How can we be more FAIR in science? A look at the *Aspergillus fumigatus* field.

Sibbe Bakker[1] *   Mariana Santos Couto Silva[1] **

Anna Fensel[2] ***

[1]Genetics department
[2]Data Competence Centre

Wageningen University
January 15, 2024

Abstract

Adherence to findable, accessible, inter-operable & re-usable (FAIR) principles means sharing your data in such a way that it is easy for others to use. These practices are not yet widespread in academia. In this thesis, I have investigated how to make the scientific research more FAIR by working on a FAIR usecase in mycology: research into the genetics and spread of azole resistance in *A. fumigatus*. This ascomycete lives as a saprotroph in decaying plant material. Since fungicides are applied in agricultural domains against fungal pathogens. This over-usage leads *A. fumigatus* to evolve azole resistance. *A. fumigatus* frequently infects immunocompromised people, which makes azole resistance an issue of global emergent health concern. Mycologist are creating lots of genotyping and resistance phenotyping data for *A. fumigatus*. The ability to share between these researchers is lacking. To increase the quality of data sharing, the use of an existing FAIRification platform–FAIRDS–has been explored. While investigating the needs of the azole community, researchers indicated apprehension regarding the standardisation and sharing of 'their' data. While working on the FAIRDS, progress was made with increasing awareness of FAIR principles within the *A. fumigatus* field. The main output of this thesis is a list of recommendations of how to build a central FAIR data repository for *A. fumigatus*.

*Email: `sibbe.bakker@wur.nl`.
**Daily supervisor and project leader.
***Master thesis advisor.

Key words — Data management, FAIR data & Aspergillus fumigatus.

Further information: Please see the git.wur page for the latest progress and documents (https://git.wur.nl/aspar_kr/). The prototype ASPAR_KR platform is available on https://aspar-kr.bioinformatics.nl/.

Glossary

SPARQL protocol and RDF query language SPARQL protocol and RDF query language (RDF) is a programming language with the purpose of searching graph databases. 2, 3, 8, 28

R Language for statistical analysis . 10, 14, 17, 21, 27

azole resistance Azole resistance in *A. fumigatus* has different meanings depending on context: In the medical setting, an isolate shows resistance when its growth is inhibited by a minimal inhibitory concentration (MIC) under a specific breakpoint. Environmental azole resistance however is simply when azoles do not affect the growth of the isolate. 5, 25

clustered regularly interspaced short palindromic repeats-CAS-9 An enzyme complex that is guided by RNA to a cutting site in DNA. 3, 17

cold spot Oposit of a hot spot, where *A. fumigatus* finds it hard to proliferate, and cannot easily develop azole resistance. *cf.* hot spot . 2, 6

geoSPARQL protocol and RDF query language geoSPARQL is a set of functions for SPARQL to run geo-spacial queries. This is a standard described by the open geogspacial consortium: https://www.ogc.org/standard/geosparql/. 4, 7

hot spot An environement that is ideal for the proliferation of *A. fumigatus*, and has a high presence of azoles–driving the selection for azole resistance. *See* Schoustra *et al.* [64]; *cf.* cold spot . 2, 6, 25

metadata Data about data (hence metadata), concerns all the context and extra information one needs to understand a given data point. For a publication, the metadata is the list of authors who published it. For a lab assay, it may be the lab where the result was obtained and the protocol that was used. 5–7

ontology Studer, Benjamins, and Fensel [69] describes an ontology as being a ‘formal, explicit specification of a shared conceptualisation’. This refers to the understanding of a topic, being a machine readable version of a model obtained by consensus. An ontology is commonly used to store the rules of a graph database. 2, 7, 8, 11, 21

regex A language for matching patterns of text. Available in most text editors and viewers. See https://regex101.com/ to try it out . 17

resource description framework A file format designed initially to store metadata. Currently in use as a popular linked data file format. Data is encoded as a subject-predicate-object triple. See `https://www.w3.org/RDF/` for more information. 3, 5

shapes constraint language shapes constraint language (SHACL) provides a way of specifying whether an resource description framework (RDF) file adheres to a pattern. 3, 20

taxonomy A taxonomy is a hierarchical definition of terms and their relations. An example of an taxonomy is the way scientific projects are planned: with the project having sub questions answered with experiments. Each experiment tracks experimental units and samples. 8

uniform resource identifier An unique identifier for a resource, such as a file. 3, 7

uniform resource locator A path that links the computer to a place where an uniform resource identifier (URI) is found. 4

vocabulary A list of terms with no relation between them. 11

Acronyms

`ASPAR_KR` *A. fumigatus* azole resistance knowledge resource. 2–37

CRISPR-CAS-9 clustered regularly interspaced short palindromic repeats-CAS-9. 17, *Glossary:* clustered regularly interspaced short palindromic repeats-CAS-9

FAIRDS FAIRDS-data station. 1, 4, 11–14, 17, 19–23, 27, 29

FAIR findable, accessible, inter-operable & re-usable. 1, 4, 6, 8–11, 13, 17, 21–23, 27, 29

ISA investigation, study and assay. 5, 8, 11

JERM just enough results model. 5, 7, 11

MIC minimal inhibitory concentration. 2, 4, 17, 18, 20, 27

MIxS minimum information about any (x) sequence. 4, 5, 7, 8

NGINX engine-x. 13, 14, 26

PPEO plant phenotyping experiment ontology. 7, 11

RDF resource description framework. 2, 3, 5, 8, 9, 13, 14, 20–22, 27, 28, *Glossary:* resource description framework

SHACL shapes constraint language . 3, 20, 21, 28, *Glossary:* shapes constraint language

SPARQL SPARQL protocol and RDF query language. 2, 3, 8, 10, 13, 16, 17, 20, 21, 27, 28, *Glossary:* SPARQL protocol and RDF query language

UML unified modeling language. 10, 11

URI uniform resource identifier. 3, 7, *Glossary:* uniform resource identifier

URL uniform resource locator. *Glossary:* uniform resource locator

YODA YAML observations data Archive. 13, 21

MIxS-sa minimum information about any (x) sequence (MIxS) symbiont associated. 7

geo-SPARQL geoSPARQL protocol and RDF query language. 7, 13, 14, *Glossary:* geoSPARQL protocol and RDF query language

schema.org schema.org web architecture. 5, 7, 11

Contents

1 Introduction

1.1 Importance of improving data sharing solutions

Computational methods in biology are improving year by year. Both the computational hardware, as well as the software, are be becoming more accessible and capable. Upcoming analysis tools such as generative pre-trained transfomers (GPT) [86] are powerful in the analysis and prediction of sequential data, such as weather [45] or a coding assignment. Such models and need large amounts of high quality data. One aspect of quality data that is often forgotten in the life science is that quality data must be placed into context: when and where was it collected? What other observations are related? &c. Even though computers take over more pattern recognition tasks, it remains the analyst who needs to find the dataset, and judge its quality. For this the analyst needs more data: Who collected the data? How many samples were taken with what method? Do the signals in the data make sense from a domain perspective?

These questions can be answered with metadata: the data about data. Standardisation systems exist for basic information such as author data, publishers, licensing, and general dataset structure: schema.org web architecture (schema.org), dublin core [23] and resource description framework (RDF)-schema [21]. These metadata standards are supplemented with domain specific standards to produce standards that are useful to a field. For the lifesciences, standards for biological data include investigation, study and assay (ISA), just enough results model (JERM) and minimum information about any (x) sequence (MIxS). Such standards keep track of experimental design, data type and citation data.

Data management is an important aspect of modern life science. The time where data could just be kept on paper is over. Now, data sets are getting bigger and more complex and researchers are expected to be data managers & analysts: Most journals and funders expect data to be shared in a way that other people can easily use it. Most researchers (85 %) feel a need to share their data [74]. Most researchers however, have issues with turning their data into a form that can be easily shared. A large chunk of researchers (29.8 %) keep their data on an thumb-drive for general storage [74].

When data is mismanaged by not using standardised metadata, there are difficulties in sharing it. Tedersoo *et al.* [71] found that, of all authors in *nature* and *science* who indicated that their data is available open request, only on average $69.5\,\%$ were able to provide the full dataset within 60 days. Tedersoo *et al.* [72] recommend that data should be available open publication. Yet even when data was readily available, [72] still had to contact a lot of authors to get all the metadata clear. Findable, accessible, inter-operable & re-usable (FAIR) data sharing may currently not be occurring because a lack of awareness and technical ability.

1.2 ASPAR_KR will prototype a platform for *A. fumigatus* data standardisation

To determine what issues might arise when advocating for FAIR data in a lifescience domain, the case of *A. fumigatus* research will be studied (see Appendix A). The issues encountered and solutions to them will be useful for improving data sharing for both the *A. fumigatus* domain and the lifesciences as a whole. The *A. fumigatus* domain is varied, with data sets coming from extensive field sampling or from well defined strains in the laboratory. One project that is interesting from a metadata perspective is schimmelradar, where the azole resistance of airbore spores is determined in various locations in the Netherlands. In the case of the SchimmelRadar project, being able to inter-operate their data with the Dutch geographical open data platforms [32] will make it easier for geographical relations to be understood. What happens when quality disease data is not available for epidemiological study is clear: during the COVID-19 pandemic, a lot of 'worthless' sequence data has been generated: the data that was uploaded to the databases, 25% did not include any geographic information [65]. Now these sequence data cannot be used to analyse how they fit in the COVID-19 pandemic.

To solve the data sharing issues in the *A. fumigatus* community, a platform for data and knowledge sharing will be prototyped: the *A. fumigatus* azole resistance knowledge resource (ASPAR_KR). ASPAR_KR should facilitate data sharing, give guides on reproducible data analysis and outline community agreed-on definitions. These basic requirements will be extended and implemented based on a survey of community needs. After the basic requirements have been extended, the prototype will start with a programme to standardise data from researchers. Such a data standardisation programme will be implemented after a survey of available programmes.

When *A. fumigatus* data is standardised and stored centrally, re-analysis is possible: for example to track the dispersion of cold spots and hot spots. Besides the data sharing options that can be offered by the ASPAR_KR platform, the experience of developing a community platform for sharing FAIR data can be applied to other domains where the user base is similar. An example of such a similar user base is the *Candida* domain.

2 Related work

2.1 Metadata management practices

2.1.1 What standardised metadata is

Metadata, data about data, is central to understand the context of a dataset and to index it in a database [57]. Examples of metadata are the treatments applied to a sample, or the instrument in a measurement. If the metadata is not of sufficient quality users of the dataset cannot understand the data well enough to use it. It might also be that missing metadata lead to data not being found, a shame in a climate where citations are so important. Furthermore, the lack of standardisation makes it hard for computer systems to automatically explore and analyse datasets on the internet.

Even if metadata is present, its usefulness may still be limited: perhaps it is in a different language, uses a different spelling or date format: Of the 11.4 million records of biological samples stored in NCBI and EBI genomic databases most do not adhere to any standardisation [29]. To work around these standardisation issues, metadata can be specified in a 'metadata standard'. A metadata standard is specified in such a way that its validity can checked by a computer. When such a standard is available, it is important for biologist to use such a standard.

For a data point to be described well, it must be unique, and clearly named [34]. This can be achieved by giving each record an uniform resource identifier (URI). There are various kinds of metadata to describe a datapoint [42]: (*i*) Unstructured metadata like free text, or structured terms from a controlled vocabulary. (*ii*) Automatic metadata such as last edit date, or manually asserted, such as the title of a document. (*iii*) Textual, like the name of author, or numeric, such as the depth at which a soil sample was taken. Between these types of metadata there is a trade off, free text values are more likely to be filled in, but are difficult to analyse with an automated system.

