Tracing the Biological Origin of Food

H.E.J.M. Heuer MSc BA, S. van Leeuwen MSc, Prof. Dr. Ir. J.L. Top

PUBLIC

WAGENINGEN
UNIVERSITY & RESEARCH

Tracing the Biological Origin of Food

2020 deliverable of the Trusted Source project in cooperation with Questionmark

Authors: H.E.J.M. Heuer MSc BA, S. van Leeuwen MSc, Prof. Dr. Ir. J.L. Top

Institute: Wageningen Food & Biobased Research

This study was carried out by Wageningen Food & Biobased Research, subsidised and commissioned by the Dutch Ministry of Agriculture, Nature and Food Quality and funded by partners of TKI AF-16201/MIP Trusted Source.

Wageningen Food & Biobased Research
Wageningen, September 2021

Public

Report 2189

WFBR Project number: AF-16201
Version: Final
Reviewer: dr. R.J. (Rutger) Vlek
Approved by: dr.ir. H. (Henk) Wensink
Subsidised by: the Dutch Ministry of Agriculture, Nature and Food Quality
Funded by: partners of TKI AF-16201/MIP Trusted Source
Commissioned by: the Dutch Ministry of Agriculture, Nature and Food Quality
This report is: public

The research that is documented in this report was conducted in an objective way by researchers who act impartial with respect to the client(s) and sponsor(s). This report can be downloaded for free at https://doi.org/10.18174/553616 or at www.wur.eu/wfbr (under publications).

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

Contents

Summary

In the current consumer climate there exists a growing interest in the origin of our food, because of, for example, sustainability or human and animal welfare reasons. Therefore, initiatives are formed to accommodate transparency in food chains. The Trusted Source project intends to offer consumers and other supply chain actors access to information about food products and their provenance. This report focuses on the biological origin of ingredients of food products. The overall goal is to provide a way to clarify the origin of ingredients as being plant of animal-based, for as many products that can be bought at the supermarket as possible.

The report presents a way of disclosing the biological origin of food products. In order to express the biological origin of food products and to make the information needed for a digital platform as widely accessible and consistent as possible, a linked data approach was used. Using global references called 'Uniform Resource Identifiers (URIs)' an ontology with many food products was made, which includes a label marking them as animal based, plant based or of unknown origin: 'animal', 'plant' or 'unknown'. An algorithm was then created to check the origin of specific food products. This algorithm returns the biological origin of all of the ingredients of the input product. The ingredients of a small set of reference products were classified correctly.

1 Introduction

1.1 Background

Choosing the right food in accordance with our morals, our needs and wishes can be a tedious, time consuming, and most of the times even an impossible job. Proper information can help consumers make conscious decisions about food. This report focuses on providing information that is needed to know whether the ingredients of a product originate from plant or animal material.

The Trusted Source project, carried out by Wageningen Food & Biobased Research, subsidised and commissioned by the Dutch Ministry of Agriculture, Nature and Food Quality and funded by partners of TKI AF-16201/MIP Trusted Source, intends to offer consumers and other supply chain actors access to information about food products and their provenance. This project is divided into five pilots. This report belongs to the Consumer Pilot. The consumer's choices about sustainable and healthy food require awareness and insight. This can be obtained by providing information and data about the properties and the origin of a product. The philosophy behind this is that consumers become more aware of sustainable and healthy food products. The prediction is that as a consequence food producers will adjust to the wishes of the consumers and thus produce food in a more sustainable and healthy way.

The mission of Questionmark is to provide transparency of the sustainability efforts of food brands to customers, companies and other interested parties. Their main goal is aligned with the goal of this project: not to change consumer behaviour, but rather bring about change in the supply of food products. One of the methods with which Questionmark wants to achieve this, is by providing four "Superlists". These lists rank supermarkets of the Netherlands on four different themes: health, environment, human rights and animal welfare. In this specific sub-project of the Trusted Source project Questionmark is developing the digital platform to make this information available.

