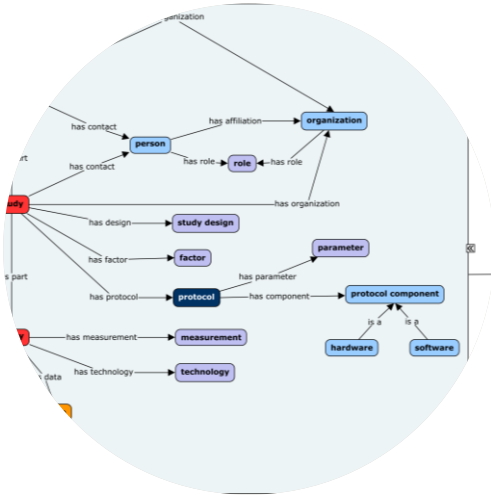


# FAIR Data & ISA standard



Using the ISA standard for collecting and sharing data  
2024, Sven Warris & Rick van de Zedde



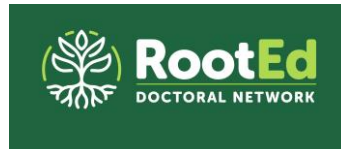
# Data challenges in genomics & phenomics

- 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

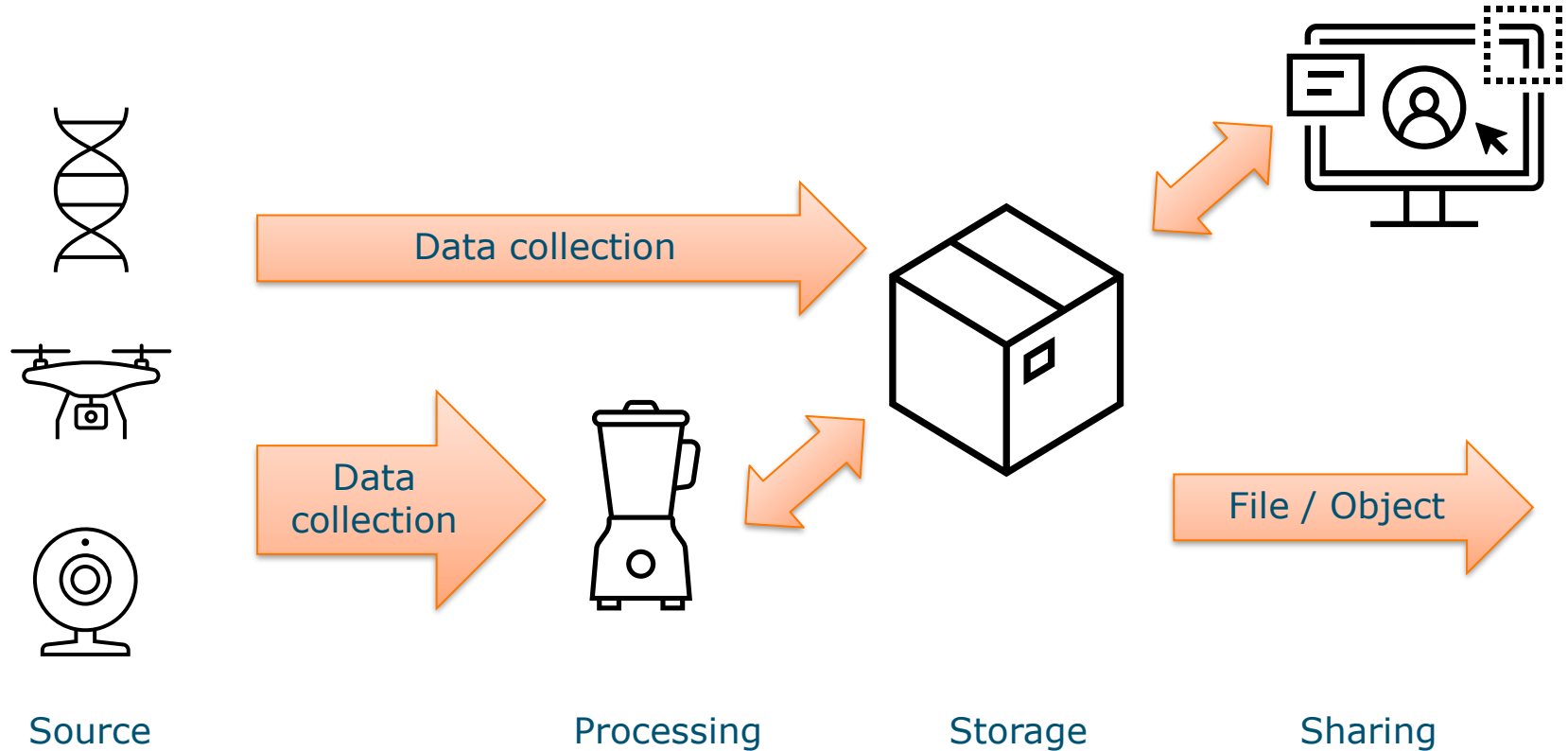
# Project partners



- Bioscience
- Biointeractions
- Biometrics

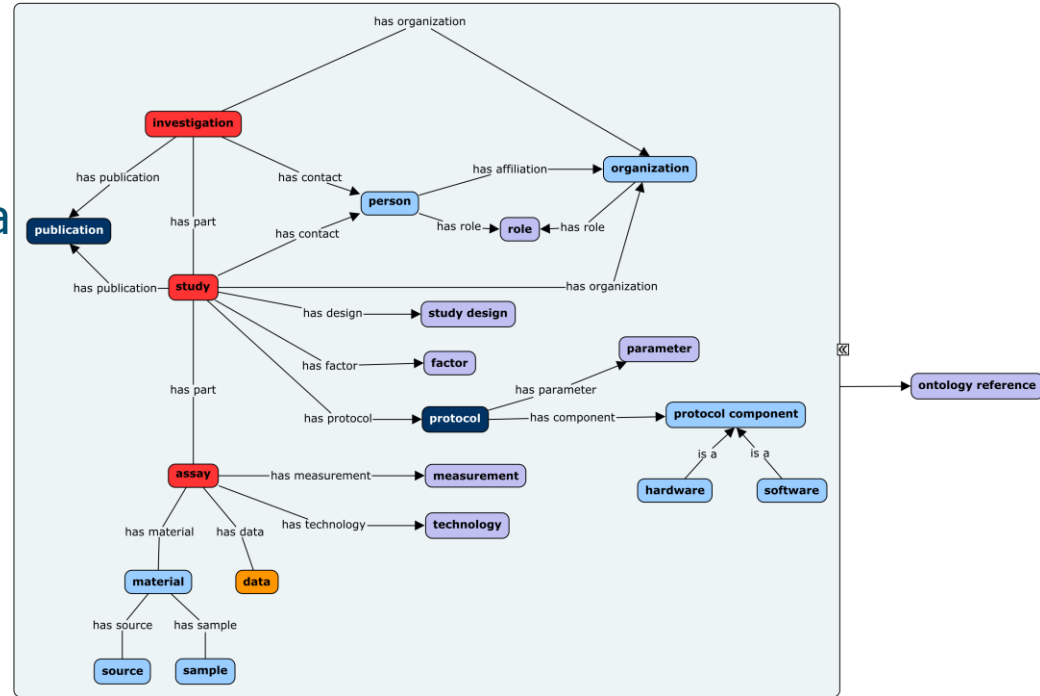


# Data flow



# Investigation / study / assay

- 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

# 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

- **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 20220111-WUR_test24-metadata.json 20220111-WUR_test24-metadata.json 20220111-WUR_test24-metadata.json
1 {
2   "Name": "Minimum I
3   "Models": [
4     {
5       "Name": "Enviro
6       "Definition": "
7       "Attributes": [
8         {
9           "Name": "A
10          "Line": "EM
11          "Definition
12          "Value": "
13          "Format": "
14          "Marker": "
15        },
16        {
17          "Name": "O
18          "Line": "EM
19          "Definition
20          "Value": "
21          "Format": "
22          "Marker": "
23        },
24        {
25          "Name": "L
26          "Line": "EM
27          "Definition
28          "Value": "
29          "Format": "
30          "Marker": "
31        },
32      ],
33    },
34  ],
35}
```