2.1.2 Popular metadata standards

To make metadata more standard, there are various systems for various types of metadata. Basic bibliographic metadata, such as author information and publishing details, may be represented using schema.org [62]. Basic datatypes like text, numbers and dates can be indicated using the XML schema [17]. Geolocation data, such as surfaces, points, routes &c can be encoded using geoSPARQL protocol and RDF query language (geo-SPARQL) [52].

The JERM standard [35] In the just enough results model it is possible to describe aspects of a biological investigation, such as a strain that was studied, or technique with which a sample was measured.

The plant phenotyping experiment ontology (PPEO) standard The PPEO is a minimal metadata model specifically for plant phenotyping experiment by Ćwiek-Kupczyńska *et al.* [20]. This ontology was specifically designed for keeping track of the interactions between genotype and the environment often encountered during plant phenotyping studies [50].

The MIxS standard [87] The MIxS standard allows for the description of sequencing experiments where samples are taken from various types of environments such as human skin, or soil. It also allows for description of genome sequences from natural and artificial sources. For each type of sequence, there is a minimum information that must be provided. For example, for a soil sample, one must provide sampling depth in metres. The MIxS standard is continuously updated, with the last update being MIxS symbiont associated (MIxS) [37] in 2022.

The ISA standard [60] allows for description of an experimental workflow that includes an Investigation, samples, and assays. It overlaps with the MIxS standard for a large portion. The ISA format uses subsets of MIxS for certain contexts.

2.2 FAIR data in biology

Data management can be coordinated using the findable, accessible, inter-operable & re-usable (FAIR) principles [82]. These principles are not ridged points, but more guide lines to help people determine if their data is usable by others. The FAIR principles are [82]:

Findable Information must be indexed so that it can be searched through. This may be achieved using a database.

Accessible The data must be available though a programme that is open and available everywhere. Data can be kept private and accessed by a select few only, but even then that programme should work on every computer. This may be done by allowing access through a standard query language, such as SPARQL protocol and RDF query language (RDF).

Interoperable Data must be easily combined with other datasets by using formal and shared terms. For example, terms from an ontology may be used for field names in the databse.

Reusable Licensing is permissive enough to allow data usage. The data must be described well enough, and community quality standards must be met. For example using an creative commons attribution licence [18].

When the FAIR data is available on the web, more technologies (Figure 1) come into the picture. Using ontologies and taxonomies the meaning of terms and how they relate can be stored in an accessible manner, for both researchers and the machines they use. Since FAIR data is supposed to be machine readable and flexible. The RDF format is a good carrier of such data. Since it is a graph format, it handles connections between data well. Missing data is handled by simply not including it in the file. Various data types can be handled: from time series to facts about a person. The RDF file can be indexed and searched using a graph database. Graph databases are a popular method in biology to expose knowledge to researchers. A list of the most important graphs is maintained by Yummydata [85]. Notable graphs include (*i*) Allie [1] that holds information on lifescience acronyms; and (*ii*) Rea [6], a database of biochemical reactions.

2.3 Data standardisation programmes

There are plenty of research groups that made programmes for FAIR management of laboratory data, examples include:

SEEK by Wolstencroft *et al.* [84] a database that allows users to share templates and data using those templates.

COPO by Shaw *et al.* [66] is a programme that can manage and share data online.

BioShare by McQuilton *et al.* [47] is a similar programme to COPO, but it went out of since initial development in 2016.

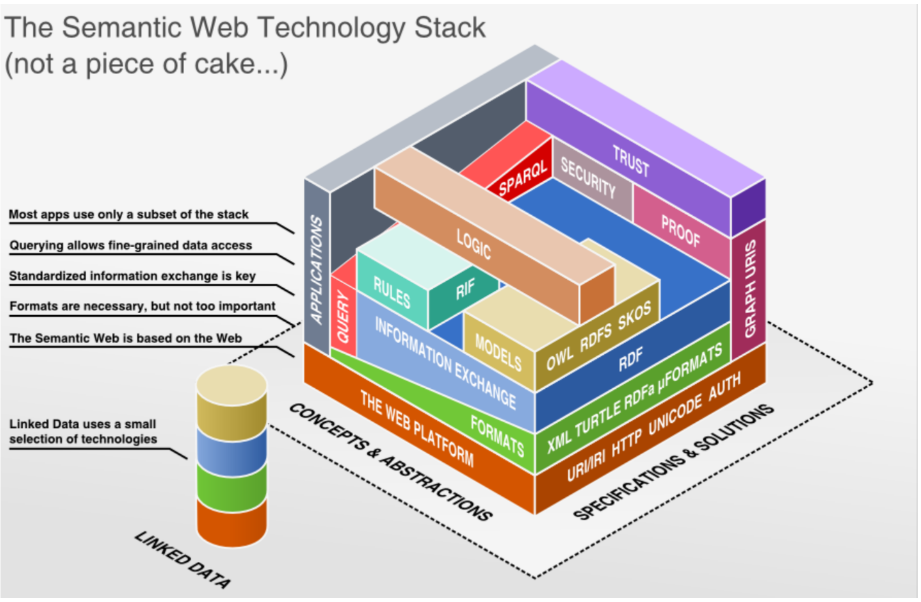

The Semantic Web Technology Stack
(not a piece of cake...)
Most apps use only a subset of the stack
Querying allows fine-grained data access
Standardized information exchange is key
Formats are necessary, but not too important
The Semantic Web is based on the Web
Linked Data uses a small selection of technologies
LINKED DATA
APPLICATIONS
QUERY
RULES
RIF
LOGIC
SPARQL
SECURITY
TRUST
PROOF
GRAPH URIS
MODELS
OWL RDFS SKOS
RDF
INFORMATION EXCHANGE
FORMATS
XML TURTLE RDFa µFORMATS
THE WEB PLATFORM
URI/IRI HTTP UNICODE AUTH
CONCEPTS & ABSTRACTIONS
SPECIFICATIONS & SOLUTIONS

Figure 1: Overview of the technologies central to FAIR linked open data. The fundament of web technology are technologies for transferring files (HTTP) saving information (unicode). Using such technologies, formats (such as RDF-turtle, see Appendix D) can be created for information exchange. Figure made by Bazzanella and Tzitzikas [9].

3 Methods

3.1 Implementing an ASPAR_KR prototype

For a central FAIR data base like the envisioned ASPAR_KR, there needs to be a way of FAIRifying data. Requirements of such a method have been determined via one-to-one interviews with prospective users ($n = 5$). In each interview, users were asked to show their current data-entry-to-analysis workflows. Additionally, a survey of data types and laboratory results was made by looking at the raw data of each researcher. Notes of the interviews are used to determine commonalities in data storage and analysis methods. Since the *A. fumigatus* domain is unfamiliar to FAIR practices, simplicity was emphasised in the requirements.

A literature search for standardisation programmes was made using the search term 'fair data management software biology' in google scholar. Besides a literature search, data management advice and software recommendations from the Wageningen data competence centre were considered. Requirements and FAIRification programme candidates were presented to the community at monthly meetings to obtain feedback. After the fairification programme was chosen, it was used to gather data for a central FAIR data prototype in a series of use cases. A prototype of a central database was implemented with the guidance of the Wageningen university IT department.

3.2 Applying the programme to use cases in *A. fumigatus*

Use cases from the *A. fumigatus* were examined. This domain was examined as the differing expertise levels between researchers in this field are typical for a life science profession: laboratory researchers use excel and only a small portion centre their workflow on scripting languages such as R. To find use cases from the *A. fumigatus* domain, laboratory biologists from the WUR genetics department were invited. The use cases were selected to ensure that they were representative of the different types of work done in the domain: fundamental biology in the laboratory, field trails, and environmental surveys. Data from use cases are available on git.wur (Appendix B-3), publicising data was done in agreement with the users.

To determine the needs of each use case, interviews were conducted, where each user was asked to provide an outline of their research topic, questions and methods. Protocols and raw datasets were also investigated. During the interview, the standardisation problems with the user their current data management practice were discussed. Together with the user, a new standard of data entry was created: first a draft was made which was shown to the user for feedback. Based on the feedback, the draft was changed. These rounds continued until there was a minimally useful standard. The standards were represented as unified modeling language (UML)[1] diagrammes. The standards were applied on new data the user had available. If needed, R-scripting was used to re-shape the dataset into the new standard.

The fuseki 2 database programme by Apache software foundation [4] was used to store and locally serve the FAIR data produced using the new standards. Promising analysis questions were discussed with the user and written to SPARQL. Data from the queries were analysed using R.

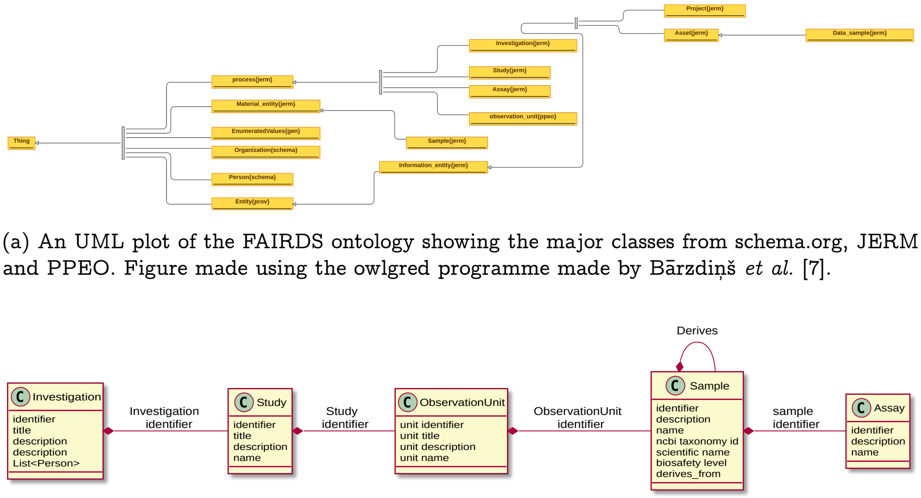
4 Results and discussion

4.1 Requirements of the *A. fumigatus* community

To determine how to promote FAIRification in the life science domain the field of *A. fumigatus* was studied. To support *A. fumigatus* researchers in datasharing and analysis, an online platform was prototyped: the *A. fumigatus* azole resistance knowledge resource (ASPAR_KR). There are two parts to such a datasharing service: collecting standardised data, and serving said data. Starting with the data collection method, requirements were collected with a series of interviews. The interviews are available in Appendix B-2.

During the interviews it was observed that the data literacy of *A. fumigatus* research community is heterogeneous. Most people use excel to analyse their datasets and a small subset uses scripting languages such as R and flat file formats (see Appendix D). There were issues aggregating data, as the researcher who were interviewed all used different layouts for their xlsx sheets. Sadly, a conclusive view of data types and data management practices could not be obtained, as no systematic survey was conducted into this topic. Not all researchers were familiar with computers or had a desire to install additional software packages for their data management besides excel.

[1] https://sparxsystems.com/resources/tutorials/uml/datamodel.html



(a) An UML plot of the FAIRDS ontology showing the major classes from schema.org, JERM and PPEO. Figure made using the owlgred programme made by Bārzdiņš *et al.* [7].

(b) Simplified schema of the FAIRDS vocabulary.

Figure 2: Datasets made using the FAIRDS model the experimental design by including the relation between Observation units, Samples and Assays.

