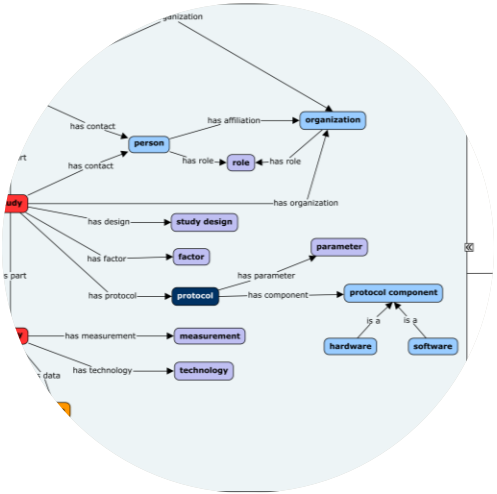
FAIR Data & ISA standard

Using the ISA standard for collecting and sharing data

2024, Sven Warris & Rick van de Zedde



WAGENINGEN
UNIVERSITY & RESEARCH

Data challenges in genomics & phenomics

- Collecting and pre-processing terabytes of data
  - Single and combined experiments
  - Many different sensors, technology platforms
  - Many different pre-processing steps
- Sharing and publishing data
- Metadata is as important as the data itself
  - Organisms, treatment, samples, etc
  - Sensor type, settings, etc