This report focuses on the environmental impact of food, in particular on the distinction between plant-based and animal-based ingredients. With the growing world population, the demand for high-quality protein rises. Currently, much of our protein consumption has animal origin. However, the production of animal protein is highly inefficient when considering the use of land, water and raw materials. Furthermore, agriculture causes large greenhouse gas emissions and in the western world there is more and more resistance against consuming animals due to the conditions in which animals are kept. Wageningen University and Research wants to aid the transition from a diet in which proteins are predominantly animal-based, towards a more plant-based diet. Wageningen Food and Biobased Research aims to realise the transparency that is needed for this transition. Therefore, in this collaboration with Questionmark, Wageningen University and Research wants to help consumers gain insight in their food choices by disclosing the biological origin of the food products they buy on a digital platform.

To make the information needed for Questionmark's intended digital platform as widely accessible and consistent as possible, the notion of linked data comes into play. The World Wide Web Consortium (W3C) is an international community that introduced this term in relationship to the Semantic Web. The main idea of the Semantic Web is to support a distributed web at the level of the data (Allemang & Hendler 2008). So instead of one webpage pointing to another, one data item can point to another, using global references called Uniform Resource Identifiers (URIs). The Web infrastructure provides a data model whereby information about a single entity can be distributed over the Web. A web application does not just publish a human-readable presentation of this information but instead a distributable, machine-readable description of the data. The data model that the Semantic Web infrastructure uses to represent this distributed web of data is called the Resource Description Framework (RDF).

1.2 Goals and research question

Together with Questionmark, Wageningen University and Research wants to gain insight into the ingredients of food products. This is done by first making an ontology database in which as many ingredients as possible are present. In this ontology, the products have to be labelled according to their biological origin: ‘plant’, ‘animal’ or ‘unknown’. For many products it is unclear whether they are animal or plant based by just looking at the product name. For example, lasagne may or may not contain meat. Therefore, the label unknown is used for those composite products. Some products, such as ‘peppermint’ or ‘beef’, can be labelled as completely animal or plant based. Once this database has been made, an algorithm will be developed to check whether a product has animal or plant ingredients. This algorithm can be used in applications for consumers.

The main research question of this report is:

What information is needed to determine if a product has animal or plant-based ingredients?

In order to answer the main research question, there are four sub-questions:

1. *How do we define the concepts ‘animal based’ and ‘plant based’?*
2. *Which products are animal based and which ones are plant based?*
3. *How can this information be expressed using linked data standards?*
4. *Which algorithm can be used to determine if a product has animal or plant-based ingredients?*

2 Method

As was said before, the goal of this research is to investigate the biological origin of ingredients of food products. The aim is to have products in a database as well as their ingredients with their matching biological origin. Only when the product is not available in the database or listed as unknown, the algorithm will get the ingredients of said product and determine their origin. In order to make a database with food products and their biological origin, which an algorithm can use for labelling ingredients, the first research question needs to be answered. The definition of when an ingredient has animal origin and when it has plant origin has to be clear. Does the emphasis lie on products containing animal based protein or is it important to use the much broader definition of whether a product contains ingredients for which animals are used or not? We decided to use a pragmatic approach and defined an ingredient as animal based if the ingredient consists of some animal part or if it is the secretion of an animal. For now, all ingredients that are atomic, and thus consist of only one type of ingredient, and that are not animal based nor of unknown origin will be categorised as plant based. During future research into the same topic the ontology can, for example, be extended with labels that categorise things as artificial or as specific animals. Ingredients can also be categorised as unknown, because information about the biological origin of a product is not always available. For example, the instance 'fruit liqueur' could include some kind of cream or milk secreted by an animal, but it can also be plant based.

To answer the second research question, which investigates the biological origin of ingredients, a database of food products and their ingredients is needed. With that data, an algorithm can look up an ingredient in the database, which then in turn can distinguish between animal and plant-based products. To build this database of food products, data sources where ingredients are stored with their biological origin have to be collected. To ensure accessibility and consistency of this database, it makes use of linked data technology. It uses a RDF database, an 'ontology', using the global references mentioned in the introduction, URIs. We can make use of the existing food taxonomy[1] the WUR has access to and we looked for data bases that can add ingredients to it. We will start with a first list of ingredients and if the algorithm is not accurate enough, we will add ingredients to the ontology, and we will look if additional information may help the classification of products.