4.2 Implementing the ASPAR_KR prototype

4.2.1 The FAIRDS-data station (FAIRDS)

The FAIRDS is a programme written by Nijsse, Schaap, and Koehorst [51] to make the production of FAIR data files easier. They developed the FAIRDS to support the unlock project. The unlock project[2] aims to understand how microbial communities behave, and to see if they can be altered to improve a biotechnological process. Examples of an unlock projects is the pig-paradigm[3] which aims to find ways to prevent intestinal infection in piglets. They aim to achieve this by understanding the pig gut microbiome better. The most important feature of the FAIRDS data format is that it models the structure of an experimental design (Figure 2b). The FAIRDS tool builds on the ISA framework with some changes to clarify the terms 'sample material' and 'source material' which Nijsse, Schaap, and Koehorst [51] found confusing to users of the FAIRDS: The 'observational unit' from PPEO is used to replace the 'source material' and 'sample material' is replaced by 'sample' from JERM.

The FAIRDS targets itself to users without much computer knowledge, since it uses `xlsx` for the initial data entry, users do not have to download a programme to use it. An `xlsx` file of the FAIRDS format, contains a sheet (Figure 3b) for each of the 5 classes that is relevant to keep track of in a wet lab dataset (Figure 2b):

Investigation Here the overarching question that the dataset is trying to answer is recorded. As well as bibliographic information such as author information.

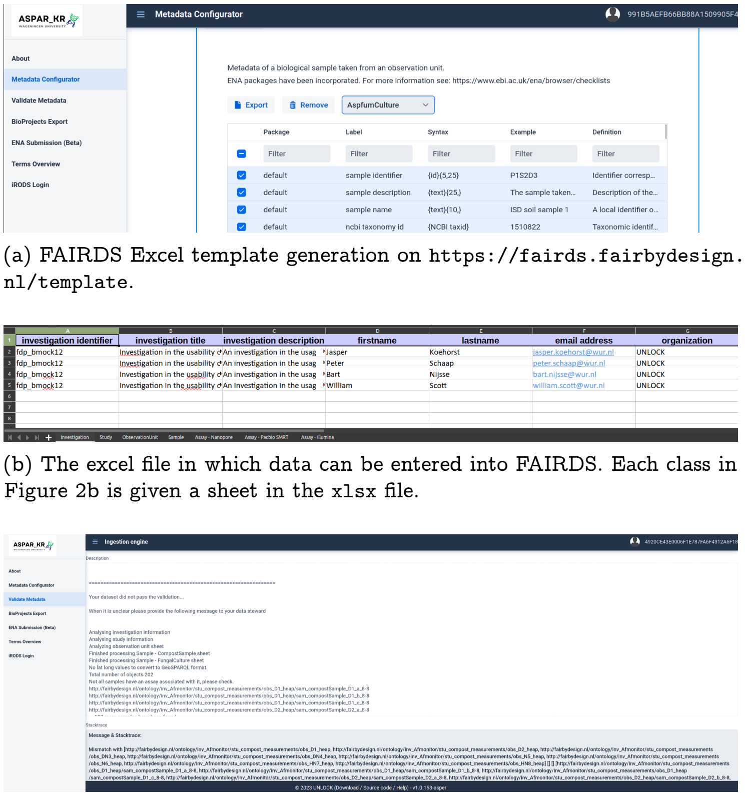
[2] `https://m-unlock.nl/`
[3] `https://projects.au.dk/pig-paradigm`

Study Here all the sub-questions (or sub datasets) that are noted with a name and a description.

Observation unit An observational unit is the thing on which observations are made and statistics are compiled [28].

Sample Here, samples and direct sample observations are recorded. For example, a Petri dish with a mould on it is a Sample, and the description of its morphology can be recorded with it.

Assay Here users may note their measurement data. For example the download link to another dataset such as sequencing reads, or a qPCR result can be recorded here.



Metadata of a biological sample taken from an observation unit.
ENA packages have been incorporated. For more information see: https://www.ebi.ac.uk/ena/browser/checklists

	Package	Label	Syntax	Example	Definition
☑	default	sample identifier	{id}{5,25}	P1S2D3	Identifier corresp...
☑	default	sample description	{text}{25,}	The sample taken...	Description of the...
☑	default	sample name	{text}{10,}	ISD soil sample 1	A local identifier o...
☑	default	ncbi taxonomy id	{NCBI taxid}	1510822	Taxonomic identif...

(a) FAIRDS Excel template generation on `https://fairds.fairbydesign.nl/template`.

	A	B	C	D	E	F	G
1	investigation identifier	investigation title	investigation description	firstname	lastname	email address	organization
2	fdp_bmock12	Investigation in the usability c	An investigation in the usag	Jasper	Koehorst	jasper.koehorst@wur.nl	UNLOCK
3	fdp_bmock12	Investigation in the usability c	An investigation in the usag	Peter	Schaap	peter.schaap@wur.nl	UNLOCK
4	fdp_bmock12	Investigation in the usability c	An investigation in the usag	Bart	Nijsse	bart.nijsse@wur.nl	UNLOCK
5	fdp_bmock12	Investigation in the usability c	An investigation in the usag	William	Scott	william.scott@wur.nl	UNLOCK
6							
7							
8							

(b) The excel file in which data can be entered into FAIRDS. Each class in Figure 2b is given a sheet in the `xlsx` file.

```
=================================================================

Your dataset did not pass the validation...

When it is unclear please provide the following message to your data steward


Analysing investigation information
Analysing study information
Analyzing observation unit sheet
Finished processing Sample - CompostSample sheet
Finished processing Sample - FungalCulture sheet
No lat long values to convert to GeoSPARQL format.
Total number of objects 202
Not all samples have an assay associated with it, please check.
http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap/sam_compostSample_D1_a_8-8
http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap/sam_compostSample_D1_b_8-8
http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap/sam_compostSample_D1_c_8-8
http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D2_heap/sam_compostSample_D2_a_8-8
```

```
Message & Stacktrace:

Mismatch with [http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D2_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements
/obs_DN3_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_DN4_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_N5_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements
/obs_N6_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_HN7_heap, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_HN8_heap] [] [] [http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements
/obs_D1_heap/sam_compostSample_D1_a_8-8, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap/sam_compostSample_D1_b_8-8, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D1_heap
/sam_compostSample_D1_c_8-8, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D2_heap/sam_compostSample_D2_a_8-8, http://fairbydesign.nl/ontology/inv_Afmonitor/stu_compost_measurements/obs_D2_heap/sam_compostSample_D2_b_8-8,
```

© 2023 UNLOCK (Download / Source code / Help) - v1.0.153-asper

(c) FAIRDS validation interface. `https://fairds.fairbydesign.nl/validation`.

Figure 3: Data entry system for the FAIRDS. First users make an excel template (Figure 3a), then they fill in the excel sheet (Figure 3b) and upload it to the validation service (Figure 3c).

In each work sheet of the data entry excel file, each term has its own column. These are either optional, or required. If a user does not fill in a required column or the patterns do not match, the data is not deemed of sufficient quality and will be rejected. Users can also include column names not yet known by the programme, these columns will be processed as well and added to the RDF file without undergoing any checks. When the user has prepared the xlsx file, they can upload it to the FAIRDS validation programme. The validation programme checks the dataset for errors, if no errors are found, the data is parsed to a FAIR format: RDF. This RDF file can then be loaded into a graph database, and used for data analysis by requesting information using SPARQL. While the FAIRDS is available on-line, it can also be downloaded and ran on a locally, meaning that it can be used also with sensitive data that cannot be uploaded to a server. As output, the FAIRDS generates RDF files with a graph of the user their data. Extending the FAIRDS vocabulary is also done by editing a xlsx sheet. Optionally, this data can contain links to other data sources, such as sequences, and other raw files. These files can be hosted on open databases such as ENA, or institutional cloud services such as YAML observations data Archive (YODA).

All in all, the FAIRDS is suitable for use in the ASPAR_KR project because of its simplicity in usage. Furthermore, it fits the requirement of being oriented towards xlsx users.

4.2.2 The ASPAR_KR prototype

The RDF files created with the FAIRDS were served with the fuseki 2 graph database[4] by Apache software foundation [4]. This database was chosen for its geo-SPARQL support. The fuseki 2 database was also used for serving the RDF files for the use cases (see next subsection 4.3). Using the FAIRDS as a basis for FAIRification and fuseki 2 as an hosting service, a prototype web-app was build. The prototype uses the engine-x (NGINX) webserver by Reese [58] to forward the user their requests to either the database or the FAIRDS (Figure 4). The web service is deployed as a docker network [48]. Using the cURL programme by Hostetter *et al.* [33]–available on most computers, the database part of the web service can be accessed. Source code to the web service is available in Appendix B-1.

4.3 Collecting data from each usecase

During development of the ASPAR_KR prototype, potential users were contacted in the Genetics department of Wageningen university. During interviews with each user, a data management plan was set up where the FAIRDS plays a central role.

4.3.1 SchimmelRadar citizen science Air sampling

The schimmelradar project [63] aims to establish the background resistance fraction of *A. fumigatus* spores in the air at more than 300 locations in the Netherlands. These samples are taken according to Kortenbosch *et al.* [40]: spores land on a sticky surface with in a ‘delta trap’. The sticky strip is used in a two layer fungal culture to test antibiotic resistance. First a minimal medium is added, and the spores are allowed to grow. This represent the ‘total spore count’. Then a medium with antifungals is added, the spores that grow here are ‘resistant‘. Using the ‘total’ and ‘resistant’ spore count, a resistance fraction will be determined. Besides this, the number of viable

[4] Latest version from https://hub.docker.com/r/stain/jena-fuseki

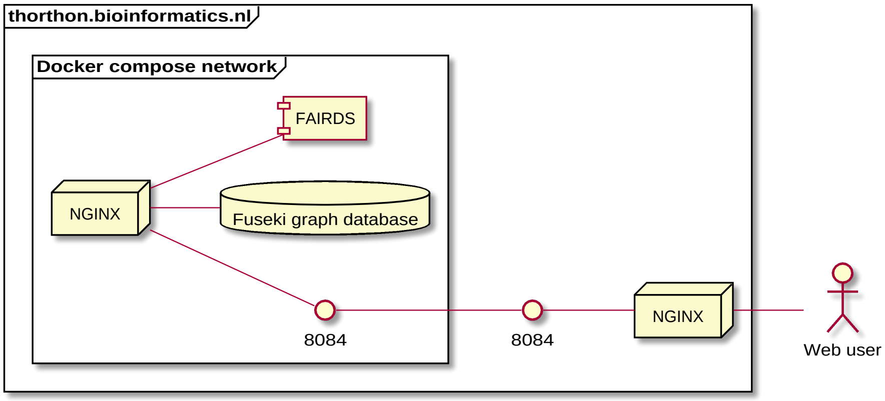


Figure 4: The architecture of the ASPAR_KR webservice. A request from the user is send to the NGINX programme of the thorthon server which sends it to the 8084 outer port of the network. The outer port connects to the inner 8084 port at the ASPAR_KR NGINX service. This NGINX service forwards FAIRDS request to an FAIRDS instance, and the database request to the fuseki database endpoint.

spores in total–without fungicides–is also assessed using a separate sticky strip. The coordinate location of the sticky strip is also recorded. Using this information, the schimmelradar team is interested in determining whether variance in resistance fraction can be explained by spacial differences.

Standardising the schimmelradar data with the FAIRDS would be used for geo-SPARQL queries and for publishing the dataset. The standardisation plan was made together with Hylke Kortenbosch, who supplied a set of test data to standardise. Geolocation data of each Dutch address was requested from google using via GeoPy [26]. Using the FAIRDS the data was standardised according to a schema drawn up with Hylke Kortenbosch (Figure 5). The RDF file was loaded into an Apache Jena geo-SPARQL database. Using R, queries to the database were made and analysed.