Project partners

WAGENINGEN
UNIVERSITY & RESEARCH

- Bioscience
- Biointeractions
- Biometrics

WUR is serious about Data

PHENET
PHENOTYPING & ENVIROTYPING
SOLUTIONS FOR AGROECOLOGY

NPEC

UNLOCK
MICROBIAL POTENTIAL

NARO
National Agriculture and Food Research Organization

vlpb
Virtual
Laboratory
Plant Breeding

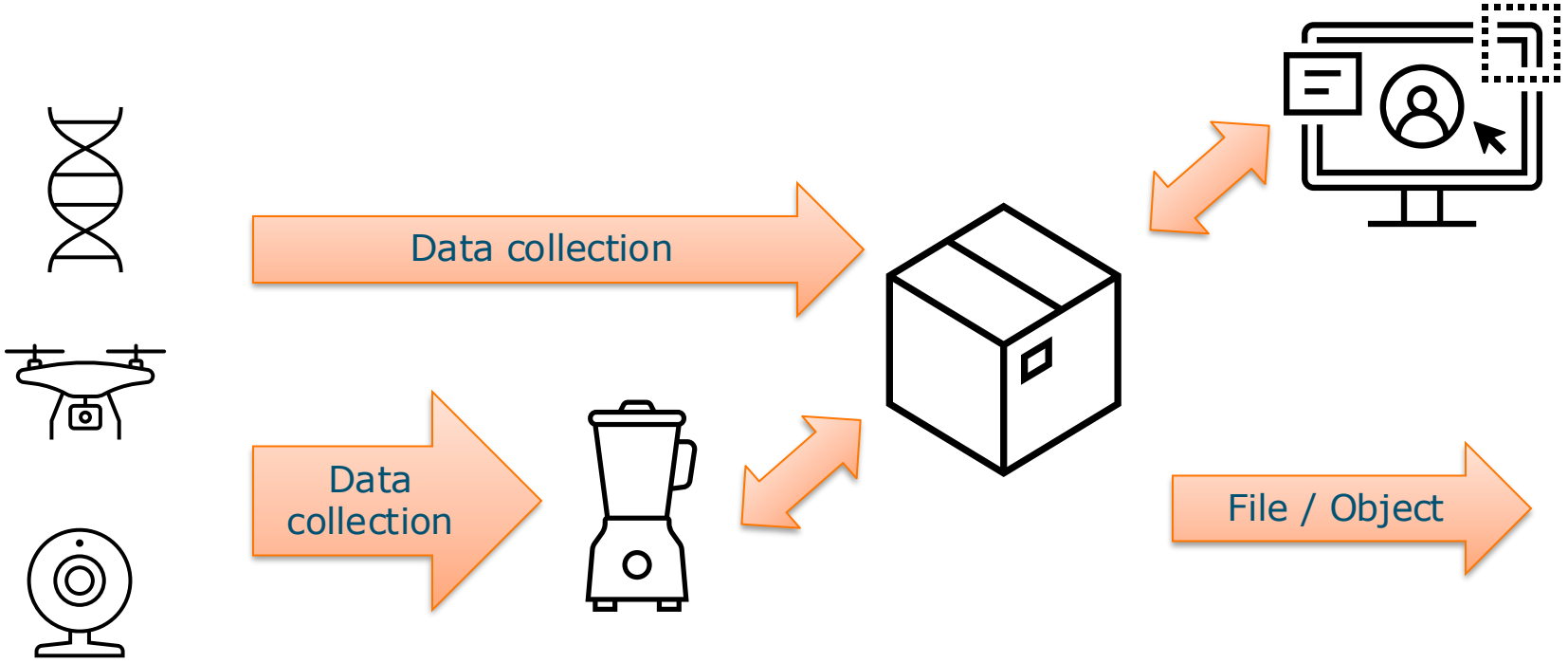
RootEd
DOCTORAL NETWORK

FB-Information Technology

WDCC
