RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

RISIS Tool Demonstration Event: The OpenAIRE Research Graph: an Open Access resource for research on research

Claudio Atzori, Miriam Baglioni, Alessia Bardi

October 26 2022

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n° 824091

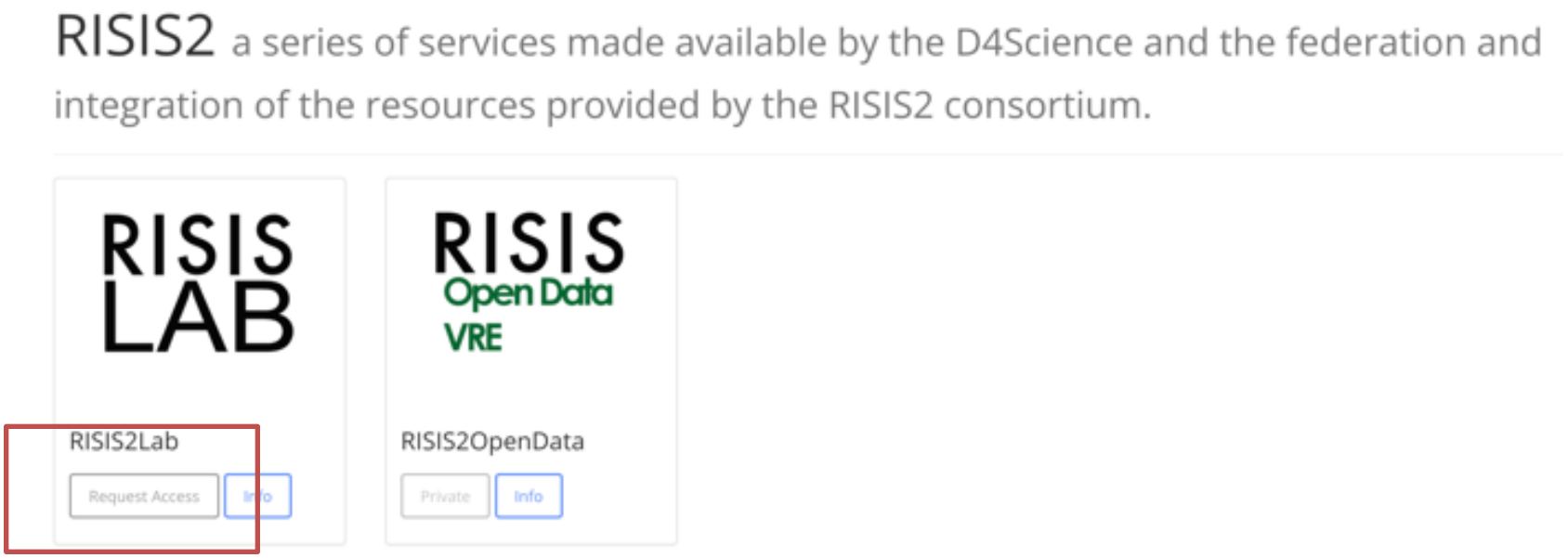
Getting started

RISIS

- Ensure you can enter the RISIS Lab Virtual Research Environment of the D4Science infrastructure

Request access (if needed)