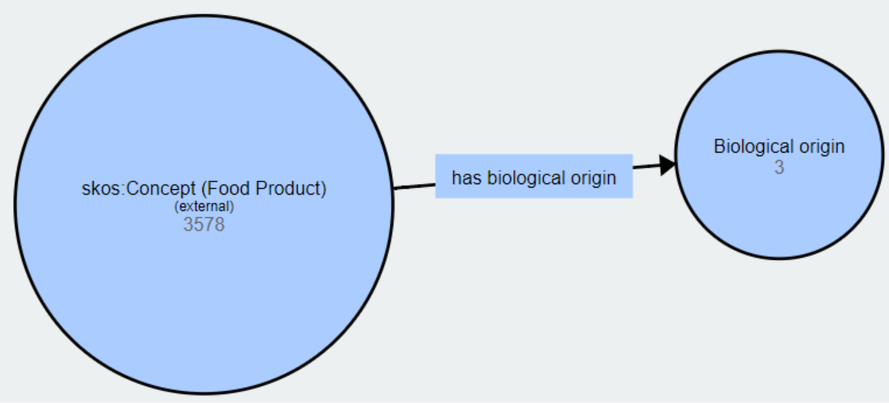

Figure 1: Structure of the ontology. The number of instances is shown in grey. The biological origin class contains three instances: 'plant', 'animal', and 'unknown'.

[1] http://foodvoc.org, accessed on 3-11-2020

The food taxonomy will be edited by adding the ingredients we find in the databases. The existing food taxonomy of the WUR is a SKOS ontology, which stands for Simple Knowledge Organization System. SKOS was designed to allow modelers to create modular knowledge organisations that can be reused and referenced across the web (Allemang & Hendler, 2008). However, a SKOS ontology just uses tags like ‘narrower’ and ‘broader’, whereas the current ontology needs a property that expresses the biological origin of the ingredient in the ontology. Therefore, the property ‘hasBiologicalOrigin’ with the class ‘BiologicalOrigin’ which includes three instances with labels ‘plant’, ‘animal’ and ‘unknown’ were added to the ontology, see Figure 1. All products were labelled. We used the software package TopBraid made by TopQuadrant to edit the ontology.

To formulate an algorithm that can decide whether a product contains animal or plant based ingredients, we use the output of an ingredient parser from Questionmark as input (Van Engen, 2020). They provided examples of products and their ingredients. From these lists it should be possible to determine whether a product contains animal-based ingredients, plant-based ingredients or whether it is uncertain whether it contains animal-based ingredients. For example, the ingredient parser lists the ingredients of the product ‘Becel Olie blend classic’ as follows:

```
{
   "contains": [
      {
         "name": "Plantaardige oliën",
         "contains": [
            {"name": "zonnebloemolie"},
            {"name": "koolzaadolie"},
            {"name": "lijnzaadolie"},
         ],
      },
      {"name": "vitamine E"},
   ]
}
```

3 Results

3.1 Data sources

As mentioned in the introduction, there are several data sources from which a list of animal and plant-based ingredients could be drawn. The sources which are used in this project are described below.

Foodvoc[2]
As it says on the Wageningen University and Research website, 'the objective of foodvoc.org is to publish vocabularies and directly interface with them'. This portal contains a food taxonomy vocabulary, or 'ontology', where many food products and ingredients can be found.

Vegan Wiki[3]
At the vegan wiki website, information can be found about vegan products in the Netherlands. Their website is kept up to date by volunteers of the Dutch Association of Veganism (Nederlandse Vereniging voor Veganisme, NVV). They also provided us with a list of E numbers that are animal based, are potentially animal based, are rarely animal based or are never animal based. Unfortunately, the website does not provide further details on how the classification was created.

Websites AH, Jumbo[4]
Regulation (EU) 1169/2011 is the main law of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers. [5] It is an amendment to earlier versions of this law. This regulation defines to what extent food information must be provided with products. Here it could be found that a fresh product with only one ingredient does not need an ingredient list, as long as it is clear what product it contains. Therefore, we looked at the fresh products from supermarket websites AH and Jumbo to add their fresh products to the ontology as well.

Food Standards Agency
The Food Standards Agency (2010) lists the current approved additives and e numbers of the EU. All E numbers could be added to the food taxonomy and the Vegan Wiki could tell which E numbers were plant or animal based.

Questionmark
Questionmark provided us with ingredient spreadsheets that included many dairy, meat and fish products with their ingredients.