To obtain good results with the air sampling method: it is important that the citizen scientist expose the seals for around 30 days, and that they do not take too long to arrive at the laboratory. This can be determined with a geo-SPARQL query made from R (Appendix B-3): each experimental unit is queried for its location, experimental duration and transfer duration. Calculation of duration in second are automatically handled by fuseki 2. The DAY function is used to convert the duration in seconds to days. Additionally, the distance to Wageningen is also requested from wikidata. Here the functions provided by geof and uom are used to calculate distances between Points and convert it to kilometres.

```
PREFIX schema: <http://schema.org/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX uom: <http://www.opengis.net/def/uom/OGC/1.0/>
```

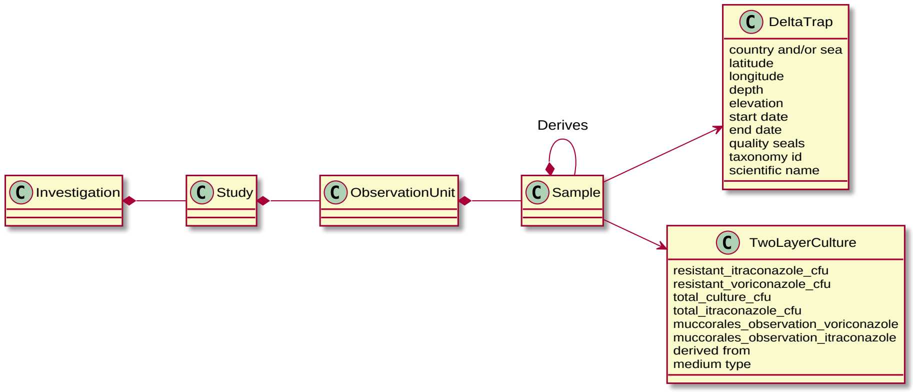


Figure 5: Schema of schimmelradar minimal information. The schimmelradar uses a delta trap and a 'two layer' culture. These experiments have been made subclasses of Sample.

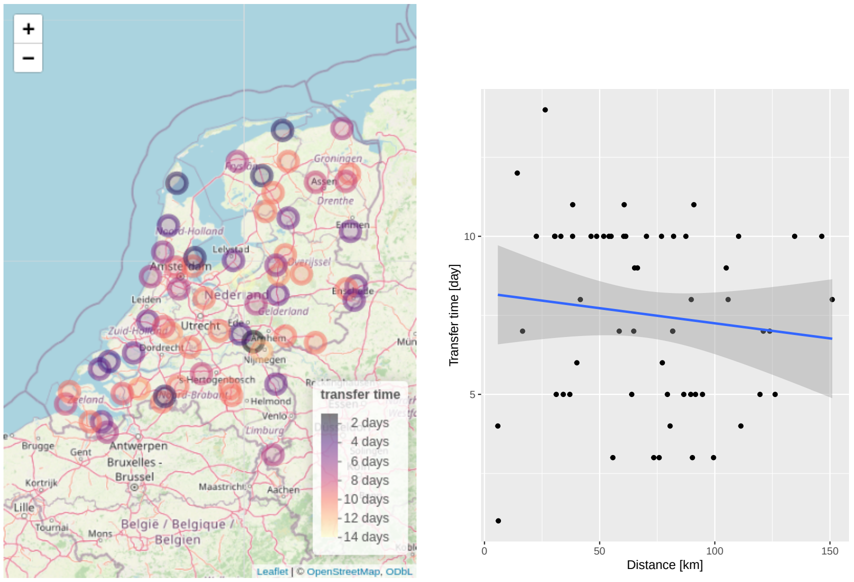
```
PREFIX fair: <http://fairbydesign.nl/ontology/>
PREFIX jerm: <http://jermontology.org/ontology/JERMOntology#>
SELECT *
WHERE {
    # Get properties of the DeltaTrap.
    ?deltaTrap fair:packageName 'DeltaTrap' ;
        <https://w3id.org/mixs/terms/0000011> ?arrival_date ;
        # the point that connects to the DeltaTrap.
         geo:hasGeometry/geo:asWKT ?point .

    # From which subset of the Netherlands does the sample come from?
    ?observationalUnit jerm:hasPart ?deltaTrap ;
        schema:identifier ?obsId .

    # Get properties of the culture
    ?deltaTrap fair:derives ?twoLayerCulture ;
        fair:start_date ?start_date ;
        fair:end_date ?end_date ;

    # What is the amount of time the seals were exposed?
    BIND(DAY(?end_date - ?start_date) AS ?air_exposure_days)
    # What is the transfer time?
    BIND(DAY(?arrival_date - ?end_date) AS ?transfer_time)

    # Distance from Wageningen
    SERVICE <https://query.wikidata.org/sparql> {
        # What things are a municipality?
        ?municipality wdt:P31 wd:Q2039348.
        # What things have a place?
```



(a) Transfer time for each air sample to the WUR laboratory.

(b) Relation between transfer time and distance

Figure 6: Relation between distance to the WUR laboratory.

```
        ?municipality wdt:P625 ?placeOfInterest .
        # Take only the thing that is the place of interest.
        FILTER(?municipality = wd:Q1305) . # Wageningen
    }

    # Distance between our place of interest (WUR) and every DeltaTrap.
    BIND(geof:distance(?point, ?placeOfInterest, uom:kilometre) AS ?d_km)
}
```

After requesting the information as a comma separated values table from fuseki 2 using httr [80], a dataframe is constructed with each variable in the query being a column.[5] The dataframe is converted to a sf object using tidyverse [81] and sf [54, 55]. Leaflet [30] is then used to plot the location of each delta trap on a map, and ggplot2 [79] is used to plot the scatter. These plots show that the average air exposure is 28 days and that the average time in the postal service is 8 days. There is no evidence for a relation between geographic location and transfer time to the WUR laboratory (Figure 6). The calculations to determine these relationships could be done with scripting languages, but the ease of requesting the data and derived calculations from a single database makes the use of a graph database a worthwhile option.

[5] In SPARQL, a variable is prefixed by the ? sigil.

4.3.2 MIC phenotyping data for Δ*pef* mutants

For the structure of the cell membrane, the PEF protein plays an important role. Martin Weichert is a cell biologist interested in researching this protein in a FAIR manner. To investigate the PEF protein, Martin made knockout mutants (Δ*pef*) using clustered regularly interspaced short palindromic repeats-CAS-9 (CRISPR-CAS-9). Using resazurin he determined the the minimal inhibitory concentration (MIC) value with a photospectrometer. Resazurin is a non-fluorescent molecule that can be converted by live fungal cells to a fluorescent derivative.

Together with Martin Weichert, a data management plan and minimal metadata package was drawn up. The data management plan consists of a git.wur repository where files from a photo spectrometer can be uploaded together with the metadata–all in comma separated files. An `R` script reads all the data into a set of `R` objects. These `R` objects can then be parsed to a FAIRDS compatible `xlsx` file. The git repository is available in Appendix B-3.

The schema for Martin his measurements was made in accordance to what metadata were both relevant biologically and relevant for reproducibility (Figure 7). The CultureCollection class is important for people who want to reproduce the dataset, as it gives contact information about the set of fungi that was tested and where samples can be requested. Metadata such as inoculum spore concentration is relevant as inoculation density can affect fungal growth [56]:

Querying the data can be done with a SPARQL query such as the one below. The query shows that SPARQL is also compatible with regex, this makes it possible to transform text strings that are in the database.

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX uom: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX fair: <http://fairbydesign.nl/ontology/>
PREFIX jerm: <http://jermontology.org/ontology/JERMOntology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?name ?storage ?measureParts ?fluor ?c ?molecule
       ?file ?fluorControl ?controlParts ?cC ?concentrationControl
       ?unit
WHERE {
    # Get the sample units from the database
    ?sample rdf:type jerm:Sample ;
        schema:name ?name ;
        fair:packageName 'Strain';
        fair:storage_date ?storageDate .
    ?control rdf:type jerm:Sample ;
        schema:name ?nameSample ;
        fair:packageName 'Controls';
        schema:identifier 'sterile' .
```

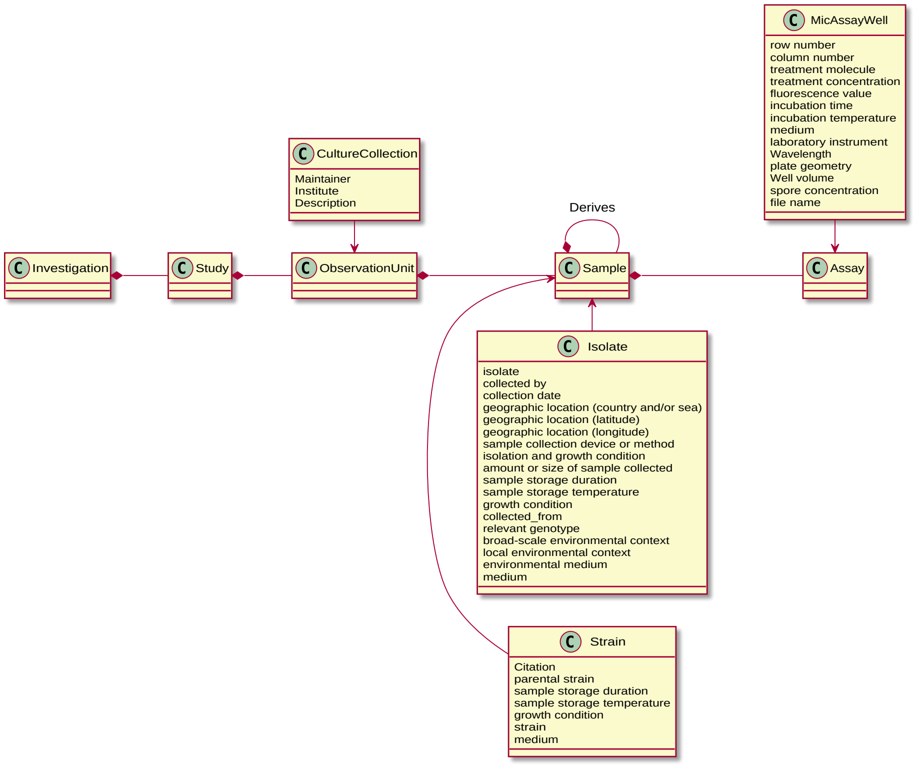


Figure 7: Minimum information for the MIC test devised by Martin. For his experimental data it was needed to define a CultureCollection class to store information on what fungi collection are analysed. Isolate and Strain were newly defined samples. The samples are analysed with an MicAssay, a photospectrometer based assay of MIC on a wells plate.

```
    ?control jerm:hasPart ?controlParts .
    ?controlParts fair:value ?fluorControl ;
        fair:treatment ?concentrationControl ;
        fair:file_name ?fileControl ;
        fair:treatment_molecule ?moleculeControl .

    ?sample jerm:hasPart ?measureParts .
    ?measureParts rdf:type jerm:Assay ;
                  schema:dateCreated ?assayDate ;
                  fair:packageName 'MIC' ;
                  fair:value ?fluor ;
                  fair:file_name ?file ;
                  fair:treatment_molecule ?molecule ;
                  fair:treatment ?concentration .