https://risis2.d4science.org/explore

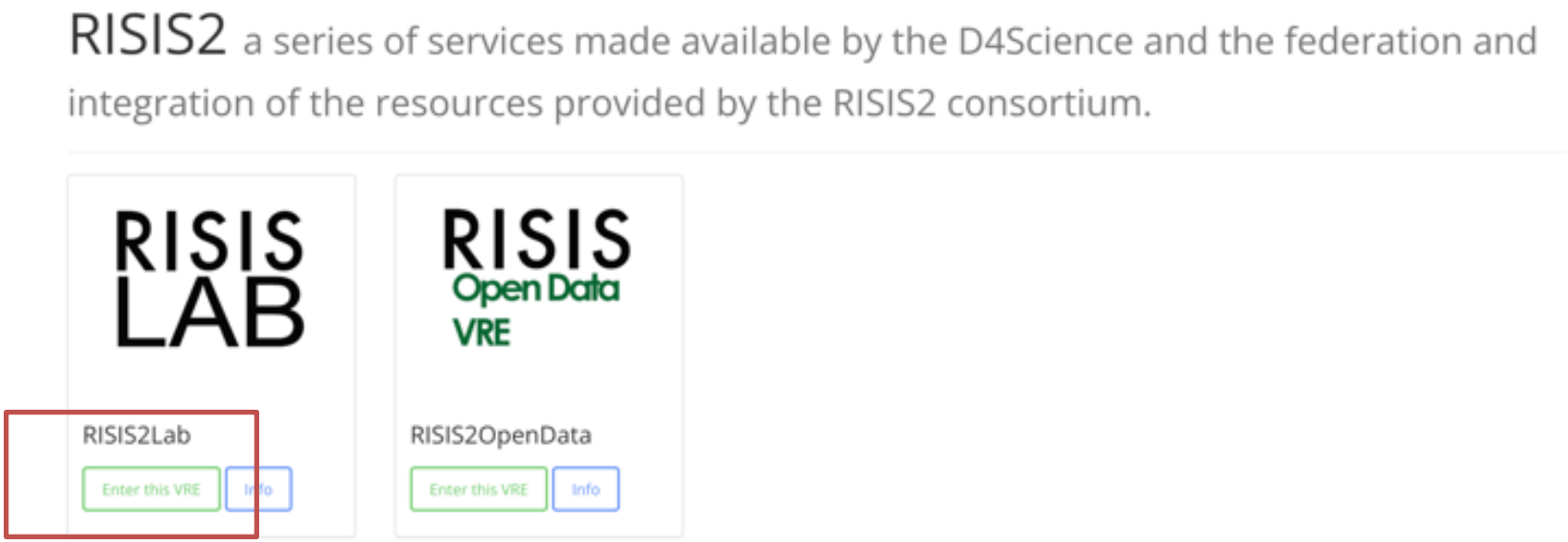


Request access, wait for the acceptance email. 10 minutes later you should be able to enter the VRE.

Get into the VRE

RISIS

- Click on « Enter this VRE »
- You can also access directly to https://risis2.d4science.org/group/risis2lab



Agenda

RISIS

12:35 – 13:00 OpenAIRE Research Graph

- Concept and data model
- The json dumps

13:00 – 14:15 Practical session

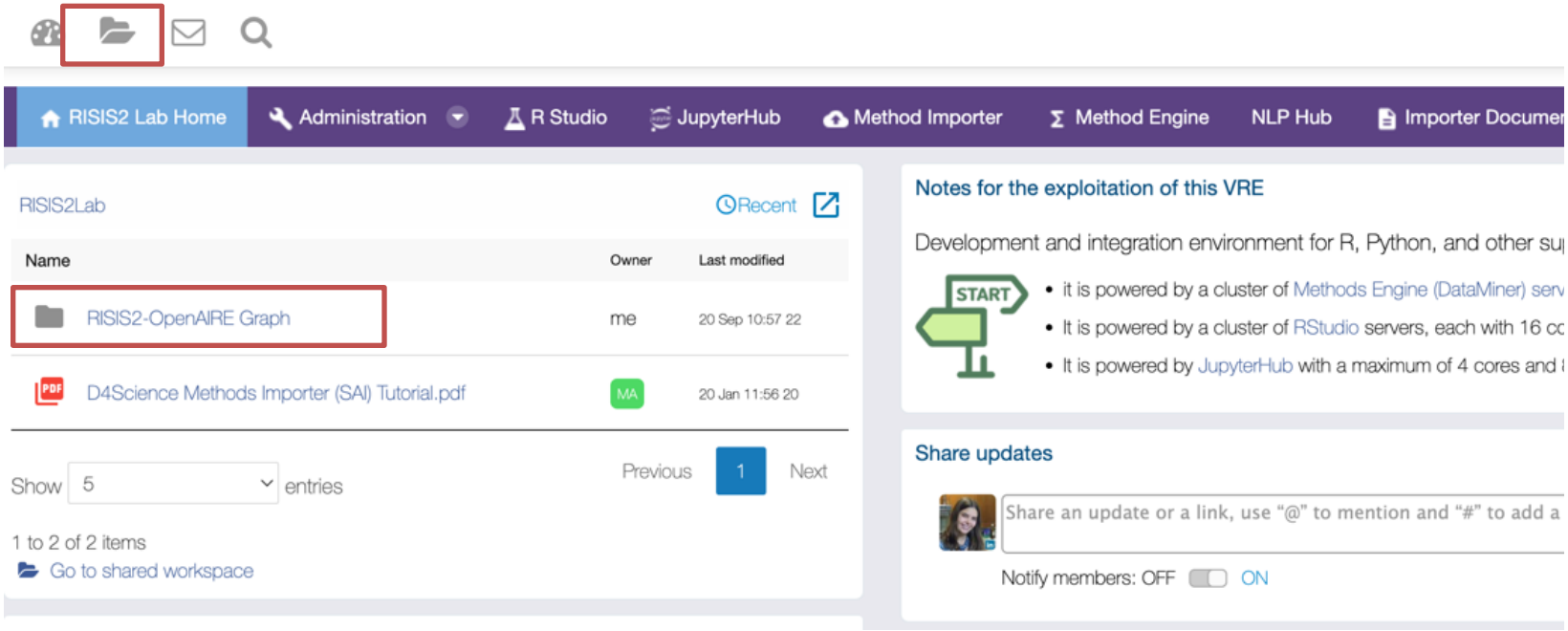
- Dataset description (selection of the subset, format)
- Dataset queries with JupyterHub