The resulting ontology containing 3578 instances for food products and ingredients with their biological origin can be found at GitHub[6]. Access to this information can be requested from Görkem Simsek-Senel (gorkem.simsek-senel@wur.nl).

[2] http://foodvoc.org
[3] Veganwiki.nl
[4] ah.nl and jumbo.nl
[5] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32011R1169
[6] git.wur.nl/FoodInformatics/food-taxonomy/-/blob/develop/Product.ttl

3.2 Algorithm

The overall goal of the algorithm is to take a list of ingredients as input and output the biological origin of those ingredients. For each ingredient, the algorithm has to check in the ontology what its biological origin is or move one level deeper in terms of ingredients. The algorithm is structured as follows in pseudocode:

Algorithm used for ingredient classification in pseudocode

```
INPUT ingredient list

CREATE empty dictionary for ingredients and their biological origin {}
CREATE empty list for additions []

FOR EACH ingredient in the list:
        CHECK with a SPARQL query in the ontology what the biological origin is
        IF the query has found a biological origin in the ontology:
                ADD it to the dictionary
        ELSE:
                ADD ‘not available’ to the dictionary
        IF an ingredient consists of multiple other ingredients, all of those are checked
        again recursively UNLESS the label is ‘plant’ OR ‘animal’

RETURN dictionary with the origin of all ingredients AND list with ingredients that should be
added to the ontology
```

The actual code can be found in Appendix 1. See the below example that shows the input of the algorithm and the returned output for four cases:

Input:

```
becel = {"contains":[{"name":"Plantaardige oliën", "contains": [{"name": "zonnebloemolie"},
{"name":"koolzaadolie"},{"name":"lijnzaadolie"}]}, {"name":"vitamine E"}]}
jumbo = {"contains":[{"name":"Natuurlijk mineraalwater"}, {"name":"koolzuur"}]}
smaakt = {"contains":[{"name":"Linzen","marks":["*"]}],"notes":["*=biologisch"]}
parma = {"contains":[{"name":"Varkensvlees"},{"name":"zout"}]}
```

Output:

```
>>>>  {'plantaardige oliën': 'not available', 'zonnebloemolie': 'plant', 'koolzaadolie': 'pl
ant', 'lijnzaadolie': 'plant', 'vitamine e': 'unknown'}
>>>>  {'natuurlijk mineraalwater': 'plant', 'koolzuur': 'plant'}
>>>>  {'linzen': 'plant'}
>>>>  {'varkensvlees': 'animal', 'zout': 'plant'}
```

Each ingredient is listed with its biological origin. Once a food item is categorised as ‘plant’ or ‘animal’, the algorithm does not go deeper into the nested structure, but it assumes all its ‘sub ingredients’ are animal or plant based as well.

4 Discussion

In the ontology, identification and labelling of ingredients are two separate things. Although the ontology is filled with as many ingredients as we could get from our sources, there are still ingredients missing. Especially the number of plant-based ingredients can be improved. When considering the labelling of the ingredients, semantic technology is really useful for multilingual support of the same knowledge. Currently, the ontology consists mostly of Dutch labelled products and ingredients. To make the database more internationally deployable, English and other language labels should be added to the ingredients. An auto-translation for multi-lingual support can be used, which a native speaker can then check. Furthermore, the naming of ingredients is not static. New ingredients are invented or found, and names change. A term frequency analysis on the output of the parser can tell which new ingredients pop up and require addition to the ontology. To improve the database on the above mentioned points we suggest that it should be possible for volunteer contributors to add labels to the ontology and make someone the ontology administrator that can accept the suggested changes.

To make the database even more specific, labels such as ‘artificial’ and specific animal or plant labels such as ‘sheep’, ‘cow’, etc. should be added to the Biological Origin class in the ontology. This way ingredients and products can even be more specific as to what kind of animal is contained in or used for the product, or whether the product contains artificially made ingredients. In the current setup, we have used the label ‘unknown’ to express the fact that we do not know whether a product is completely plant or animal-based. In hindsight we think it is not desirable to have an ‘unknown’ label in the biological origin. After applying our algorithm, for any food product for which the biological origin label is still missing, we know that either we don’t know its ingredients, or it has not been classified explicitly as being plant-based, animal-based or artificial. This product would need further attention from the data owners. Only food products that have a known biological origin (by itself or at some level of its ingredients) are informative for the consumer using the application.