connecting data

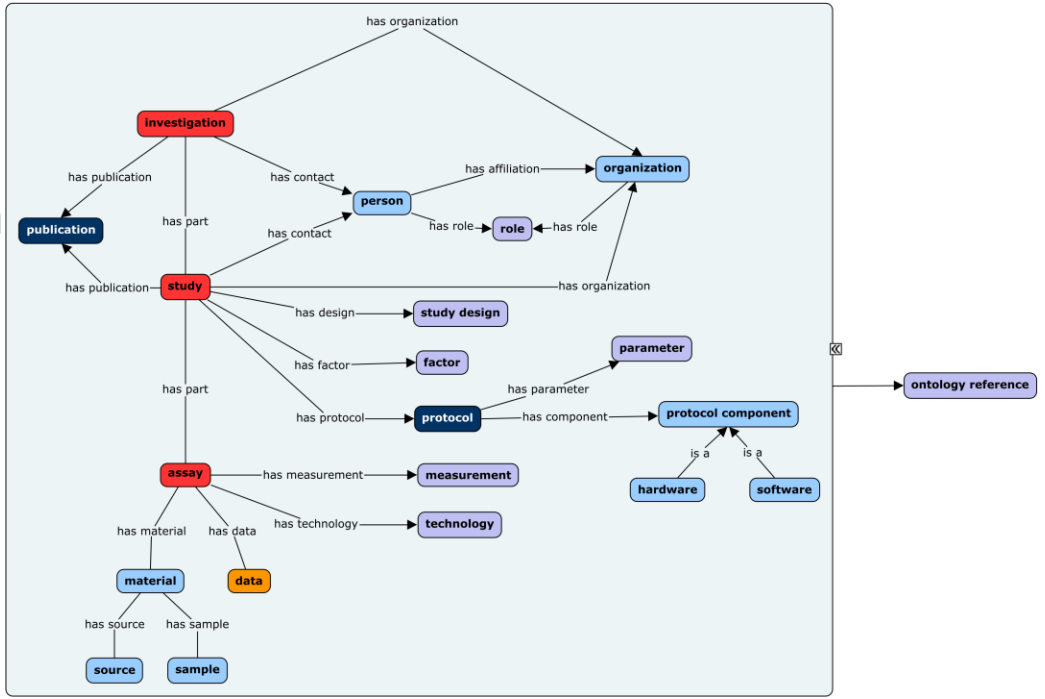
WAGENINGEN
UNIVERSITY & RESEARCH

Data flow


Data collection
Data collection
File / Object
Source
Processing
Storage
Sharing

Investigation / study / assay (ISA)

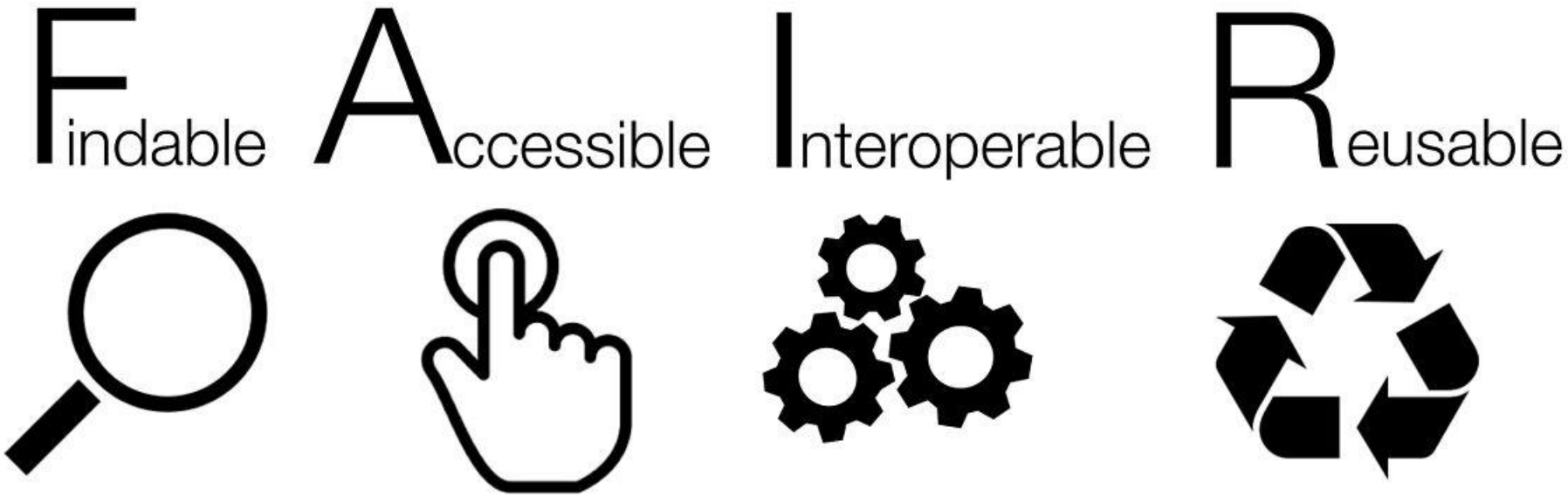
- Standardized way of structuring project metadata
- Data files, documents, etc
- Ontology-based
- Important entities:
    - Person, Assay, Data