    # We only take files that describe the 48h timepoint.
    FILTER(regex(?file, '48h')).
    FILTER(regex(?fileControl, '48h')).
    # Extract the number part from a string such as
    # 12 mm.
    BIND(xsd:double(REPLACE(?concentration,
        '(.+) (.+)', '$1')) as ?c)
    BIND(xsd:double(REPLACE(?concentrationControl,
        '(.+) (.+)', '$1')) as ?cC)
    BIND(xsd:double(REPLACE(?concentrationControl,
        '(.+) (.+)', '$2')) as ?unit)
           BIND(xsd:date(?assayDate) AS ?date)
    BIND(day(?storageDate - ?date) AS ?storage)
}
```

By applying a few extra transformations on the data obtained from the query, a plot of anti-fungal activities is obtained (Figure 8).

4.4 Prototype of ASPAR_KR

4.4.1 Effectiveness of FAIRDS as a standardisation method

Effectiveness of a programme is judged best by seeing how well people are able to use it independently from the developer. On this front, the FAIRDS performs poorly, as researchers have not adopted the xlsx templates (Figure 3a) *independently* and have not used them for data entry in the laboratory–as is the original design of FAIRDS. When asked for their decision not to adopt the template in their laboratory work, researchers in the *A. fumigatus* domain state various reasons: (*i*) It conflicts with their current workflow. (*ii*) It would take extra time. (*iii*) Data sharing is not seen as a priority. This apparent unease of users to employ the FAIRDS for data standardisation could be solved by better integrating the FAIRDS in their current workflow. Integrating the FAIRDS into elab electronic laboratory management system is being developed

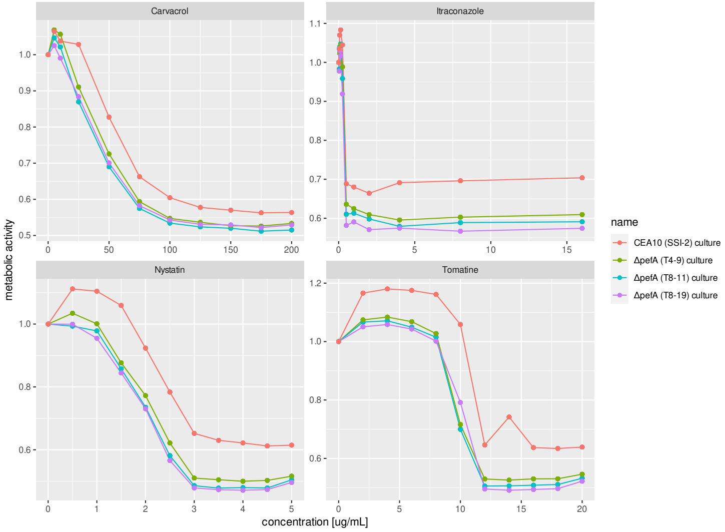


Figure 8: Results from Martin his MIC assay.

by Jasper Koehorst. Use of elab labjournals is mandated in the genetics department. Another potential source of difficulty in FAIRDS adoption is the RDF format produced by the programme. While RDF is useful for the informatically inclined, biologists lack the knowledge to deal with this data format, since it requires the user to know SPARQL. This is the major difference between the usecase at the synthetic systems biology (SSB) department and the usecase of the genetics (GEN) department: at SSB, there are bioinformaticians that are able to incorporate the FAIRDS files into an automated pipeline, at GEN, most biologists are responsible for their own data analysis.

Besides user experience, experience of the developer, who is supposed to implement new features into `ASPAR_KR` is also important, especially for maintaining the programme over a longer period of time. On this front, the FAIRDS also performs poorly, as it is a programme that does not have a design that is oriented towards easily fixing errors and implementing new features. While FAIRDS is an open source programme, it is not likely that external opensource developers will contribute code, as there is no good documentation available within the codebase. This issues could all stem from a lack of clearly defined roles for people working on the FAIRDS project [36].

While the FAIRDS allows for easy conversion of format and basic syntax checking, it is not fully rigorous in its ability to validate. For example, there is no check to determine which combination of experiments actually makes sense, for example, samples that are not microbial samples, such as a histological sample of a patient, can still have an associated MIC assay. Currently, the FAIRDS has no way to easily check these relations besides stating them explicitly within another sheet of an `xlsx` document. Making the ontology of FAIRDS able to support shapes constraint language (SHACL) validation would make checking arbitrary patterns in the data set more ergonomic. The use of `xlsx` as a data carrier for validation data has another drawback: `xlsx` is a format with quite arbitrary restrictions, a sheet name in Excel can only be 31 characters long. This puts a limit on what package names are possible.

4.4.2 General recommendations for improvement

Promotion of FAIR data literacy is a complex issue. Besides the technical issues encountered, FAIR implementation commonly has to deal with cultural issues as well [83]. The lack of expertise in data sharing; limited funding; fear of having your research 'scooped' are the main reasons why researchers have troubles sharing data in a FAIR manner [11]. By providing a service to FAIRify data *without* automatic upload to a public database, researchers who use the FAIRDS can keep their data FAIR yet private from the start of their research. With the use of excel sheets, FAIRDS sets itself apart from other FAIR data solutions such as COPO [66] and BioSharing [47].

The issues that are encountered when applying the FAIRDS to a new domain can be overcome. Improved incentives for using the FAIRDS together with improved user-friendliness of the FAIRDS would help to speed up adoption. Two methods to achieve these goals are by creating automated analysis workflow for data in the FAIRDS format, and to make the FAIRDS more closely aligned to the user's workflow with elab integration. The acceptance of the technology will be eased if there would be a committee of senior researchers that decide on standards for common types of experiments in the field. Additionally, collaborating with researchers who are working on the development of new methods, such as Martin Weichert and Hylke Kortenbosch is a benefit: not only is their data is more FAIR, the templates that result from their work, could be used by anyone who wants to replicate their work. By releasing a `R` script that can take the data produced with Martin his template and analyse it directly, adoption would be incentivised. This would also solve the issue of biologists not being able to work with SPARQL them selves, as the `R` scripts may automate this process. Finally, the ability to link data files to a FAIRDS excel sheet, currently requires a YODA account from the WUR library, a better system needs to be put in place so that file upload is also possible for external collaborators.

The experience of the developer can also be alleviated by paying more attention to software design principles and software design techniques. Sources such as refactoring guru [59] programming language agnostic guides to make code more maintainable. Adding more functional test cases will also help developers, as this provides both example on how to use the code, as a detection mechanism for errors Basili and Selby [8]. Furthermore, by giving each contributor a clear responsibility and making it clear how new code will be placed in to the codebase, conflicts and misunderstandings can be avoided. This would make it easier to add new features to the FAIRDS and increase the possibility of third party contributors joining the project. Additionally, converting the FAIRDS vocabulary from `xlsx` to an ontology in RDF, implementing pattern checking and validation logic can be made easier for the developer using SHACL.

5 Conclusion

Over the course of this project, an existing programme to generate FAIR data was applied in a new domain. This programme, the FAIRDS, was originally designed for system biologists by Nijsse, Schaap, and Koehorst [51]. During my year of work on promoting FAIRification of *A. fumigatus* data, no widespread change in data management culture could be achieved. Sadly, such a lack of progress is not uncommon: Verbeke [77] build a FAIR and open database for researchers in the field of water purification which still did not see any active submissions for a year after its release. The main reason for the lack of FAIRDS adoption is that the needs of the system biology department are different of the mycologists: in the system biology department, biologists manage and analyse data with help from bioinformaticians; the mycologists at the genetics department are often responsible for their own data management and analysis, making the use of specialised technologies, like RDF hard to promote.

If researchers would be able to use the FAIRDS–or a similar tool–to make their data FAIR from the planning stages on using data entry templates. Research groups would be more productive as data from past research projects would searchable and understandable. This has a clear monetary value: less work would be repeated. Going further, the databases from each research group can be combined into a domain specific database. In the case of *A. fumigatus*, such a database would be the envisioned ASPAR_KR platform. Having an online FAIR data repository which is curated by domain experts would increase the quality of scientific outputs: (*i*) The data of a study would be specifically reviewed. (*ii*) There would be high quality and trusted datasets available for the community that can be used to inform new experiments or to do meta analysis. Additionally, such a dataset could specifically allow the curated entry of unpublished data, so researchers can also communicate what 'did not work'. Realising such a platform is a great challenge, both socially and technically. Socially, there needs to be a change of mind set that a dataset is valuable on its own–like a publication, and that the data also belongs to the community at large. Technically, there needs to be a server that is capable of hosting all the data from the domain, and a programme capable of entering/serving it. Hosting a community SEEK instance on a university server is good start.

Funding for such a database is difficult to arrange, as there are many steps that need to be completed until there is a 'critical mass' (Figure 9): there need to be committed scientists that gather good (meta)data, then there must be trust in the organisation to host all the data, lastly there needs to be a stable source of funding. Asking users for money is not practicable: it must be freely and openly accessible for all, or it would violate Openscience principles. Getting funds and support for running a database is a long standing issue [19]: The database must be run by an institution with a suitable mandate, be well integrated with other trusted databases, and must motivate scientist to edit and update it. For ASPAR_KR, this might mean collaborating with FUNDB. In case funding cannot be found, or dries up along the way, data can still be kept save by migrating to larger (uncurated) open science platforms (like FAIRDOM-hub) [89].

This work showed that FAIRifying data in the *A. fumigatus* domain by relying on the researchers own initiative does not lead to direct success; as researchers are not able to FAIRify data them selves with simple programmes, like the FAIRDS. To achieve the goal of FAIRification, the following additional components may be needed: (*i*) FAIRification must above all be simple and cannot ask the user to make changes to their work flow too much; and (*ii*) There must be a better incentive structure for FAIR data. Data citation, the practice of considering data to be a publication on itself [67] is hypothesised to be this 'missing incentive'. Data citation however, may not yet be ready for this purpose [11, 70]: data citations should be considered with the same

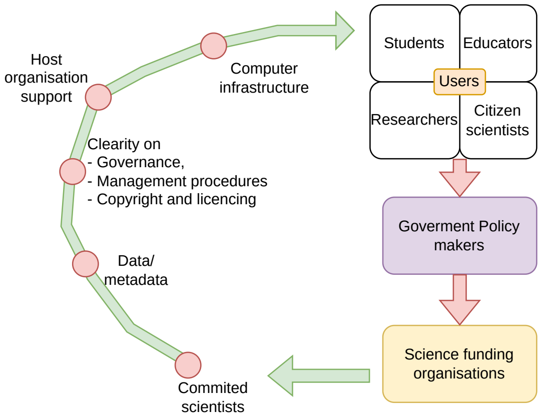