```
56 {
57   "Name": "Plot",
58   "Definition": "
59   "Attributes": [
60     {
61       "Name": "Ge
62       "Line": "DM
63       "Definition
64       "Value": "5
65       "Format": "
66       "Marker": "
67     },
68     {
69       "Name": "Ge
70       "Line": "DM
71       "Definition
72       "Value": "5
73       "Format": "
74       "Marker": "
75     },
76     {
77       "Name": "Ge
78       "Line": "DM
79       "Definition
80       "Value": "9
81       "Format": "
82       "Marker": "
83     },
84     {
85       "Name": "Ob
86       "Line": "DM
87       "Definition
88       "Value": "W
89       "Format": "
90       "Marker": "
91     },
92     {
93       "Name": "Pl
94       "Line": "
95       "Definition
96       "Value": "5
97       "Format": "
98       "Marker": "
99     },
100    ],
101  ],
102}
```

```
166 {
167   "Name": "Imag
168   "Definition": "
169   "Attributes": [
170     {
171       "Name": "Ev
172       "Line": "DM
173       "Definition
174       "Value": "W
175       "Format": "
176       "Marker": "
177     },
178     {
179       "Name": "Ev
180       "Line": "DM
181       "Definition
182       "Value": "2
183       "Format": "
184       "Marker": "
185     },
186     {
187       "Name": "Pl
188       "Line": "
189       "Definition
190       "Value": "5
191       "Format": "
192       "Marker": "
193     },
194     {
195       "Name": "Pl
196       "Line": "
197       "Definition
198       "Value": "C
199       "Format": "
200       "Marker": "
201     },
202     {
203       "Name": "Pl
204       "Line": "
205       "Definition
206       "Value": "5
207       "Format": "
208       "Marker": "
209     },
210    ],
211  ],
212}
```

```
559 "Name": "LIDAR",
560 "Definition": "Settings and specifications of LIDAR",
561 "Attributes": [
562   {
563     "Name": "LIDAR used",
564     "Line": "",
565     "Definition": "",
566     "Value": "Yes",
567     "Format": "Text",
568     "Marker": "DeviceSetting"
569   },
570   {
571     "Name": "LIDAR type",
572     "Line": "",
573     "Definition": "",
574     "Value": "Sick LMS400-2000",
575     "Format": "Text",
576     "Marker": "DeviceProperty"
577   },
578   {
579     "Name": "LIDAR product number",
580     "Line": "",
581     "Definition": "Manufacturers product number",
582     "Value": "1041725",
583     "Format": "Text",
584     "Marker": "DeviceIdentity"
585   },
586   {
587     "Name": "LIDAR serial number",
588     "Line": "",
589     "Definition": "Serial number of sensor",
590     "Value": "14420197",
591     "Format": "Text",
592     "Marker": "DeviceIdentity"
593   },
594   {
595     "Name": "LIDAR viewing angle",
596     "Line": "",
597     "Definition": "",
598     "Value": "70",
599     "Format": "Degrees",
600     "Marker": "DeviceProperty"
601   },
602   {
603     "Name": "LIDAR framerate"
604   }
605 ],
606 }
```

# 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

- ▼ Ryegrass\_Experiment21\_Gantry
  - ▼ Experiment21
    - animations
    - derived
  - ▼ Pot
    - ▼ NPEC54.20220822TW.CK1.Bar52.Drought.1
      - ▼ 20220822T140847
        - ▼ Imaging
          - ▼ PlantEye
            - derived
            - pointcloud
    - > 20220822T230620
    - > 20220823T060618

- Name ^
- f00067\_20220822T140847\_full\_sx000\_sy000.ply.gz
  - f00067\_20220822T140847\_full\_sx000\_sy000.ply.gz.ndvi.PNG
  - f00067\_20220822T140847\_full\_sx000\_sy000.ply.gz.png
  - f00067\_20220822T140847\_mg\_sx000\_sy000.ply.gz
  - f00067\_20220822T140847\_mr\_sx000\_sy000.ply.gz
  - f00067\_20220822T140847\_sl\_sx000\_sy000.ply.gz

# 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

## Assay (part)