- https://isa-specs.readthedocs.io/en/latest/isamodel.html

Investigation / study / assay

- Structure is the same in, for example:
  - eLabJournal
  - FAIRDOM-seek
- Python support through isatools:
  - ISAJSON / ISATAB
- R support:
  - ISATAB
  - ISAJSON currently being developed


Findable
Accessible
Interoperable
Reusable

FAIRDOM-seek

Open-source platform designed for the cataloguing and sharing of diverse scientific research data, including datasets, models, simulations, processes, and outcomes.

- **Preserves Associations**: It maintains relationships between various research components along with information about the involved people and organizations
- **ISA Infrastructure**: FAIRDOM-SEEK structures how experiments are part of broader studies and investigations
- **Configurable Structure:** structure is adaptable, making it suitable for various scientific fields

FAIRDOM-seek

- **Sharing Permissions:** flexible and detailed sharing permissions, supporting collaboration at different research stages, from initial collaboration to publishing final results
- **DOI Generation**: Digital Object Identifiers for individual items or entire collections packaged as Research Objects
- **Semantic Technology**: Advanced queries over its content
- **Metadata Collection:** standard Excel tools and processes (RightField)
- **MIAPPE** support: metadata from *Minimum Information About Plant Phenotyping Experiments*

Metadata implementation (MIAPPE)

20220111-WUR_test24-metadata.json

```
{
  "Name": "Minimum Information About Plant Phenotyping Experiment",
  "Models": [
    {
      "Name": "Enviror
      "Definition": "E
      "Attributes": [
        {
          "Name": "Air
          "Line": "ENV
          "Definition"
          "Value": "14
          "Format": "D
          "Marker": "E
        },
        {
          "Name": "Org
          "Line": "ENV
          "Definition"
          "Value": "14
          "Format": "D
          "Marker": "E
        },
        {
          "Name": "Liq
          "Line": "ENV
          "Definition"
          "Value": "98
          "Format": "m
          "Marker": "E
        },
```

20220111-WUR_test24-metadata.json

```
    {
      "Name": "Plot",
      "Definition": "Plot position and s
      "Attributes": [
        {
          "Name": "Geographic location (
          "Line": "DM-19",
          "Definition": "Coordinate poin
          "Value": "51.988212150",
          "Format": "Degrees in the deci
          "Marker": "PlotIdentity"
        },
        {
          "Name": "Geographic location (
          "Line": "DM-20",
          "Definition": "Coordinate poin
          "Value": "5.661708225",
          "Format": "Degrees in the deci
          "Marker": "PlotIdentity"
        },
        {
          "Name": "Geographic location (
          "Line": "DM-21",
          "Definition": "INS height of p
          "Value": "9.8",
          "Format": "m (metre)",
          "Marker": "PlotIdentity"
        },
        {
          "Name": "Observation unit ID",
          "Line": "DM-70",
          "Definition": "Plotname",
          "Value": "WUR_test24",
          "Format": "Text",
          "Marker": "PlotIdentity"
        },
        {
          "Name": "Plot coordinate A (Lo
          "Line": "",
          "Definition": "",
          "Value": "5.661030201",
          "Format": "Degrees in the deci
```

20220111-WUR_test24-metadata.json

```
    {
      "Name": "ImagingSession",
      "Definition": "Imaging sessing parameters",
      "Attributes": [
        {
          "Name": "Event accession
          "Line": "DM-66",
          "Definition": "Foldernam
          "Value": "WUR_test24-202
          "Format": "Text",
          "Marker": "ImageIdentity
        },
        {
          "Name": "Event date",
          "Line": "DM-68",
          "Definition": "Date and
          "Value": "20220111-09:58
          "Format": "Date/Time (IS
          "Marker": "ImageContext"
        },
        {
          "Name": "Plot entry coor
          "Line": "",
          "Definition": "",
          "Value": "5.661021788",
          "Format": "Degrees in th
          "Marker": "ImageContext"
        },
        {
          "Name": "Plot entry coor
          "Line": "",
          "Definition": "",
          "Value": "51.989074658",
          "Format": "Degrees in th
          "Marker": "ImageContext"
        },
        {
```

