenance data maintenance, knowledge extraction, etc. can be performed by external servers and best possible providers. The RICHFIELDS app can be developed

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

independently and stores data on a server, which connects with the RDP in the same way as other app servers. More the RDP is modular, easier it can be upgraded or degraded to serve more or less functionality, respectively. This means that it is up to the RICHFIELDS management to decide which external parts of the RDP in which stage (CO, G, M) will be under the development and management of contractors and which will be inside the scope of RICHFIELDS.

The construction of the RDP with the actions that are listed on the roadmap should be done in cooperation with / learning from other EU-funded projects and initiatives (e.g. EIT Quisper, BBMRI-ERIC, Elixir, ENPADASI, MetroFood, etc.), and exploit the developments in the European Science Cloud to a maximum.

As this roadmap is closely related to other RICHFIELDS deliverables, we cited all the deliverables that include detailed information. However, the final RICHFIELDS roadmap should integrate all the information.

References

Agile manifesto. Available at: http://agilemanifesto.org (accessed on 15th June 2018).

DOI. Available at: https://www.doi.org (accessed on 15th June 2018).

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

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

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

GO FAIR Initiative. Available at: https://www.go-fair.org/fair-principles (accessed on 15th June 2018).

OpenAIRE Guidelines for Data Archives. Available at: https://guidelines.openaire.eu/en/latest/data/index.html (accessed on 15th June 2018).

Rams, D. Design by Vitsœ**.** Available at https://www.vitsoe.com/files/assets/1000/17/VITSOE_Dieter_Rams_speech.pdf (accessed on 15th June 2018).

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

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.