```
{
  "@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

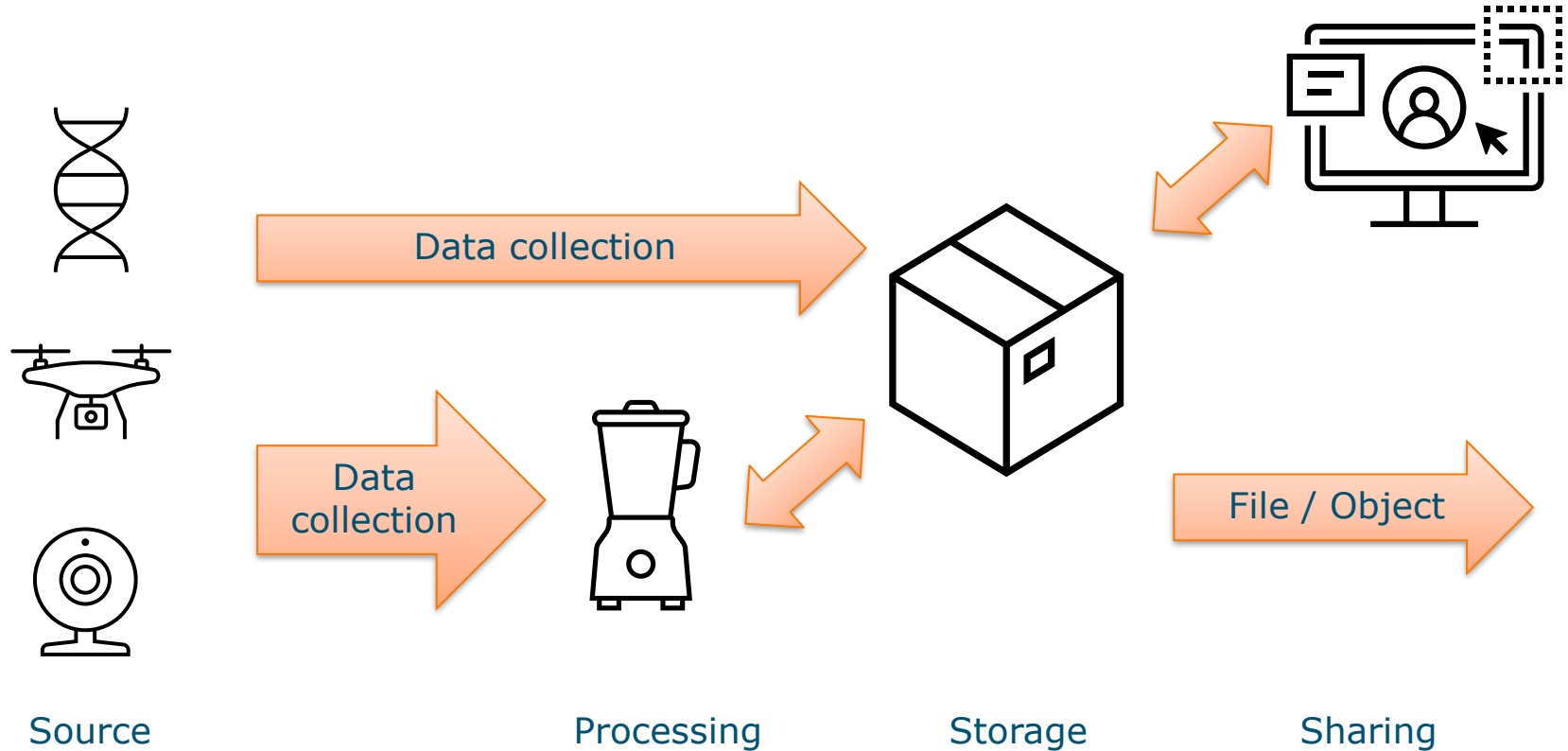
```
{
  "@id": "#data_file/0f65eadf-f045-4c0c-af40-bf91970627a9",
  "comments": [
    {
      "name": "fullPath",
      "value": "running_exp28_Gantry4/exp28/Pot/F2_5/20230725T111409/Imaging/PlantEye/derived/20230725T111409_psri.csv"
    }
  ],
  "name": "20230725T111409_psri.csv",
  "type": "Derived Data File"
},
{
  "filename": "20230725T111409",
  "materials": {
    "otherMaterials": [],
    "samples": [
      {
        "@id": "#sample/a575albe-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": ""
  }
},
```

