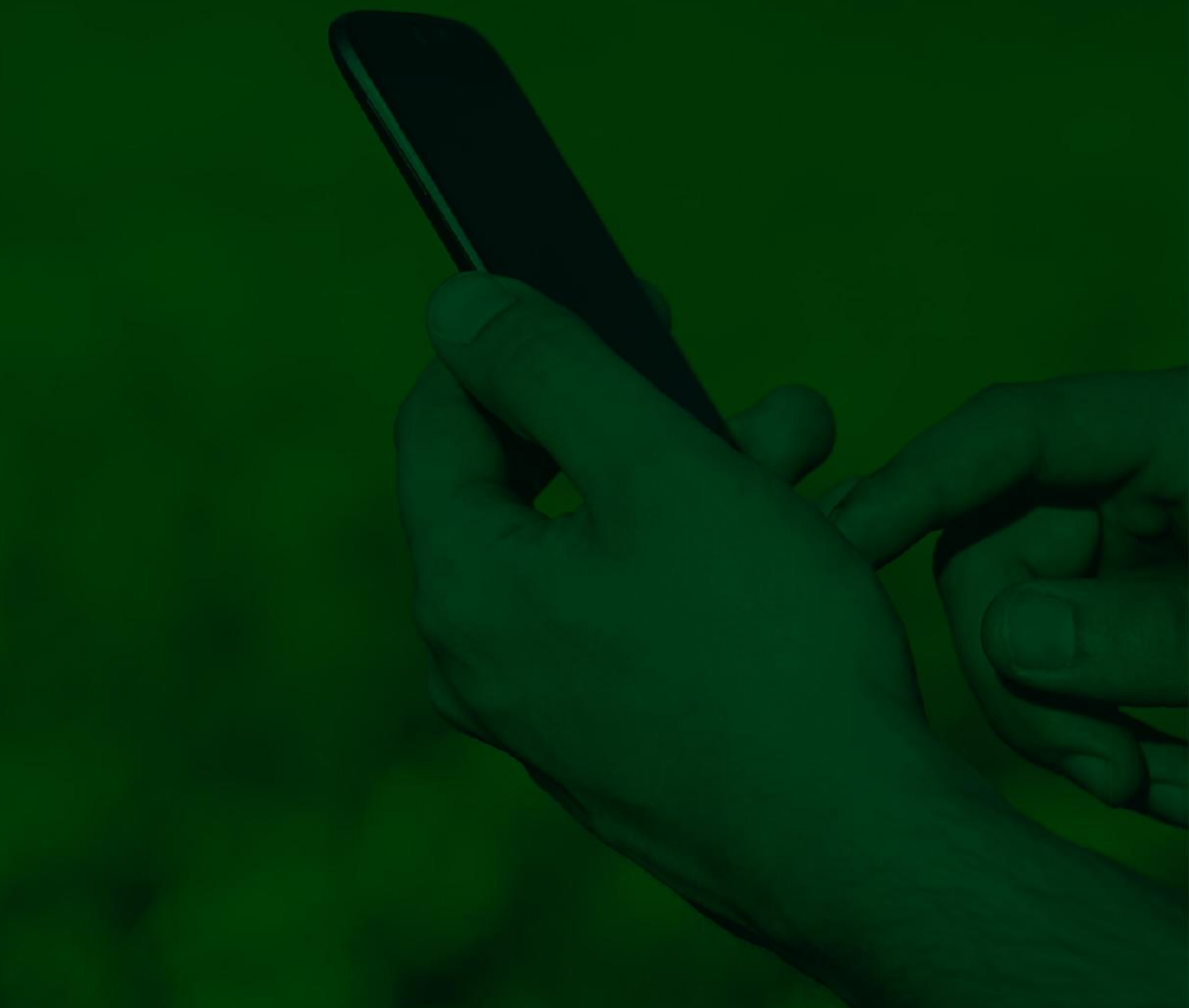
Ref. Ares(2020)6188760 - 31/10/2020

FAIRshare

DIGITAL TOOLS FOR FARM ADVISORS



FAIRshare
DIGITAL TOOLS FOR FARM ADVISORS

D1.2: Web-based semantic search engine

Authors: Kevin Jacobs (WR), Julian Bianco-Martinez (WR)