Figure 9: Steps that are needed for funding acquisition of a open science database. For a database to get any funding, there must be a community of committed scientists that makes high quality data. Clarity on how the database operates (licensing, governance &c.) will draw researchers to the database. Then there must be enough funds to host the data. Users can then come to the database. An active user base will attract more funding. Figure redrawn from Costello *et al.* [19].

regard as a publication and data should be completely machine actionable. As found in this thesis, machine actionability on data is still a bridge too far, as we are only starting the use of FAIR data.

Despite the difficulties is important to keep working on FAIRification, in this regard the FAIR cookbook by Rocca-Serra *et al.* [61] and the guide by Welter *et al.* [78] make for good reference materials: FAIRification should be specific, incremental, with realistic expectations and a multidisciplinary team eager to FAIRify. This incremental approach can be targeted to a specific part of the *A. fumigatus* domain where standardisation is needed most [25, 68]: genome sequences, as there is currently no way to relate genome sequences of *A. fumigatus* across places and environments in the same way as can be done for bacteria or viruses using platforms like nextstrain [31]. The FAIRDS tool would already suffices for FAIRifying new and existent genome data, as it is able to describe microbial sequences from the environment. Now it rests on the community to sequence FAIRly!

Appendices

A Background information on the *A. fumigatus* domain

A-1 Public health and fungi

Fungi are an important factor in human well being: they are important in nutrient cycling [41], are symbionts of our crops, and are used in fermentation [49]. However, some fungi cause diseases in humans and major societal damage by infecting crop plants [2]. Around 150 000 000 severe fungal infections are occur each year, resulting in about 1 700 000 deaths per year — on the same level as tuberculosis [39]. These severe infections are caused by only four genera of fungi [12]: (*i*) *Aspergillus*, (*ii*) *Cryptococcus*, (*iii*) *Candida*, (*iv*) *Pneumocystis*. Of these fungi, *A. fumigatus* is estimated by European Centre for Disease Prevention and Control [24] to have caused 22 thousand admissions to the intensive care within the European union.

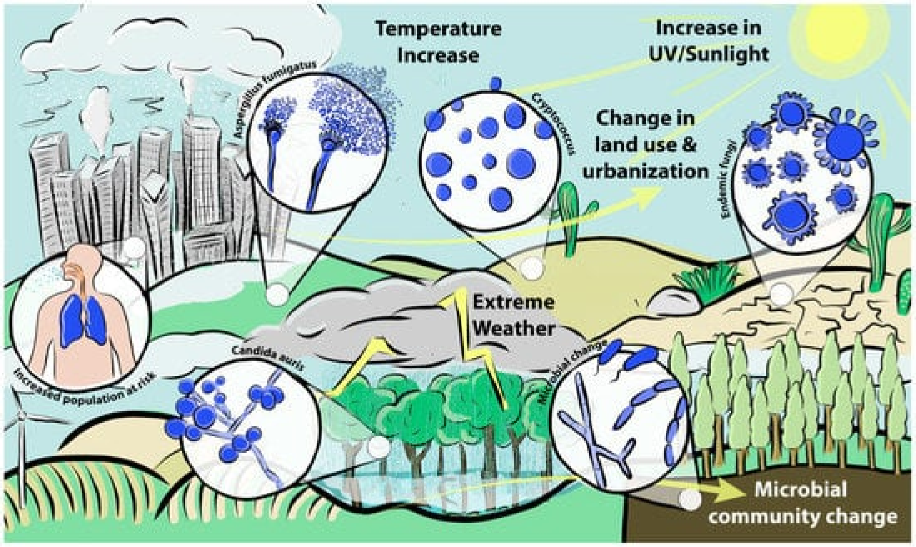


Figure 10: With the changing climate, land use and human demographics, fungal pathogens will be of increasing importance to human health. Figure made by Van Rhijn and Bromley [75].

A. fumigatus is an ascomycete opportunist pathogen, of special interest for its flexibility: it is able to live saprophytically in agricultural waste, and can infect plants and immunocompromised animal hosts [73]. In people with a working immune system, *A. fumigatus* spores are removed by the innate immune system [44]. Since immunosuppressive therapies–used during organ transplant or cancer treatment-are more effective, the number of immunocompromised people is increasing [44]. Depending on the immune status of the individual exposed to *A. fumigatus* adverse effects can include an allergic response, or aspergillosis. When *A. fumigatus* spores enter the lung of an immuno-compromised individual, *A. fumigatus* can cause aspergillosis [15]:

A. fumigatus starts growing in the lungs of the patient, it could even spread outside the lungs and infect other organs. Besides aspergillosis, an allergic response to *Aspergillus* is also dangerous: an allergy to *Aspergillus* triggers inflammation of the lungs, and embolisms of the heart [43].

Compost heaps are an effective substrate for *A. fumigatus* [88]. The heaps are an evolutionary pressure cooker: Because of their high temperatures ($\geq 50\,^{\circ}\mathrm{C}$ not being uncommon), they select for the growth of the more thermophilic *A. fumigatus* [64]. Since compost heaps are formed from plant waste treated with (azole) fungicides, such as tulips, the fungicide pressures are also high. These areas are known as hot spots [64]. This means that *A. fumigatus* isolated from such compost heaps are more likely to be resistant to azoles [64]. From these heaps, *A. fumigatus* may spread to patients. With climate change, *A. fumigatus* spread will become a bigger issue: Van Rhijn *et al.* [76] found that the amount of *A. fumigatus* spores correlates with an increased temperature.

A-2 Azole resistance in *A. fumigatus*

In the late 1970$^{\text{ies}}$, antifungals of the azole class were first brought to market [3]. These chemicals are composed of pentagonic, heterocyclic ring with at least one non-carbon atom in it [3]. Because countries with a more intensive agro-cultural system use more azoles, azole resistance also develops more quickly in *A. fumigatus* populations in these countries [10, 16]. For example, in the NL, azole resistance in clinical *A. fumigatus* isolates increased from $0.79\,\%$ in 1995 to $7.04\,\%$ in 2016 [13]. This increase in resistance has prompted the WHO to put *A. fumigatus* on the list of critically high priority pathogens [53].

Azoles target the CYP51 protein, which plays a vital role in the ergosterol synthesis pathway [27]. Ergosterol plays an important role in the integrity of the fungal cell membrane, by maintaining fluidity [22]. Point mutations in the coding region of the gene, may reduce affinity to azoles, and mutations in the *cyp51a* promoter region may cause over expression (Figure 11). Lastly, it is possible for the efflux transport proteins to be more efficient, or more highly expressed, leading to a detoxification of the cytoplasm. This means that the relation between genotype and phenotype is critical to understanding when *A. fumigatus* may be resistant to azoles and how this resistance has occurred.

Little is known about the geographic spread of azole resistance in *A. fumigatus*: there is a sampling imbalance making it difficult to know where hot spots are absent [14]. Currently, the Genetics department works on the SchimmelRadar project[6] which involves sampling air 300 locations in the Netherlands to investigate spread and azole resistance fraction of *A. fumigatus*. They want to investigate what spacial factors may account for the variance in azole resistance and *A. fumigatus* spore frequency. Specifically, they are interested in determining how agricultural industries affect the azole resistance fraction of *A. fumigatus*. Furthemore, the schimmelRadar team is assessing the *cyp51a* sequence of each azole resistant isolate. These types of investigation will be valuable to better understand the epidemiology of azole resistant *A. fumigatus*. The data from such studies must be available and open for the whole field to use.

[6] `https://schimmelradar.wordpress.com/`

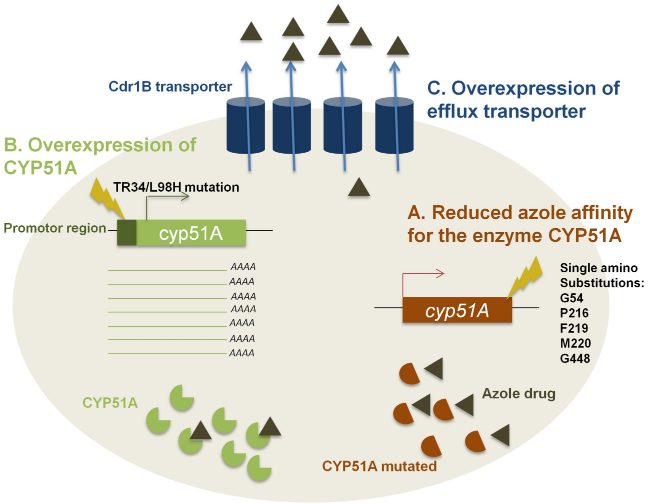

Cdr1B transporter
C. Overexpression of efflux transporter
B. Overexpression of CYP51A
TR34/L98H mutation
Promotor region
cyp51A
AAAA
AAAA
AAAA
AAAA
AAAA
AAAA
AAAA
A. Reduced azole affinity for the enzyme CYP51A
cyp51A
Single amino Substitutions:
G54
P216
F219
M220
G448
Azole drug
CYP51A
CYP51A mutated

Figure 11: Azole resistance in *A. fumigatus* may occur through mutation in the *cyp51a* gene (A & B). Other mechanisms, such as the over expression of efflux pumps (C) can also contribute to resistance. Figure from Berger *et al.* [10].

B Code and data availability

All products of the thesis such as presentations, documents, data, and computer code are available in a documented form on the WUR gitlab servers: `https://git.wur.nl/aspar_kr`. All repositories that do not store data that is under embargo are available for use without authorisation. The files and source code to build this thesis can be found on `https://www.overleaf.com/project/6579cab457f11615b4d278ef`. Repositories of special interested to this thesis are described in the subsections to this annex.

B-1 The `ASPAR_KR` prototype

Help from Rick Janssen to set up the reverse proxy of NGINX is acknowledged (commit `a5074f21`). The docker compose file is available in a git repository: `https://git.wur.nl/aspar_kr/programme/aspar`.

B-2 Survey repository

The one-to-one conversations and interviews about the `ASPAR_KR` project are reported in `https://git.wur.nl/aspar_kr/documents/survey`.

B-3 Usecase repositories

Because of privacy constraints, it was not possible to place all usecase data in the `aspar_kr` repository instead the work for the use cases is spread between the following repositories.

MIC assays of the 'Pef' collection Contains a test dataset of an experimental procedure that will eventually be used to profile the MIC value of each member of the 'Pef' collection. There are `R` scripts to convert the test data in a FAIRDS excel format.
`https://git.wur.nl/aspar_kr/martin-data`.

Schimmelradar air sample measurements This repository contains a subset of the data collected from the schimmelradar project along with scripts to load them into a fusiki 2 database and make queries with SPARQL from `R`. Reach out to Hylke Kortenbosch for access permission.
`https://git.wur.nl/sibbe.bakker/hylke-data`.

C Statement on generative pre-trained transformer usage

Generative artificial intelligence technologies like CHATGTP where employed during development of the `ASPAR_KR` platform. CHAPTGPT was used when dealing with the apache jena library in java and giving instructions for null pointer handling. The model was employed to explain code blocks, and to give suggestions on implementing features. Generative AI was not used to generate the text of this master thesis, or text of the tutorial. Specific issues where a model was used to inform a solution were: (*i*) `Java typecasting`. (*ii*) `formatting cURL requests`. No sensitive data was shared with the model.

D File formats

A file type is a specification that tells a computer how information should be stored and read. There are many different aspects to a file type: (*i*) Openness of the licence, some file formats like `docs` are licenced by companies like MicroSoft. These formats are controlled exclusively by one institution who owns the licence to it. (*ii*) Simplicity, file formats that are just a single text file are simpler than file formats that are collections of text files and images. A single text file is more readily indexed and searched. (*iii*) Whether a file is a binary or not determines whether a user needs a specialised programme to open and edit it.

Not all file types are equally useful in every context. `Xlsx` files [46] for example, are easy for users to edit and use (if they have the excel programme) but have some caveats with their usage. They encode information–like formatting–that are not usually used to encode any data fields. This information makes the file extra large and makes it easier for the file to be corrupted. Furthermore, since an excel file is a collection of separate text files and images with a zipped folder, searching through it is not trivial. This means that within the context of FAIR data, `xlsx` is best suited as a *data entry* tool, where after the contents of the `xlsx` file are converted to an more suitable format.

The RDF data format is an official W3C standard, adopted by modern computer infrastructure. The RDF format, it its most basic form, is a syntax for specifying logical statements composed of triples. This format is simpler than `xlsx`, as only contains text encoding the triples:

```
# A RDF turtle file.
@base <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rel: <http://www.perceive.net/schemas/relationship/> .
<#green-goblin>
    rel:enemyOf <#spiderman> ;
    a foaf:Person ;    # in the context of the Marvel universe
    foaf:name "Green Goblin" .
<#spiderman>
    rel:enemyOf <#green-goblin> ;
    a foaf:Person ;
    foaf:name "Spiderman", "spinnen man"@nl .
```

For the above example file, the graph shows that the relation between `green goblin` and `spider man` is quite hostile (Figure 12). These RDF files can be easily and quickly searched using graph databases such as fuseki. A graph database stores the tripples in a datastructure optimised for searches (or queries) using SPARQL protocol and RDF query language (SPARQL) [5].

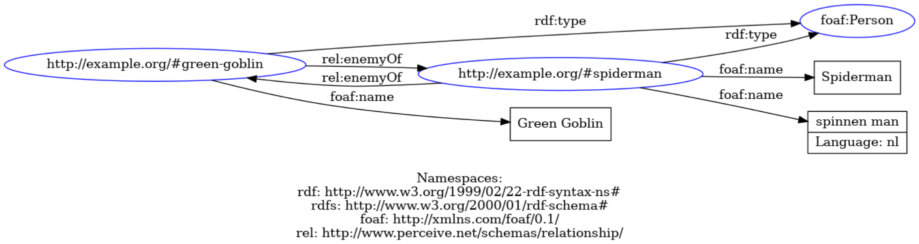

rdf:type
foaf:Person
rdf:type
http://example.org/#green-goblin
rel:enemyOf
rel:enemyOf
http://example.org/#spiderman
foaf:name
Spiderman
foaf:name
foaf:name
Green Goblin
spinnen man
Language: nl
Namespaces:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
foaf: http://xmlns.com/foaf/0.1/
rel: http://www.perceive.net/schemas/relationship/

Figure 12: The RDF graph that belongs to the example RDF file.

Such a SPARQL query might like like this