The scope of this project was to derive origin solely from the name of an ingredient. Given that ingredient names do not unlikely identify a specific substance, but can refer to a group of substances, it is not always possible to provide unambiguous information about origin (as in the liqueur example). In the future, origin could alternatively be shared through the supply chain as a form of structured data, linked to a more uniquely identified ingredient in a uniquely identified product. By combining this information, the origin could be retrieved when the origin of an ingredient is ambiguous. Furthermore, digital food product data might contain additional sources of information that could help establish origin of the product. In this subproject, ingredient names were considered the most valuable source of information and hence kept in scope. Some statements on the labels, such as a mark that says whether a product is vegan or vegetarian could enhance the categorisation of food products. Legally required statements about fishing region also may corroborate information about ingredients being animal based. This is something that can be investigated in further research.

Overall, once the above mentioned additions are implemented, the desired structure of the ontology can be found in Figure 2. Here, the input products can be listed in a SKOS ontology. On the other hand, it is also linked to the biological origin database, where the origin of each ingredient can be disclosed to as much detail as required. Then questions like ‘I want a product that is at least 30 percent plant based’ can be answered by the information in the database.

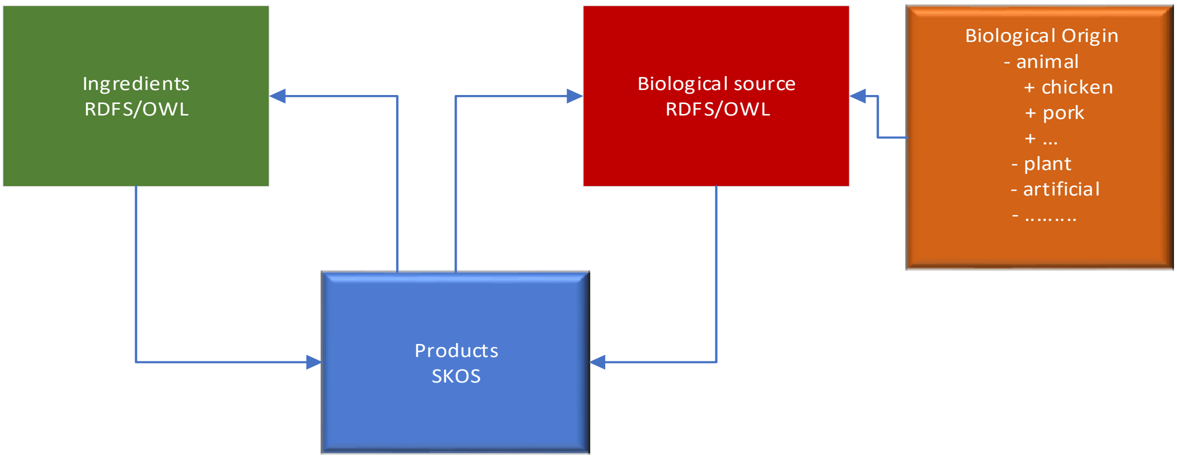

Figure 2: Desired information flow of the ontology. In blue, the input products are structured in a product ontology. In green, the ingredients of the product can be found. In red, the ontology that has been made in this research project can be found, with a link to a biological origin class, where all possible biological origins are stored.

At WUR there is ongoing research into personalised dietary advice and sustainability indicators for food, which can be improved using the biological origin information provided by this model in combination with the proposed algorithm. Also other attributes are added, such the amount of fibre in a product or its taste and nutritional value.

5 Conclusions

Figuring out the biological origin of food products can be a hard job for consumers. Icons like 'vegan' or 'vegetarian' may help them slightly, but not all food producers include that in their labels. Moreover, these terms are not used consistently. Nowadays, there is an abundance of products and information on those products that make it difficult for consumers to keep an overview and choose the right product. Proper product information can help them make conscious decisions about food, which is one of the goals of the Trusted Source project.