Contributions from: Lorijn van Rooijen (WR), Gorkem Simsek-Senel (WR), Hein ten Berge (WR), Jan Top (WR)

Technical References

Project Acronym	FAIRshare
Project Title	Farm Advisory digital Innovation tools Realised and Shared
Project No.	H2020 818488
Project Coordinator	TEAGASC
Contact	Tom Kelly (Tom.Kelly@teagasc.ie) and John Hyland (John.Hyland@teagasc.ie)

Deliverable No.	D1.2
Deliverable Name	Web-based semantic based search engine
Dissemination level	Websites, patents filling, etc.
Work Package	WP1
Task	T1.2: Develop and implement a semantic webtool for advisors to find digital solutions that match farmers' challenges
Lead beneficiary	Wageningen Research
Contributing beneficiaries	
Deliverable type	Documentation, API
Due date of deliverable	31/10/2020
Actual submission date	31/10/2020

v	Date	Beneficiary	Author
1	31/10/2020	Wageningen Research	Julian Bianco-Martinez (WR), Kevin Jacobs (WR)

FAIRshare
DIGITAL TOOLS FOR FARM ADVISORS

Table of Contents

1. Introduction 4
2. Concept completion 5
3. Search 6
4. Integrated functionalities 8
5. Future Steps 9

FAIRshare

Semantic based Search Engine

1. Introduction

This document presents the results of the T.1.2 task related to FAIRShare web base Search engine API.

For this task, DATS and other domain relevant documents were used as an initial corpus to extract a vocabulary. Based on this vocabulary, domain experts were involved in the development of an ontology used by the search engine.

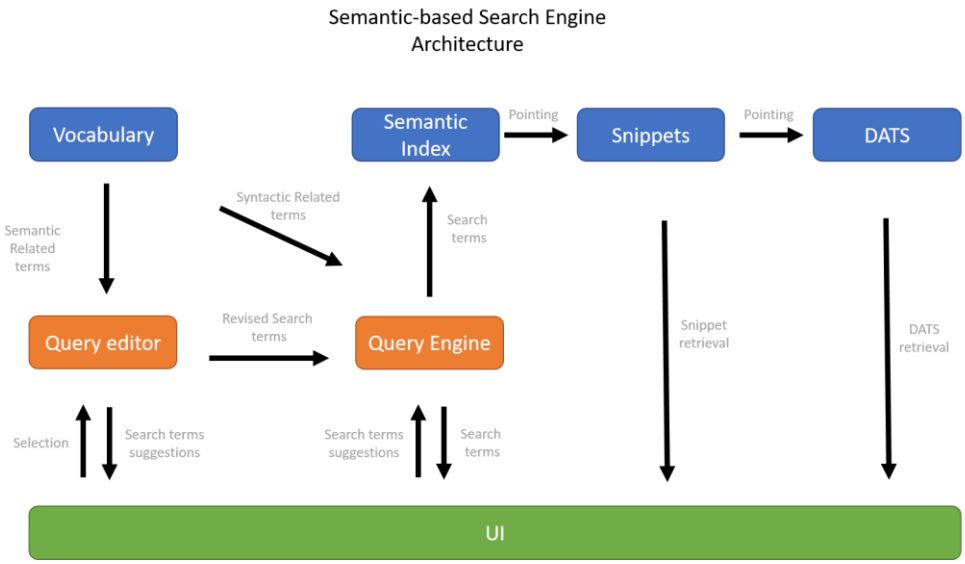


Figure 1. Diagram of the semantic based search engine architecture. The colors of the different nodes represent different general functionalities (see text).