14:15 – 14:30 Final discussion

Supporting Material

RISIS

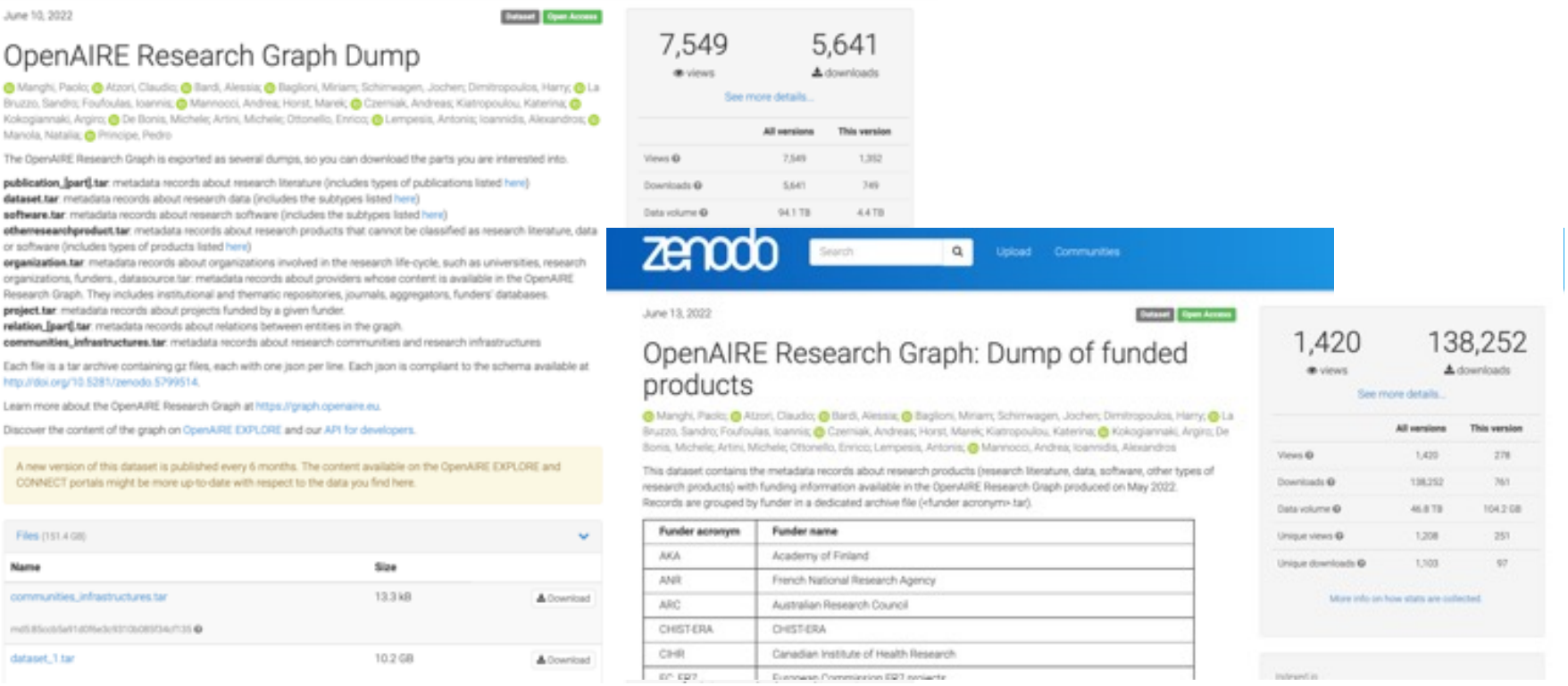
- You can find everything in the VRE workspace, under the folder RISIS2-OpenAIRE Graph:
  - Slides
  - Examples