```
# Who are the enemies of the Green Goblin?
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
prefix foaf: <http://xmlns.com/foaf/0.1/> .
prefix rel: <http://www.perceive.net/schemas/relationship/> .
prefix ex: <http//example.org/>
SELECT ?enemy
WHERE
  {
    ?person a foaf:Person ;
      foaf:name ?name ;
      rel:enemyOf ?enemy .
    FILTER(STRSTARTS(?name, "Green Goblin")) .
  }
```

This query would return `ex:spiderman`.

Another benefit of using RDF is that it can be validated with SHACL [38]. SHACL is a programming language that can be used to check if an RDF file has a correct structure. For example, it can check if all people have a birth date and two biological parents:

```
ex:PersonShape
    a sh:NodeShape ;
    sh:targetClass ex:Person ;    # Person must have one birthday
    sh:property [
            sh:path ex:birthday ;
            sh:maxCount 1 ;
            sh:datatype xsd:date ;
    ] ;
sh:targetClass ex:Person ;    # Person must have one father.
    sh:property [
            sh:path ex:father ;
            sh:maxCount 1 ;
            sh:datatype ex:Person ;
    ] ;
sh:targetClass ex:Person ;    # Person must have one mother.
    sh:property [
            sh:path ex:mother ;
            sh:maxCount 1 ;
            sh:datatype ex:Person ;
    ] ;
```

E Acknowledgements

I am thankful for Sijmen Schoustra and Anna Fensel for reviewing an early draft of this thesis and providing useful feedback on the structure. I want to thank Jasper Koehorst for his help during my contributions to the FAIRDS. I also thank all the researchers from the *A. fumigatus* field who worked with me during the master thesis and gracefully shared their data and time: Martin Weichert, Hylke Kortenbosch, Mariana Couto Silva, Johanna Rhodes & Bart Fraaie. For the discussions on the topic of FAIR data I thank Christopher Wandabwa and Murambiwa Nyati. These critical discussions were useful the point where I felt stuck.

Besides the people I directly worked with, I feel indebted to all the developers of open source programmes that I've used but were not able to cite because of space limitations.

References

[1] *Allie RDF Data Portal - Integbio Database Catalog.* URL: https://integbio.jp/dbcatalog/en/record/nbdc01192 (visited on 01/15/2024).

[2] Fausto Almeida, Marcio L. Rodrigues, and Carolina Coelho. "The Still Underestimated Problem of Fungal Diseases Worldwide". In: *Frontiers in microbiology* 10 (2019).

[3] David R. Andes and William E. Dismukes. “Azoles”. In: *Essentials of Clinical Mycology*. Ed. by Carol A. Kauffman *et al.* New York, NY: Springer, 2011, pp. 61–93. DOI: 10.1007/978-1-4419-6640-7_5.

[4] Apache software foundation. *Apache Jena - Home*. 2023. URL: https://jena.apache.org/ (visited on 11/30/2023).

[5] Apache software foundation. *Apache Jena - TDB Architecture*. URL: https://jena.apache.org/documentation/tdb/architecture.html (visited on 01/10/2024).

[6] Parit Bansal *et al.* “Rhea, the reaction knowledgebase in 2022”. In: *Nucleic acids research* 50.D1 (Jan. 7, 2022), pp. D693–D700. DOI: 10.1093/nar/gkab1016.

[7] Jānis Bārzdiņš *et al.* “UML Style Graphical Notation and Editor for OWL 2”. In: *Perspectives in Business Informatics Research*. Ed. by Peter Forbrig and Horst Günther. Red. by Will Van Der Aalst *et al.* Vol. 64. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 102–114. DOI: 10.1007/978-3-642-16101-8_9.

[8] V.R. Basili and R.W. Selby. “Comparing the Effectiveness of Software Testing Strategies”. In: *Ieee transactions on software engineering* SE-13.12 (Dec. 1987), pp. 1278–1296. DOI: 10.1109/TSE.1987.232881.

[9] Barbara Bazzanella and Yannis Tzitzikas. “Interoperability Objectives and Approaches (APARSEN Deliverable D2501)”. In: *Aparsen deliverable* (Feb. 28, 2013).

[10] Sarah Berger *et al.* “Azole Resistance in *Aspergillus fumigatus*: A Consequence of Antifungal Use in Agriculture?” In: *Frontiers in microbiology* 8 (2017), p. 1024. DOI: 10.3389/fmicb.2017.01024.

[11] Christine L. Borgman and Philip Bourne. “Why It Takes a Village to Manage and Share Data”. In: *Harvard data science review* 4.3 (July 28, 2022). DOI: 10.1162/99608f92.42eec111.

[12] Gordon D. Brown *et al.* “Hidden Killers: Human Fungal Infections”. In: *Science translational medicine* 4.165 (Dec. 19, 2012). DOI: 10.1126/scitranslmed.3004404.

[13] Jochem B. Buil *et al.* “Trends in Azole Resistance in *Aspergillus fumigatus*, the Netherlands, 1994–2016”. In: *Emerging infectious diseases* 25.1 (Jan. 2019), pp. 176–178. DOI: 10.3201/eid2501.171925.

[14] Caroline Burks *et al.* “Azole-resistant *Aspergillus fumigatus* in the environment: Identifying key reservoirs and hotspots of antifungal resistance”. In: *Plos pathogens* 17.7 (July 29, 2021), e1009711. DOI: 10.1371/journal.ppat.1009711.

[15] Jose Cadena, George R. Thompson, and Thomas F. Patterson. “Aspergillosis: Epidemiology, Diagnosis, and Treatment”. In: *Infectious disease clinics* 35.2 (June 1, 2021), pp. 415–434. DOI: 10.1016/j.idc.2021.03.008.

[16] Duantao Cao *et al.* “Prevalence of Azole-Resistant Aspergillus fumigatus is Highly Associated with Azole Fungicide Residues in the Fields”. In: *Environmental science & technology* 55.5 (Mar. 2, 2021), pp. 3041–3049. DOI: 10.1021/acs.est.0c03958.

[17] Jeremy J. Carroll and Jeff Z. Pan. *XML Schema Datatypes in RDF and OWL*. URL: https://www.w3.org/TR/swbp-xsch-datatypes/ (visited on 12/19/2023).

[18] *CC BY 4.0 Deed | Attribution 4.0 International | Creative Commons*. URL: https://creativecommons.org/licenses/by/4.0/ (visited on 12/19/2023).

[19] Mark J. Costello *et al.* “Strategies for the sustainability of online open-access biodiversity databases”. In: *Biological conservation* 173 (May 1, 2014), pp. 155–165. DOI: 10.1016/j.biocon.2013.07.042.

[20] Hanna Ćwiek-Kupczyńska *et al.* “Measures for interoperability of phenotypic data: minimum information requirements and formatting”. In: *Plant methods* 12.1 (Nov. 9, 2016), p. 44. DOI: 10.1186/s13007-016-0144-4.

[21] Dan Brickley and R. V. Guha. *RDF Schema 1.1*. URL: https://www.w3.org/TR/rdf-schema/ (visited on 01/15/2024).

[22] Lois M. Douglas and James B. Konopka. “Fungal Membrane Organization: The Eisosome Concept”. In: *Annual review of microbiology* 68.1 (2014), pp. 377–393. DOI: 10.1146/annurev-micro-091313-103507.

[23] *Dublin core*. Dec. 14, 2023. URL: https://www.dublincore.org/ (visited on 01/15/2024).

[24] European Centre for Disease Prevention and Control. *Risk assessment on the impact of environmental usage of traizoles on the development and spread of resistance to medical triazoles in Aspergillus species.* LU: Publications Office, 2013.

[25] Matthew C. Fisher *et al.* “Tackling the emerging threat of antifungal resistance to human health”. In: *Nature reviews microbiology* 20.9 (9 Sept. 2022), pp. 557–571. DOI: 10.1038/s41579-022-00720-1.

[26] *Geopy: Python Geocoding Toolbox*. Version 2.4.1.

[27] Mahmoud A. Ghannoum and Louis B. Rice. “Antifungal Agents: Mode of Action, Mechanisms of Resistance, and Correlation of These Mechanisms with Bacterial Resistance”. In: *Clinical microbiology reviews* 12.4 (Oct. 1999), pp. 501–517.

[28] *Glossary:Observation unit.* URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Observation_unit (visited on 12/04/2023).

[29] Rafael S. Gonçalves and Mark A. Musen. "The variable quality of metadata about biological samples used in biomedical experiments". In: *Scientific data* 6.1 (1 Feb. 19, 2019), p. 190021. DOI: 10.1038/sdata.2019.21.

[30] Christian Graul. *leafletR: Interactive web-maps based on the leaflet JavaScript library.* manual. 2016.

[31] James Hadfield *et al.* "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (Dec. 1, 2018), pp. 4121–4123. DOI: 10.1093/bioinformatics/bty407.

[32] *Home - PDOK.* URL: https://www.pdok.nl/ (visited on 01/15/2024).

[33] Mat Hostetter *et al.* "Curl: a gentle slope language for the Web." In: *World wide web journal* 2.2 (1997), pp. 121–134.

[34] Bernadette Hyland *et al. Best Practices for Publishing Linked Data.* W3C. Jan. 9, 2014. URL: https://www.w3.org/TR/ld-bp/ (visited on 12/19/2023).

[35] *JERM Ontology Overview.* URL: https://jermontology.org/ (visited on 08/25/2023).

[36] James Jiang and Gary Klein. "Software development risks to project effectiveness". In: *Journal of systems and software* 52.1 (May 15, 2000), pp. 3–10. DOI: 10.1016/S0164-1212(99)00128-4.

[37] Fátima Jorge *et al.* "MIxS-SA: a MIxS extension defining the minimum information standard for sequence data from symbiont-associated micro-organisms". In: *Isme communications* 2.1 (1 Feb. 1, 2022), pp. 1–5. DOI: 10.1038/s43705-022-00092-w.

[38] Jose E. Labra Gayo *et al.* "Chapter 5: SHACL". In: *Validating rdf data.* Vol. 7. Validating RDF Data, Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2018.

[39] Katharina Kainz *et al.* "Fungal infections in humans: the silent crisis". In: *Microbial cell* 7.6 (Jan. 6, 2020), pp. 143–145. DOI: 10.15698/mic2020.06.718.

[40] Hylke H. Kortenbosch *et al. Catching more air: An effective and simple-to-use air sampling approach to assess aerial resistance fractions in* Aspergillus fumigatus. Nov. 7, 2022. DOI: 10.1101/2022.11.03.515058. URL: https://www.biorxiv.org/content/10.1101/2022.11.03.515058v2 (visited on 05/12/2023). preprint.

[41] “Nutrient Cycling by Saprotrophic Fungi in Terrestrial Habitats”. In: *Environmental and Microbial Relationships*. Ed. by Christian P. Kubicek and Irina S. Druzhinina. The Mycota. Berlin, Heidelberg: Springer, 2007, pp. 287–300. DOI: 10.1007/978-3-540-71840-6_16.

[42] Zoé Lacroix *et al.* “Biological Metadata Management”. In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 215–219. DOI: 10.1007/978-0-387-39940-9_628.

[43] Amos Lal *et al.* “Recurrent Pulmonary Embolism and Hypersensitivity Pneumonitis Secondary to *Aspergillus*, in a Compost Plant Worker: Case Report and Review of Literature”. In: *Lung* 196.5 (Oct. 2018), pp. 553–560. DOI: 10.1007/s00408-018-0142-6.

[44] Jean-Paul Latgé. “*Aspergillus fumigatus* and Aspergillosis”. In: *Clinical microbiology reviews* 12.2 (Apr. 1999), pp. 310–350. DOI: 10.1128/cmr.12.2.310.

[45] Song Li *et al.* “A precipitation forecast model with a neural network and improved GPT3 model for Japan”. In: *Gps solutions* 27.4 (Aug. 16, 2023), p. 186. DOI: 10.1007/s10291-023-01526-1.

[46] library of congress (LOC). *XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5*. May 2, 2022. URL: https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml (visited on 01/10/2024).

[47] Peter McQuilton *et al.* “BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences”. In: *Database* 2016 (Jan. 1, 2016), baw075. DOI: 10.1093/database/baw075.

[48] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux journal* 2014.239 (2014), p. 2.

[49] Nicholas P. Money. “Chapter 12 - Fungi and Biotechnology”. In: *The Fungi (Third Edition)*. Ed. by Sarah C. Watkinson, Lynne Boddy, and Nicholas P. Money. Boston: Academic Press, Jan. 1, 2016, pp. 401–424. DOI: 10.1016/B978-0-12-382034-1.00012-8.

[50] Joseph D Napier, Robert W Heckman, and Thomas E Juenger. “Gene-by-environment interactions in plants: Molecular mechanisms, environmental drivers, and adaptive plasticity”. In: *The plant cell* 35.1 (Jan. 1, 2023), pp. 109–124. DOI: 10.1093/plcell/koac322.

[51] Bart Nijsse, Peter J Schaap, and Jasper J Koehorst. “FAIR data station for lightweight metadata management and validation of omics studies”. In: *Gigascience* 12 (Jan. 1, 2023), giad014. DOI: 10.1093/gigascience/giad014.

[52] *OGC GeoSPARQL - A Geographic Query Language for RDF Data*. URL: https://opengeospatial.github.io/ogc-geosparql/geosparql11/spec.html (visited on 08/25/2023).

[53] Dinah V. Parums. "Editorial: The World Health Organization (WHO) Fungal Priority Pathogens List in Response to Emerging Fungal Pathogens During the COVID-19 Pandemic". In: *Medical science monitor : international medical journal of experimental and clinical research* 28 (Dec. 1, 2022), e939088-1-e939088–3. DOI: 10.12659/MSM.939088.

[54] Edzer Pebesma. "Simple features for R: Standardized support for spatial vector data". In: *The r journal* 10.1 (2018), pp. 439–446. DOI: 10.32614/RJ-2018-009.

[55] Edzer Pebesma and Roger Bivand. *Spatial data science: With applications in R*. Chapman and Hall/CRC, 2023. DOI: 10.1201/9780429459016.

[56] Eva Petrikkou *et al.* "Inoculum Standardization for Antifungal Susceptibility Testing of Filamentous Fungi Pathogenic for Humans". In: *Journal of clinical microbiology* 39.4 (Apr. 2001), pp. 1345–1347. DOI: 10.1128/jcm.39.4.1345-1347.2001.

[57] Jian Qin and John D'ignazio. "The Central Role of Metadata in a Science Data Literacy Course". In: *Journal of library metadata* 10.2-3 (Aug. 31, 2010), pp. 188–204. DOI: 10.1080/19386389.2010.506379.

[58] Will Reese. "Nginx: the high-performance web server and reverse proxy". In: *Linux journal* 2008.173 (Sept. 1, 2008), 2:2.

[59] *Refactoring and Design Patterns*. URL: https://refactoring.guru/ (visited on 06/29/2023).

[60] Philippe Rocca-Serra *et al.* "7 - Investigation-Study-Assay, a toolkit for standardizing data capture and sharing". In: *Open Source Software in Life Science Research*. Ed. by Lee Harland and Mark Forster. Woodhead Publishing Series in Biomedicine. Woodhead Publishing, Jan. 1, 2012, pp. 173–188. DOI: 10.1533/9781908818249.173.

[61] Philippe Rocca-Serra *et al.* "The FAIR Cookbook - the essential resource for and by FAIR doers". In: *Scientific data* 10.1 (1 May 19, 2023), p. 292. DOI: 10.1038/s41597-023-02166-3.

[62] *Schema.org - Schema.org*. URL: https://schema.org/ (visited on 12/19/2023).

[63] *Schimmelradar*. Schimmelradar. URL: https://schimmelradar.wordpress.com/ (visited on 12/07/2023).

[64] Sijmen E. Schoustra *et al.* “Environmental Hotspots for Azole Resistance Selection of *Aspergillus fumigatus*, the Netherlands - Volume 25, Number 7—July 2019 - Emerging Infectious Diseases journal - CDC”. In: (2019). DOI: 10.3201/eid2507.181625.

[65] Lynn M. Schriml *et al.* “COVID-19 pandemic reveals the peril of ignoring metadata standards”. In: *Scientific data* 7.1 (1 June 19, 2020), p. 188. DOI: 10.1038/s41597-020-0524-5.

[66] Felix Shaw *et al.* “COPO: a metadata platform for brokering FAIR data in the life sciences”. In: 9:495 (June 2, 2020). DOI: 10.12688/f1000research.23889.1.

[67] Gianmaria Silvello. “Theory and practice of data citation”. In: *Journal of the association for information science and technology* 69.1 (Jan. 2018), pp. 6–20. DOI: 10.1002/asi.23917.

[68] Jacob L. Steenwyk, Antonis Rokas, and Gustavo H. Goldman. “Know the enemy and know yourself: Addressing cryptic fungal pathogens of humans and beyond”. In: *Plos pathogens* 19.10 (Oct. 19, 2023), e1011704. DOI: 10.1371/journal.ppat.1011704.

[69] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. “Knowledge engineering: Principles and methods”. In: *Data & knowledge engineering* 25.1 (Mar. 1, 1998), pp. 161–197. DOI: 10.1016/S0169-023X(97)00056-6.

[70] Cassidy R. Sugimoto and Blaise Cronin, eds. *Theories of informetrics and scholarly communication: a festschrift in honor of Blaise Cronin.* Berlin Boston: De Gruyter, 2016. 426 pp.

[71] Leho Tedersoo *et al.* “Best practices in metabarcoding of fungi: From experimental design to results”. In: *Molecular ecology* 31.10 (2022), pp. 2769–2795. DOI: 10.1111/mec.16460.

[72] Leho Tedersoo *et al.* “Data sharing practices and data availability upon request differ across scientific disciplines”. In: *Scientific data* 8.1 (1 July 27, 2021), p. 192. DOI: 10.1038/s41597-021-00981-0.

[73] Fredj Tekaia and Jean-Paul Latgé. “*Aspergillus Fumigatus*: saprophyte or pathogen?” In: *Current opinion in microbiology.* Host–Microbe Interactions: Fungi / Edited by Howard Bussey · Host–microbe Interactions: Parasites / Edited by Artur Scherf · Host–microbe Interactions: Viruses / Edited by Margaret CM Smith 8.4 (Aug. 1, 2005), pp. 385–392. DOI: 10.1016/j.mib.2005.06.017.

[74] Carol Tenopir *et al.* “Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide”. In: *Plos one* 15.3 (Mar. 11, 2020), e0229003. DOI: 10.1371/journal.pone.0229003.

[75] Norman Van Rhijn and Michael Bromley. "The Consequences of Our Changing Environment on Life Threatening and Debilitating Fungal Diseases in Humans". In: *Journal of fungi* 7.5 (May 7, 2021), p. 367. DOI: 10.3390/jof7050367.

[76] Norman Van Rhijn *et al.* "Meteorological Factors Influence the Presence of Fungi in the Air; A 14-Month Surveillance Study at an Adult Cystic Fibrosis Center". In: *Frontiers in cellular and infection microbiology* 11 (Nov. 26, 2021), p. 759944. DOI: 10.3389/fcimb.2021.759944.

[77] Rhea Verbeke. "FAIR and Open Data requires proper incentives and a shift in academic culture". In: *Nature water* 1.1 (1 Jan. 2023), pp. 7–9. DOI: 10.1038/s44221-022-00012-1.

[78] Danielle Welter *et al.* "FAIR in action - a flexible framework to guide FAIRification". In: *Scientific data* 10.1 (1 May 19, 2023), p. 291. DOI: 10.1038/s41597-023-02167-2.

[79] Hadley Wickham. *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York, 2016.

[80] Hadley Wickham. *Httr2: Perform HTTP requests and process the responses*. manual. 2023.

[81] Hadley Wickham *et al.* "Welcome to the Tidyverse". In: *Journal of open source software* 4.43 (Nov. 21, 2019), p. 1686. DOI: 10.21105/joss.01686.

[82] Mark D. Wilkinson *et al.* "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (1 Mar. 15, 2016), p. 160018. DOI: 10.1038/sdata.2016.18.

[83] John Wise *et al.* "Implementation and relevance of FAIR data principles in biopharmaceutical R&D". In: *Drug discovery today* 24.4 (Apr. 1, 2019), pp. 933–938. DOI: 10.1016/j.drudis.2019.01.008.

[84] Katherine Wolstencroft *et al.* "SEEK: a systems biology data and model management platform". In: *Bmc systems biology* 9.1 (July 11, 2015), p. 33. DOI: 10.1186/s12918-015-0174-y.

[85] Yasunori Yamamoto, Atsuko Yamaguchi, and Andrea Splendiani. "YummyData: providing high-quality open life science data". In: *Database* 2018 (Jan. 1, 2018), bay022. DOI: 10.1093/database/bay022.

[86] Gokul Yenduri *et al.* *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. May 21, 2023. DOI: 10.48550/arXiv.2305.10435. URL: http://arxiv.org/abs/2305.10435 (visited on 01/15/2024). preprint.

[87] Pelin Yilmaz *et al.* “Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications”. In: *Nature biotechnology* 29.5 (5 May 2011), pp. 415–420. DOI: 10.1038/nbt.1823.

[88] Jianhua Zhang *et al.* “Dynamics of *Aspergillus fumigatus* in Azole Fungicide-Containing Plant Waste in the Netherlands (2016–2017)”. In: *Applied and environmental microbiology* 87.2 (Jan. 4, 2021), e02295–20. DOI: 10.1128/AEM.02295-20.

[89] Tomasz Zielinski, Johnny Hay, and Andrew J. Millar. “The grant is dead, long live the data - migration as a pragmatic exit strategy for research data preservation”. In: *Wellcome open research* 4 (Sept. 23, 2019), p. 104. DOI: 10.12688/wellcomeopenres.15341.2.