The main components of the search engine are visualized in the architecture diagram showed in figure 1. These components are separated by functionality using a node colour code. The green node represents the front-end application. Blue nodes represent static data i.e. part of text extracted from within documents (snippets), name of DATS, etc. This data was created by analysing the DATS and other domain relevant documents. Finally, orange nodes represent software nodes. These nodes process information coming from both blue and green nodes.
The rest of the document is composed of two sections describing functionalities of API endpoints. In section 2, **Concept completion**, we present the concept completion

endpoint. When the user types a query, the API retrieves a set of possible concepts associated with the user's input (search query). Section 3, focuses on the search endpoint. This endpoint retrieves all information from DATS, including names and snippets from documents, based on the query of the user.

2. Concept completion

The first endpoint is an endpoint for concept completion. It will help to complete a concept while entering a query.

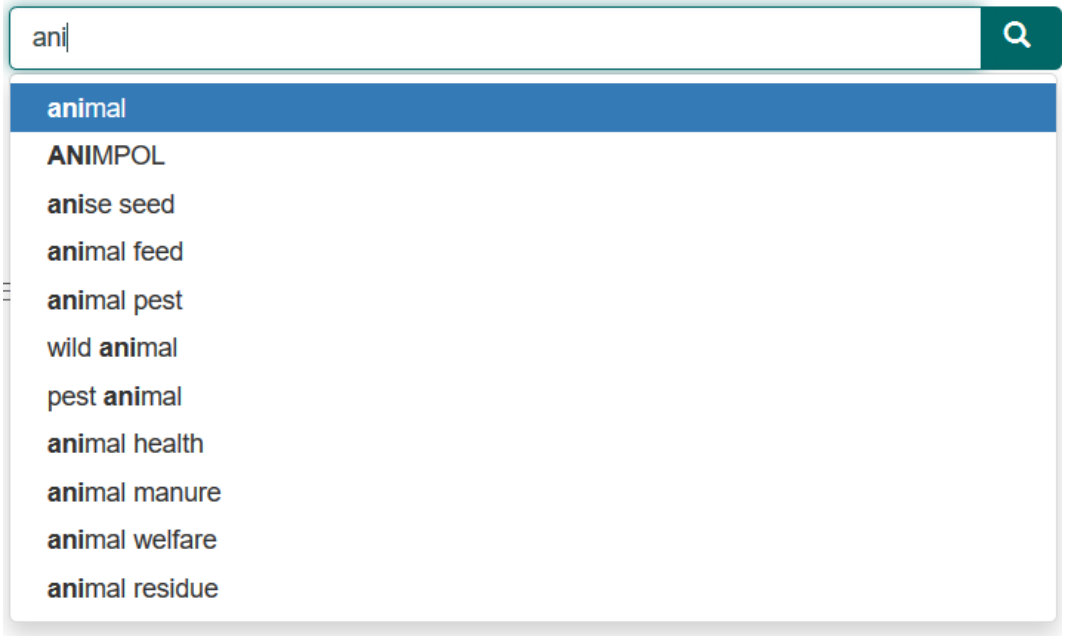


Figure 2. Concept completion example for query "ani".

The concept completion endpoint can be found at the following URL:
https://fairshare.wur.nl/api/complete-concept?query=ani

It has the following parameters:

Parameter	Description
query	The query that is entered.

The output has the following structure:

```
{
   "concepts": [
      {
         "language": "en",
         "label": "animal",
         ...
      },
      {
         "language": "en",
         "label": "ANIMPOL",
         ...
      },
      ...
}
```

Figure 3. The output of the concept completion endpoint.

The output is a JSON representation of the concepts found for the given query. Each concept is described by a unique identifier, a label and its language.

3. Search

The search endpoint will return fragments found within the documents that are about the given query.
The search endpoint can be found at the following URL:
https://fairshare.wur.nl/api/search?offset=0&query=ani

This endpoint has the following parameters:

Parameter	Description
Query	The query that is entered by the user.
Offset	The number of results to skip, used for navigating the results.