20220111-WUR_test24-metadata.json

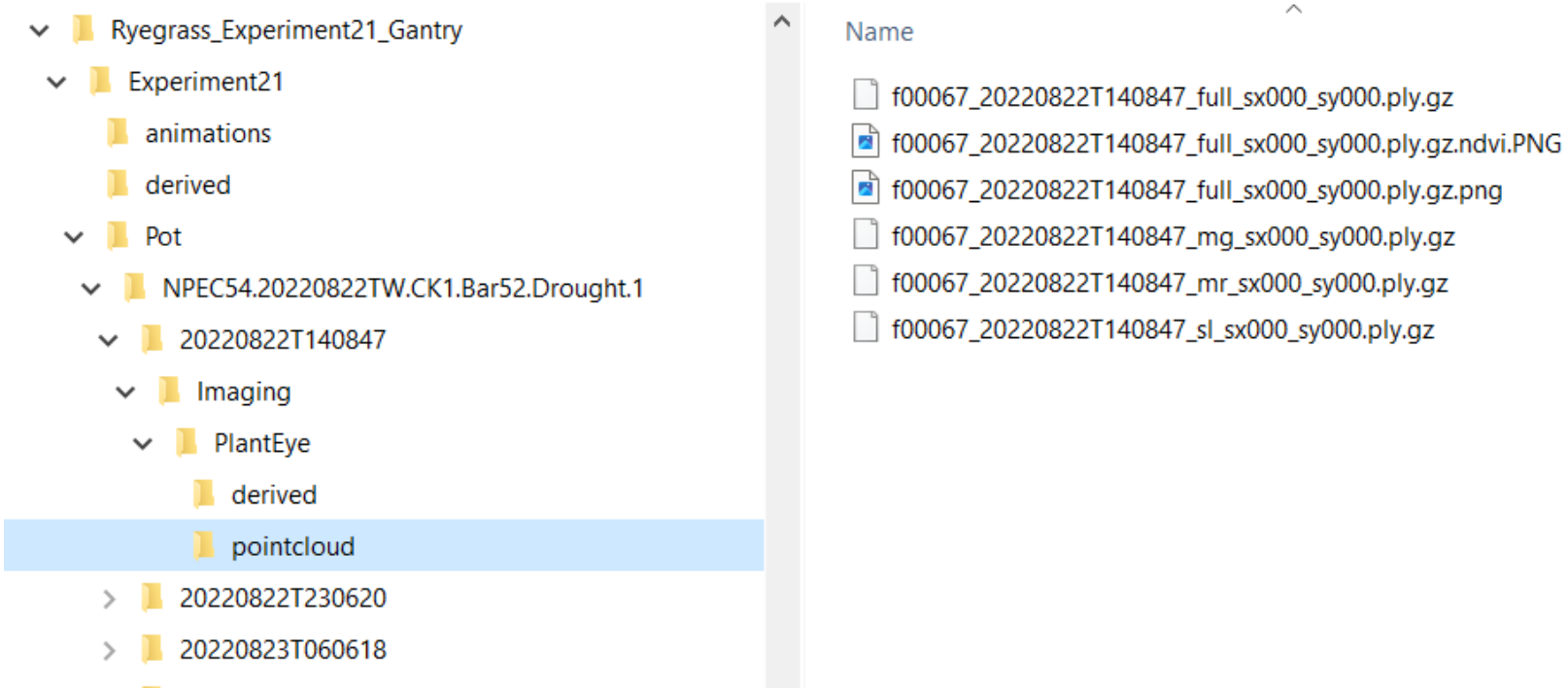
```
      "Name": "LIDAR",
      "Definition": "Settings and specifications of LIDAR",
      "Attributes": [
        {
          "Name": "LIDAR used",
          "Line": "",
          "Definition": "",
          "Value": "Yes",
          "Format": "Text",
          "Marker": "DeviceSetting"
        },
        {
          "Name": "LIDAR type",
          "Line": "",
          "Definition": "",
          "Value": "Sick LMS400-2000",
          "Format": "Text",
          "Marker": "DeviceProperty"
        },
        {
          "Name": "LIDAR product number",
          "Line": "",
          "Definition": "Manufacturers product number",
          "Value": "1041725",
          "Format": "Text",
          "Marker": "DeviceIdentity"
        },
        {
          "Name": "LIDAR serial number",
          "Line": "",
          "Definition": "Serial number of sensor",
          "Value": "14420197",
          "Format": "Text",
```