Subset of the dump

RISIS

Because the whole [OpenAIRE Research graph Dump](#) is 4.4TB and the [dump of funded products](#) is 104GB and we do not want to lose time in copying data or waiting too much to get results of our queries



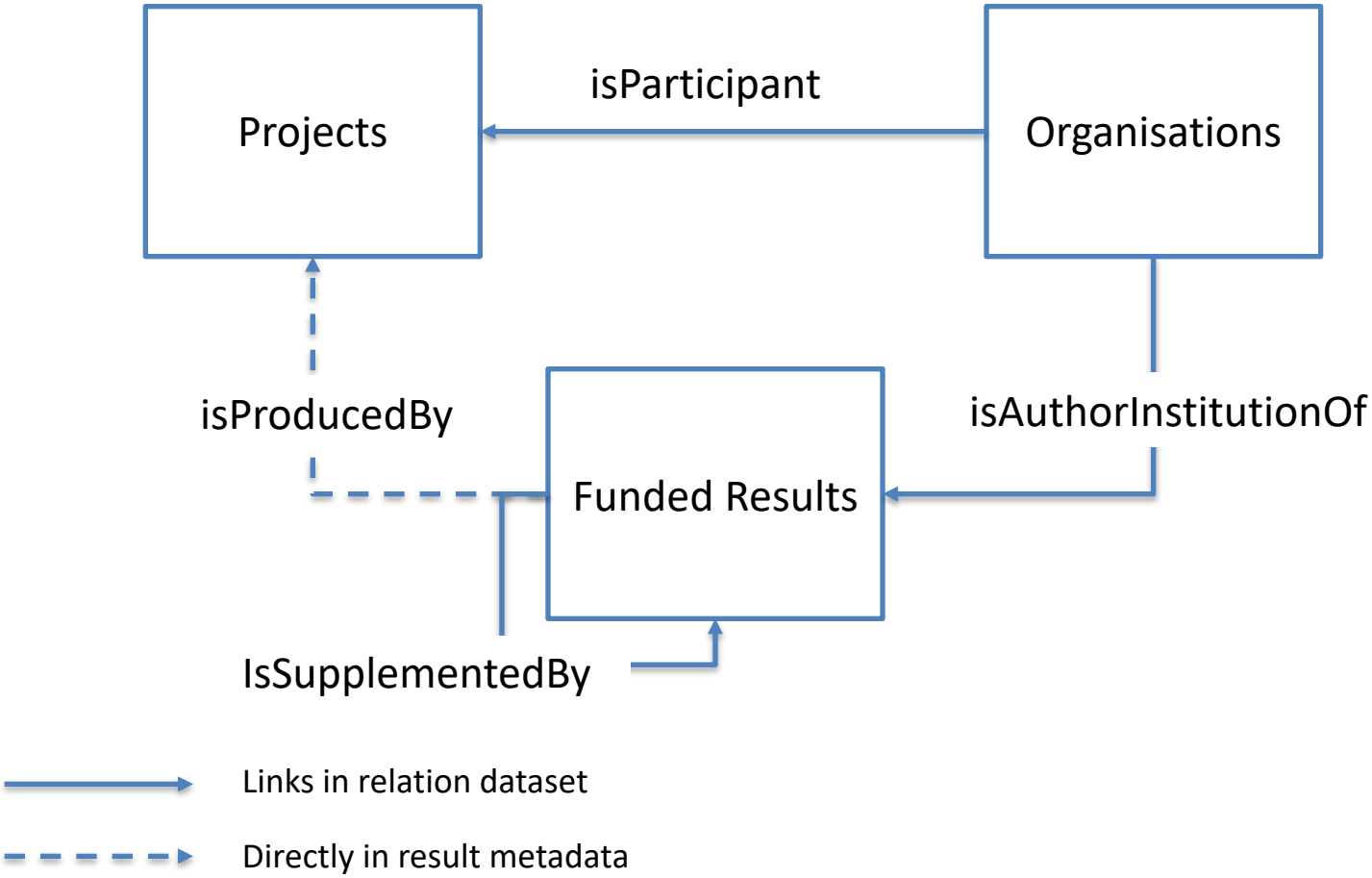
Selection of the subset

RISIS

- H2020 funded products
- Subset of products funded by a selection of 848 projects: 17369 research results
- Relationships from the whole dump:
  - Organization <isAuthorInstitutionOf> Result (where Result in the subset above)
  - Organisation <isParticipant> Project
  - Result <isSupplementedBy> Result (where both in the subset above)