# 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

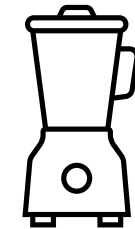


Source

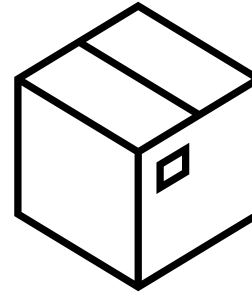


Data collection

Data collection



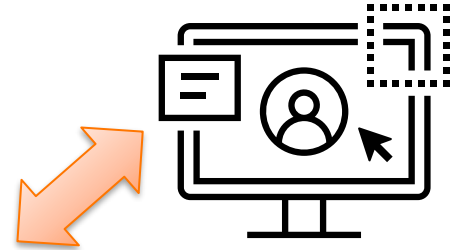
Processing



Storage

File / Object

Sharing





# Data flow implementation



NPEC



PHENET

PHENOTYPING & ENVIROTYPING  
SOLUTIONS FOR AGROECOLOGY

Source

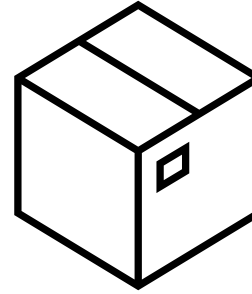


Data collection

Data  
collection



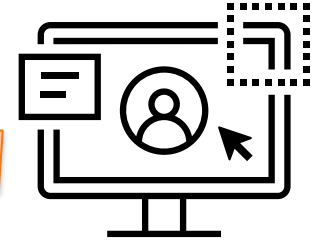
Processing



Storage

File / Object

Sharing



# Data flow implementation



Source



Data collection

Data collection



Processing

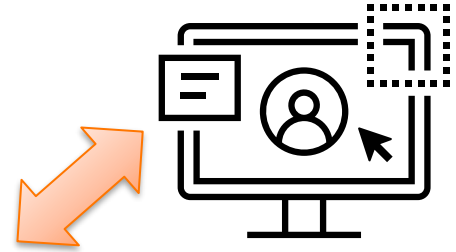
lustre®

iRODS®

Storage

File / Object

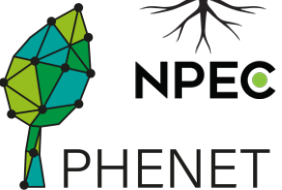
Sharing



# Data flow implementation



NPEC



PHENET

PHENOTYPING & ENVIROTYPING  
SOLUTIONS FOR AGROECOLOGY

Source



Processing



Storage



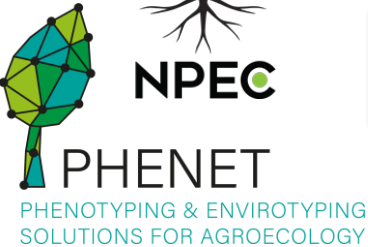
FAIRDOM



Sharing



# Data flow implementation



Source



ISA-JSON

Data collection

Data collection



Processing

lustre

iRODS

Storage

ISA-JSON

Sharing

FAIRDOM



# 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

# Question?

Balazs Brankovics

Bart-Jan van Rossum

Rick van de Zedde

Tim van Daalen

Sven Warris

Many others