WAGEN
UNIVERSITY &

Extended ISA directory structure

- With thousands of measurements per experiment, the ISA structure is not suited for file storage
    - Operating / File systems and file browsers cannot deal with that many files in a single directory
- Extended ISA structure:
  Experiment / exp[id] / [Sample type] / [Pot ID] / [Assay timestamp] / Imaging / PlantEye / [data type]
- Makes data browsable (aka Findable) again

Extended ISA directory structure



ISA-JSON

- JSON file containing:
    - Project metadata
    - Samples & organisms
    - Sensor technologies
    - Location of the data files
    - And the type for each file (raw, derived)
- Computer (and little bit human) readable format
- Readily in and out Python & R isatools

```
{
    "@id": "#sample/a721c29a-5f86-4921-a5ce-024c39833d04",
    "characteristics": [
        {
            "category": {
                "@id": "#characteristic_category/040b5fbf-5d50-4631-ad8f-8d509ddcb0a5"
            },
            "comments": [],
            "value": {
                "@id": "#ontology_annotation/cb10d77d-dac2-4d6f-a218-e571e730208d",
                "annotationValue": "Plant",
                "comments": [],
                "termAccession": "http://purl.bioontology.org/ontology/NCBITAXON/33090",
                "termSource": "NCBITaxon"
            }
        }
    ],
    "comments": [],
    "derivesFrom": [
        {
            "@id": "#source/1cf2584b-e509-46ff
        }
    ],
    "factorValues": [],
    "name": "F1_6"
},
```


Sample

Assay (part)

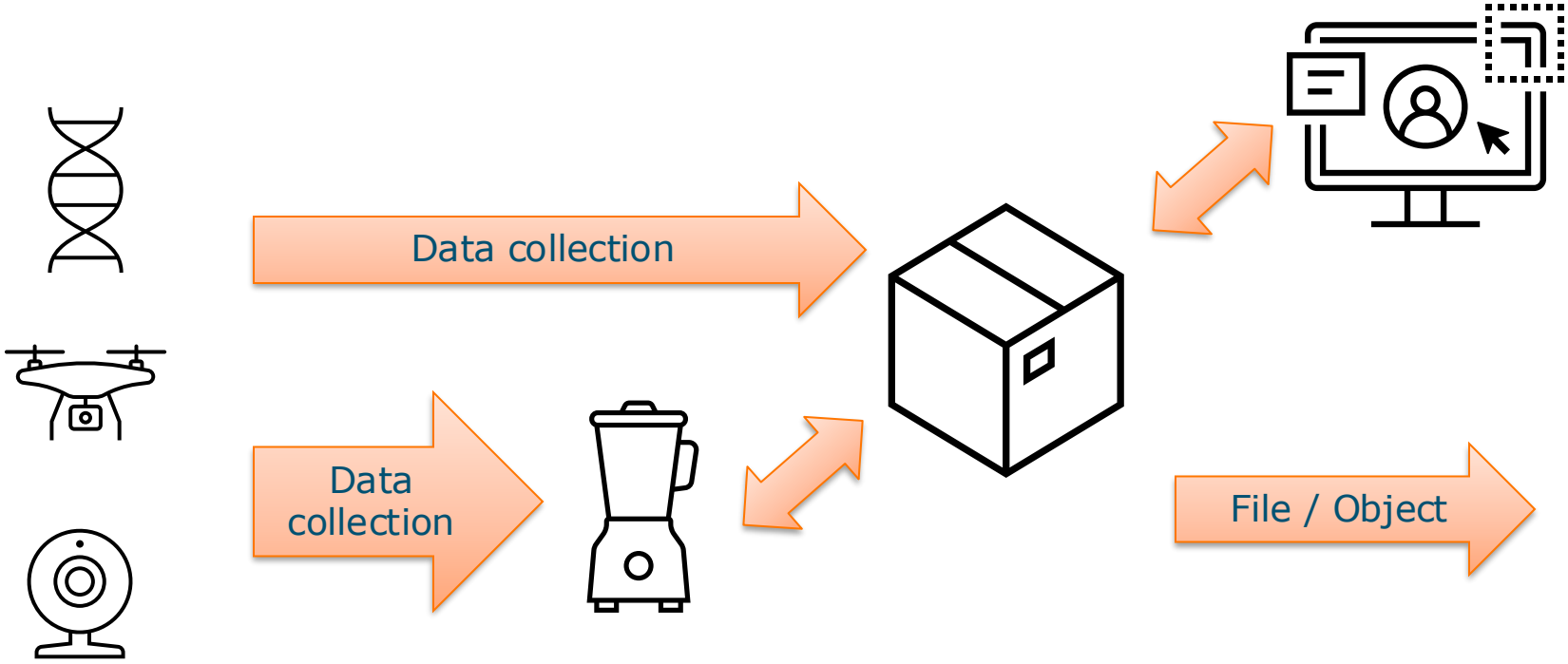

```
        {
            "@id": "#data_file/0f65eadf-f045-4c0c-af40-bf91970627a9",
            "comments": [
                {
                    "name": "fullPath",
                    "value": "running_exp28_Gantry4/exp28/Pot/F2_5/20230725T111409/Imaging/PlantEye/derived/20230725T111409_psri.csv"
                }
            ],
            "name": "20230725T111409_psri.csv",
            "type": "Derived Data File"
        }
    ],
    "filename": "20230725T111409",
    "materials": {
        "otherMaterials": [],
        "samples": [
            {
                "@id": "#sample/a575a1be-6b2b-4b41-af26-c82fcabb6aa8"
            }
        ]
    },
    "measurementType": {
        "@id": "#ontology_annotation/2279abf0-7b79-478d-a11f-7da48bf11403",
        "annotationValue": "",
        "comments": [],
        "termAccession": "",
        "termSource": ""
    },
    "processSequence": [],
    "technologyPlatform": "PlantEye",
    "technologyType": {
        "@id": "#ontology_annotation/4dbb5726-90f0-4df2-bb43-eb16af6a3b6a",
        "annotationValue": "Imaging",
        "comments": [],
        "termAccession": "http://jermontology.org/ontology/JERMOntology#Imaging",
        "termSource": ""
    },
```