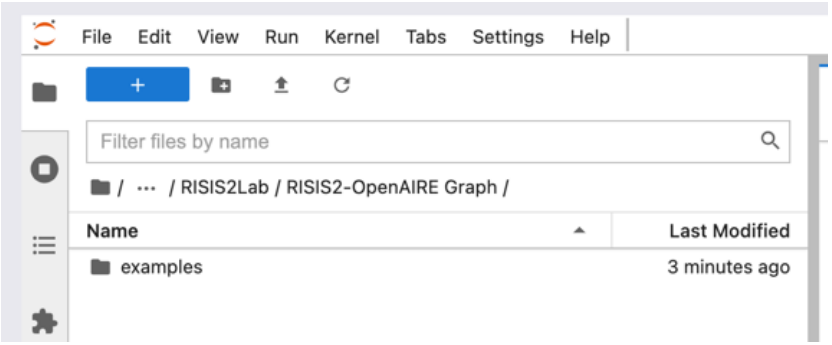
Data model

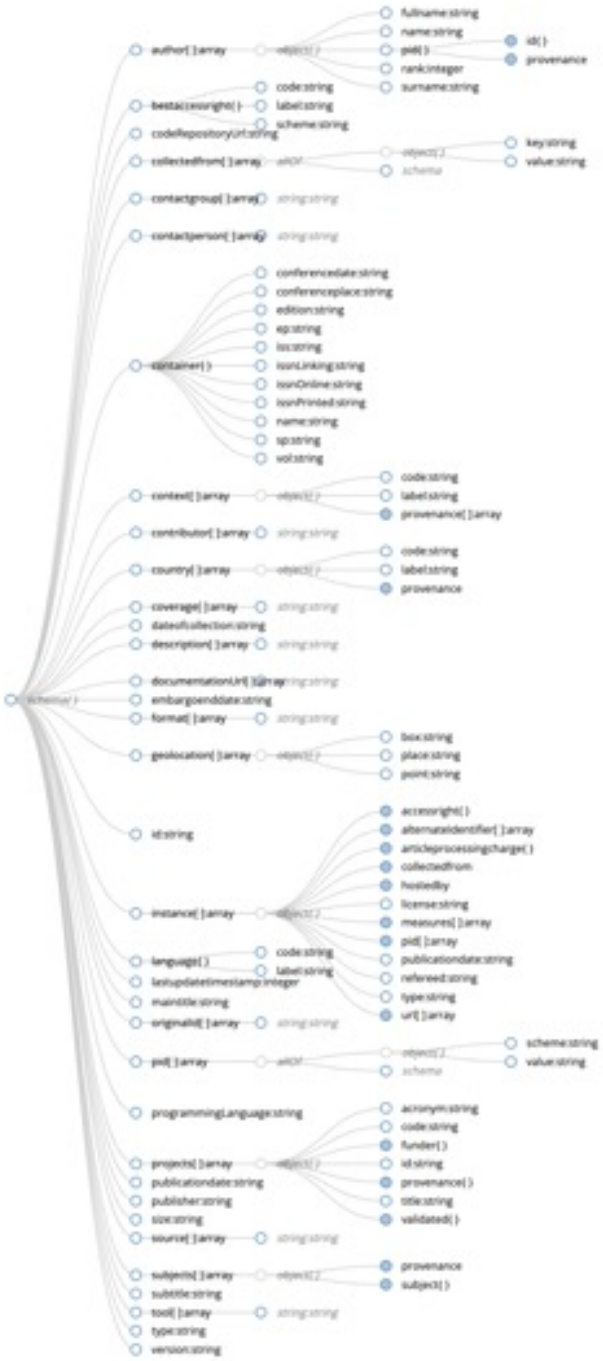
RISIS



Json schema for funded results

- Funded result schema: 10.5281/zenodo.6372977
- For a “nice” visualization you can download it and use jsonschemaviewer
- See fundedresult-example.json in /workspace/VREFolders/RISIS2Lab/RISIS2-OpenAIRE Graph





Metadata fields we will use: type

RISIS

- The type of the research product: publication, dataset, software, other types of research products
- Do not confuse with instance.type, which is the specific type of a version of the product (e.g. Article, Pre-print, Poster)

Metadata fields we will use: subjects

RISIS

- Keywords associated to the result. May be free text or terms from controlled vocabularies. May be harvested from the sources, added by MAG, or inferred by OpenAIRE


```
"subjects": [
    {
    "provenance": { "trust": "0.9", "provenance": "Harvested"},
      "subject": {"scheme": "keyword", "value": "Nature-Based Solutions"}
    },
    {
    "provenance": {"trust": "0.39543077", "provenance": "Harvested"},
    "subject": {"scheme": "MAG", "value": "business"}
    },
    {
    "provenance": {"provenance": "Inferred by OpenAIRE", "trust": "0.891"},
    "subject": {"scheme": "mesh", "value": "bacterial infections and mycoses"}
    }
]
```

Metadata fields we will use: container

RISIS

- Conference or journal where the result has been presented or published


```
"container": {
    "issnPrinted": "2071-1050",
    "name": "Sustainability",
    "vol": "",
    "sp": "",
    "iss": "",
    "edition": "",
    "issnOnline": "",