The output has the following structure:

```
{
   "concepts":[
      {
         "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
         "language":"en",
         "label":"animal"
      }
   ],
   "query":{
      "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
      "language":"en",
      "label":"animal"
   },
   "results":[
      {
         "fragment":"\u003cspan class\u003d\"keyword\"\u003eAnimal\u003c/span\u003e welfare
is definitely important to \nconsumers (but a "concerned minority")    [page 22]",
         "concepts":[
            {
               "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
               "label":"animal"
            }
         ],
         "language":"en",
         "title":"d56-practical-on-farm-welfare-assessment.pdf"
      },
      {
         "fragment":"Salmonella o Neospora y posee el mejor estado (10) \nen la enfermedad de
Johne. Todo está cer tificado y \nlos \u003cspan
class\u003d\"keyword\"\u003eanimales\u003c/span\u003e no están vacunados. \n [page 11]",
         "concepts":[
            {
               "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
               "label":"animal"
            }
         ],
         "language":"gl",
         "title":"d55-technical-leaflets-on-animal-care.pdf"
      },
      {
         "fragment":"Some of the current management practices regarding farm \u003cspan
class\u003d\"keyword\"\u003eanimal\u003c/span\u003e care and \u003cspan
class\u003d\"keyword\"\u003eanimal\u003c/span\u003e welfare \npose a reputational and
business risk to dairy farming and wider industry. It is likely that \nconsumers will
continue to see some of these management practices, no matter how well \nconducted, as
unnatural.  \n [page 4]",
         "concepts":[
            {
               "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
               "label":"animal"
            }
         ],
         "language":"en",
         "title":"d59-animal-care-rd-and-policy-recommendations.pdf"
      },
      {
         "fragment":"  Develop quantitative risk assessment methodologies, using \u003cspan
class\u003d\"keyword\"\u003eanimal\u003c/span\u003e based measures  \n [page 5]",
         "concepts":[
            {
               "resource":"http://www.foodvoc.org/resource/Valerie-9/Animal",
               "label":"animal"
            }
         ],
         "language":"en",
         "title":"d12-collation-of-topics-identified-for-regional-and-eu-rd-agendas.pdf"
```


Figure 4. Output of the search endpoint.

At the highest level, the output consists of concepts, the query and the results. The output concepts are the concepts that were deduced from the query. These concepts are used to search the documents. In the results, the name of documents are given, together with text fragments and concepts, associated with the query matched. The title field contains the filename of the document.

4. Integrated functionalities

WR has developed a FAIRshare ontology based on DATS (see figure 5) and other domain relevant documents. This ontology is the main source of information for the development of the semantic based search engine. The engine uses the ontology and taxonomy to identify related terms, then returns the most suitable documentation (DATS name) and text information (Snippets) to the user's search word.

AUA in collaboration with WR, will connect the smart search engine to the Permanent Networking Facility (PNF) platform. After this step, the current string matching search functionality will be deprecated, and replaced by the semantic based search engine.