In conclusion, this report focussed on providing information about whether the ingredients of a product are plant or animal based. A linked data approach was used in order to make the information needed for digital distribution as widely accessible and consistent as possible. This research has made efforts towards an application consumers can use to determine the biological origin of the products they buy. The application has not been made yet, but with the current ontology and algorithm, it is possible to classify parsed products with their ingredients and give them labels, including 'plant', 'animal', 'unknown' and 'not available'. The algorithm takes into account that the ontology is a work in progress by returning a list of ingredients that can be added to the ontology as well. In time, the 'unknown' and 'unavailable' labels can be dropped, because only food products that have a known biological origin are informative for the consumer using the application. This paves the way for future projects where this algorithm can be tested in practice.

Literature

Allemang, D., Hendler, J. (2008). Semantic Web for the Working Ontologist. Morgan Kaufmann.

Current EU approved additives and their E Numbers, Food Standards Agency, 26 November 2010.

Van Engen, W. (2020). *Food ingredient parser ruby*. Github. https://github.com/q-m/food-ingredient-parser-ruby

Annex 2 Tracing algorithm Python

```python
becel = {
    "contains": [
        {
            "name": "Plantaardige oliën",
            "contains": [
                {"name": "zonnebloemolie"},
                {"name": "koolzaadolie"},
                {"name": "lijnzaadolie"},
            ],
        },
        {"name": "vitamine E"},
    ]
}
jumbo = {"contains": [{"name": "Natuurlijk mineraalwater"}, {"name": "koolzuur"}]}
smaakt = {"contains": [{"name": "Linzen", "marks": ["*"]}], "notes": ["*=biologisch"]}
parma = {"contains": [{"name": "Varkensvlees"}, {"name": "zout"}]}
x = {
    "contains": [
        {
            "name": "becel",
            "contains": [
                {
                    "name": "Plantaardige oliën",
                    "contains": [
                        {"name": "zonnebloemolie"},
                        {"name": "koolzaadolie"},
                        {"name": "lijnzaadolie"},
                    ],
                },
                {"name": "vitamine E"},
            ],
        }
    ]
}

# here we import the ontology with the rdflib package. Maybe we could make this refer to the foodvoc website?
import rdflib

product_graph = rdflib.Graph()
product_ontology = product_graph.parse(
    "C:\\Users\\heuer002\\Projects\\trusted-source\\Product.ttl", format="n3"
)

def ingredient_origin(product, origin, additions):

    # a product contains several ingredients and for each ingredient we have to check the origin in the ontology
    for ingredient in product["contains"]:

        # we use a sparql query to check whether the ingredient has a perfect match with a prefLabel or altLabel
        # and get the biological origin
        sparql = product_ontology.query(
            '''''
            SELECT ?bio_origin
            WHERE
                {?product qm:hasBiologicalOrigin ?bio_origin.
                 ?product skos:prefLabel ?label.
                 OPTIONAL{ ?product skos:altLabel ?label2}
            FILTER( ?label = "'''
            + ingredient["name"].lower()
            + '''''" || ?label2 = "'''
            + ingredient["name"].lower()
            + """")
            }
            """
        )

        # if the origin of an ingredient is available, we add this to the origin dictionary
        # The name of the ingredient as the key and the biological origin as the value
        if sparql:
            for elem in sparql:
                bio_origin = elem[0].replace(
                    "http://www.foodvoc.org/resource/FoodTaxonomy/Product#", ""
                )
            origin[ingredient["name"].lower()] = bio_origin
        # if the origin is not available in the ontology, we add 'not available' as a label
        else:
            origin[ingredient["name"].lower()] = "not available"
            # we add the label to a list which could be added to the ontology
            additions.append(ingredient["name"].lower())

        # if an ingredient consists of multiple other ingredients, all of those are checked again recursively, unless
        # the label is 'plant' or 'animal', the label of the 'contains' ingredients is the same
        if ("contains" in ingredient) and (
            origin[ingredient["name"].lower()] != "plant"
            or origin[ingredient["name"].lower()] != "animal"
        ):
            origin, additions = ingredient_origin(ingredient, origin, additions)

    return origin, additions

def main(product):
    origin = {}
    additions = []
    origin, additions = ingredient_origin(product, origin, additions)
    print(">>>> ", origin)


#     print(additions)

       # these are the 4 test products
       main(becel)
       main(jumbo)
       main(smaakt)
       main(parma)
       main(x)
```

To explore the potential of nature to improve the quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
www.wur.eu/wfbr
E info.wfbr@wur.nl

Report 2189

Public

The mission of Wageningen University and Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 5,000 employees and 10,000 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.