WAGENINGEN
UNIVERSITY & RESEARCH

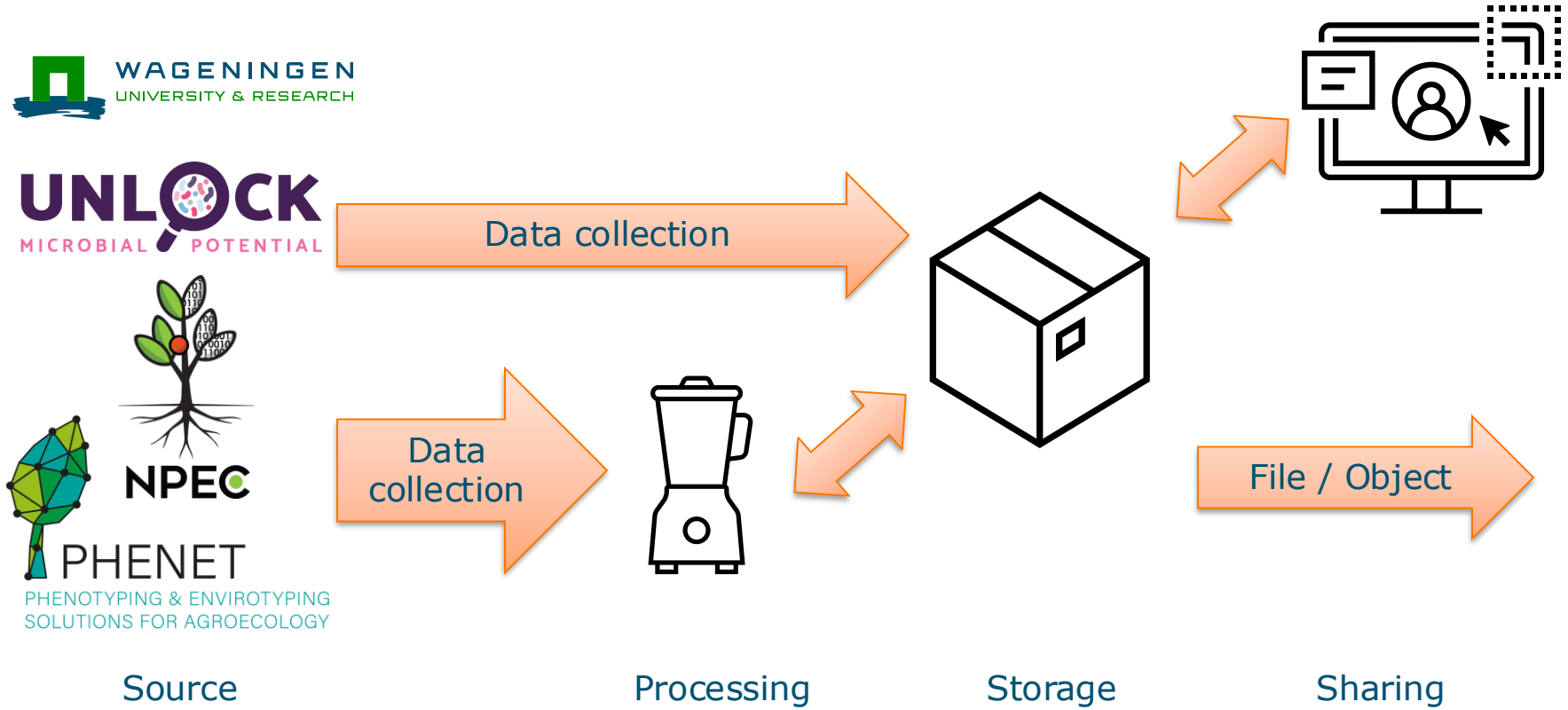
Use of ISA and ISA-JSON

- Adding pre-processing steps:
  - Get the relevant files
  - Process and store the location also in JSON
- For researchers:
  - Direct access to the relevant data files
  - Sharing makes the data more FAIR
- IT:
  - Use the metadata to store, archive or provide access to data

Data flow implementation



Data flow implementation



Data flow implementation

WAGENINGEN UNIVERSITY & RESEARCH

UNLOCK MICROBIAL POTENTIAL

NPEC

PHENET
PHENOTYPING & ENVIROTYPING SOLUTIONS FOR AGROECOLOGY

Data collection

Data collection

File / Object

Source

Processing

Storage

Sharing

WAGENINGEN UNIVERSITY & RESEARCH

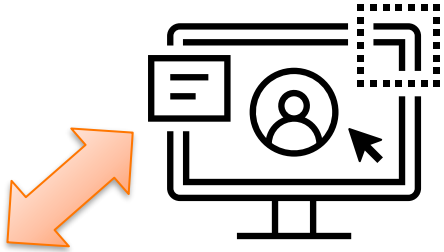
Data flow implementation

WAGENINGEN UNIVERSITY & RESEARCH

UNLOCK MICROBIAL POTENTIAL

NPEC

PHENET
PHENOTYPING & ENVIROTYPING SOLUTIONS FOR AGROECOLOGY



Data collection

lustre

Data collection

iRODS

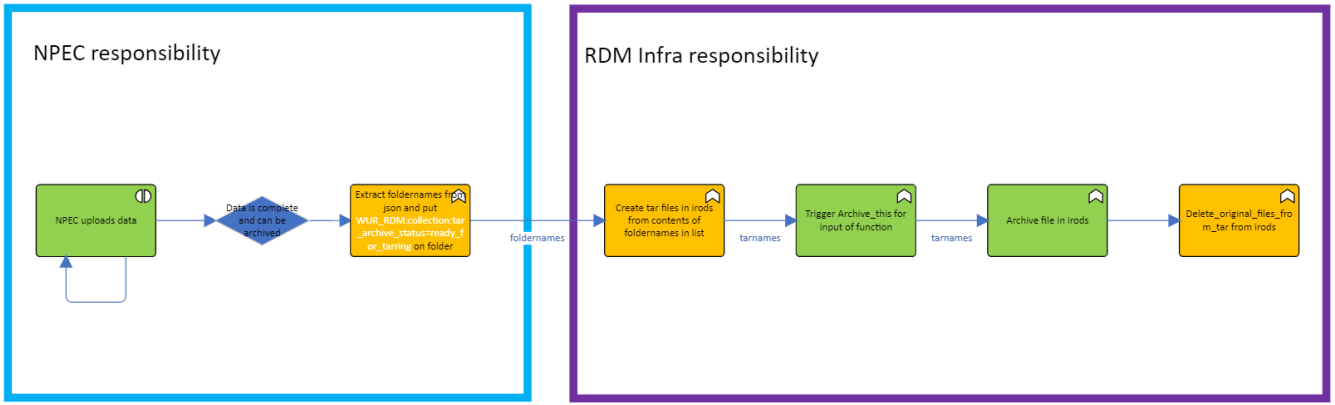
File / Object

Source | Processing | Storage | Sharing

WAGENINGEN UNIVERSITY & RESEARCH

Archiving data file – ISA-JSON - iRODS

- Cluster data files (TAR) on assay level
- Archive these TAR-files to tape
- Add metadata (original full file path of original collection name) to these TAR-files