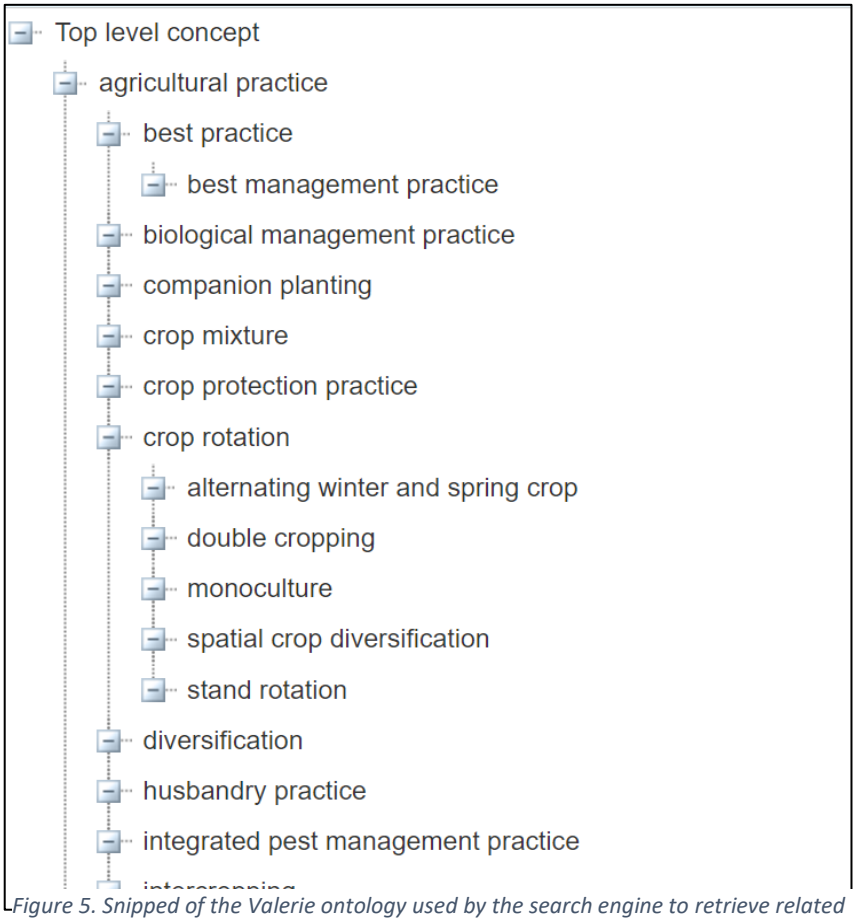


Figure 5. Snipped of the Valerie ontology used by the search engine to retrieve related ontological terms.

Both endpoints presented in this report will be implemented in the forthcoming version of the PNF. This new version will add two features: first, a search functionality with auto-completion and auto-suggestions, and second, a Valerie [1, 2] search capability.

The **Concept completion** endpoint will be part of the development of the first feature. To retrieve words used for the auto-completion feature, the endpoint scans the backend ontology and retrieves these words based on a string distance between the user's query and the ontology concepts.

Likewise, the endpoint scans the taxonomy and retrieves potentially useful terms to the user's query. The auto-suggested words depend, not only on a string comparison, but also on the taxonomic structure of the ontology.

The search functionality, accessible through the **Search** endpoint, is a functional copy of the Valerie widget. However, the initial corpus for the ontology development has been constructed using only the DATS and initial FAIRShare domain relevant documents.

5. Future Steps

Below, we present the steps to be taken to further develop the semantic based search engine. This is an iterative process that will continue throughout the FAIRshare project, and is largely dedicated to covering a much wider knowledge base, both in terms of documents and languages, and in terms of the knowledge (sub)domains covered by the ontology.

1. The current FAIRShare ontology will be enriched by the addition of new terms and the document base will be expanded by new documents. Those will include
   (i) all documentation accessible on the DATs covered by the FAIRshare DATs inventory;
   (ii) (ii) more general documentation on decision making and practices in the domains addressed by the DATs (e.g. dairy farming; open field vegetables; etc.). The FAIRshare DATs inventory itself is a moving process that should yield more and more DATs as the project proceeds.
2. The Taxonomy will be further developed by addition of hierarchical and 'related-to' relations; workshops with project domain experts and WR ontology experts will be held to establish such relations, thus enhancing the power of the query editor in the search engine, to optimize the search strategy. Since the development of the ontology is an iteration process it will require multiple sessions which will be organized on a regular basis in the remaining time of the project. The selection and prioritization of subdomains to be elaborated will be the outcome of a consultation within the project with all relevant tasks.
3. A maintenance and update process will be designed by WR to allow for continuous development of FAIRShare search capabilities. The reference

frequency for large updates – of both document base, ontology and possibly functionality - is foreseen to be annually.

6. References

1. Willems, D., Koenderink, N. J.J.P., Top, J. L., *From science to practice: Bringing innovations to agronomy and forestry,* Journal of Agricultural Informatics, 2015, Vol. 6, No. 4:85-95, doi: 10.17700/jai.2015.6.4.214
2. Bechini, L. , Koenderink, N., ten Berge, H. F. M., et al., *Improving access to research outcomes for innovation in agriculture and forestry: The VALERIE project,* Italian Journal of Agronomy, 2017, Vol. 12, No2:85-95, doi: 10.4081/ija.2016.756