Data flow implementation

WAGENINGEN UNIVERSITY & RESEARCH

UNLOCK MICROBIAL POTENTIAL

NPEC

PHENET
PHENOTYPING & ENVIROTYPING SOLUTIONS FOR AGROECOLOGY

Data collection

Data collection

lustre

iRODS

FAIRDOM SEEK

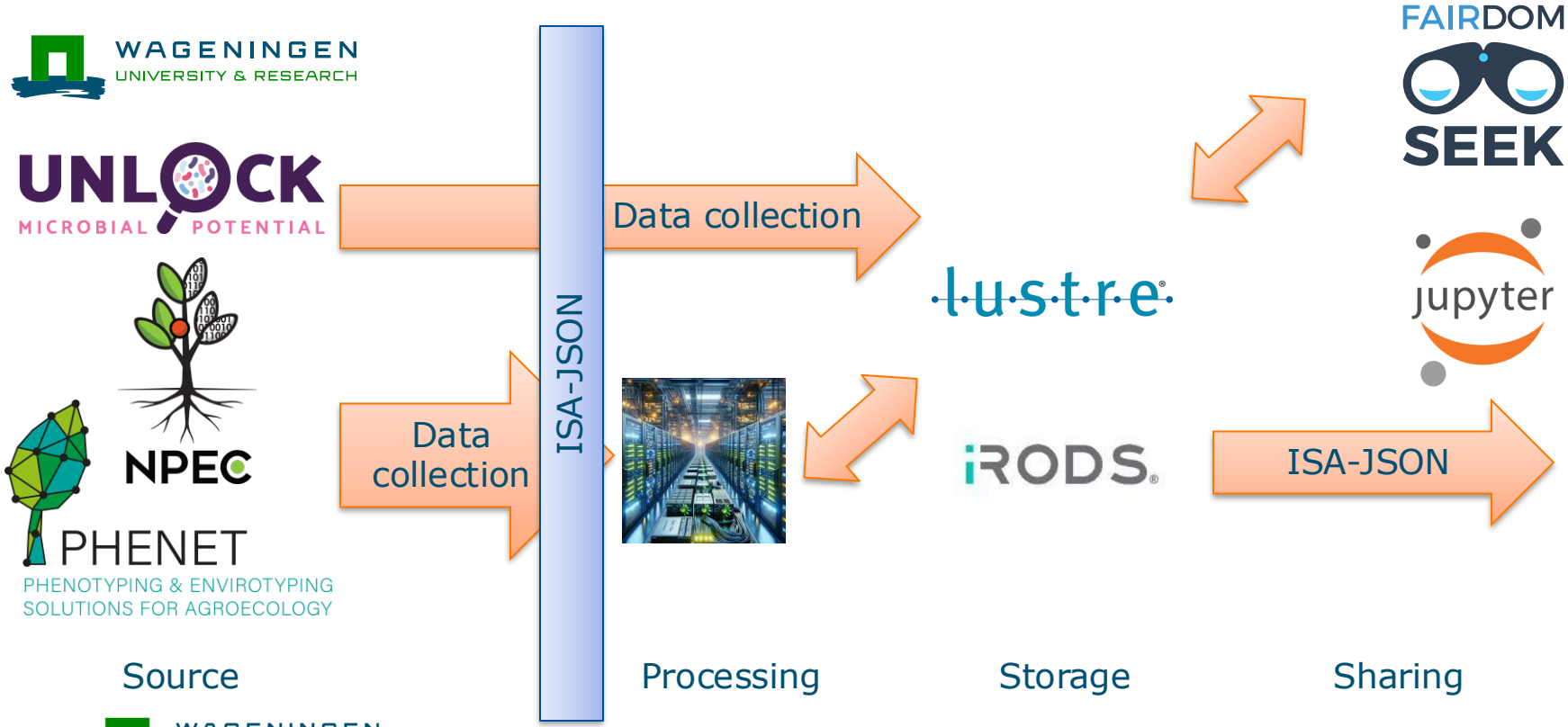
jupyter

ISA-JSON

Source | Processing | Storage | Sharing

WAGENINGEN UNIVERSITY & RESEARCH

Data flow implementation



Ongoing implementations

- Data in iRODS
    - URL of file location in iRODS
- extISA structure for many projects
    - Automatic archiving on tape
- Create experiment related ISA-JSON files
- Processing ISA-JSON & data in iRODS using Python notebooks
    - Configurable upload to FAIRDOM-seek
- Upload datasets for data publications incl DOI

Glossary

- **MIAPPE**: Minimal Information About a Plant Phenotyping Experiment.
  Thisis an open, community driven, data standard designed to harmonize data from plant phenotyping experiments. MIAPPE provides a specification including a checklist and a data model of metadata required to adequately describe plant phenotyping experiments. www.miappe.org
- **ISA-Tab**: The Investigation/Study/Assay (ISA) tab-delimited (TAB) format.
  This is a known and general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from 'omics-based' experiments employing a combination of technologies. Created by core developers from the University of Oxford, ISA-TAB v1.0 was released in November 2008.
- **JSON**: JavaScript Object Notation.
  This is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs. JSON is a language-independent data format. It was derived from JavaScript, but many modern programming languages include code to generate and parse JSON-format data. JSON filenames use the extension .json.

More info:

- https://www.npec.nl/phenotyping-modules/module-7-data/
- https://seek4science.org/about_us.html
- https://www.miappe.org/

Question?

Balazs Brankovics

Bart-Jan van Rossum

Rick van de Zedde

Tim van Daalen

Sven Warris

Many others