    "ep": "",
    "issnLinking": ""
  },
```

Metadata fields we will use: authors

RISIS

- Authors of research products. They are ordered and may come with an orcid (persistent identifier) or not.


```
"author": [
    {
        "fullname": "Gerd Lupp","rank": 1
    },
    {
        "fullname": "Aude Zingraff-Hamed",
        "pid": {
            "provenance": {
                "trust": "0.91",
                "provenance": "Harvested"
            },
            "id": {
                "scheme": "orcid",
                "value": "0000-0001-7602-7830"
            }
        },
        "rank": 2
    }
],
```

Metadata fields we will use: bestaccessright and publicationdate

RISIS

- Bestaccessright: The most open access right among those of the different versions of this result. Codes and semantics from https://vocabularies.coar-repositories.org/access_rights/


```
"bestaccessright": {
    "scheme": "http://vocabularies.coar-repositories.org/documentation/access_rights/",
    "code": "c_abf2",
    "label": "OPEN"
  },
```


- Publicationdate: main date of the research product. Typically the date of publishing, but it could also be the “issued date” or the deposition date of the pre-print (I know…metadata is not as clean as we would like…)

Metadata fields we will use: projects

RISIS

- Project grant in the context of which the result was produced. The association to a project may come from a source (Harvested), from a user (Linked by user), or inferred by OpenAIRE. If validated, it means that it was confirmed by the funder (SyGMA portal for EC projects)

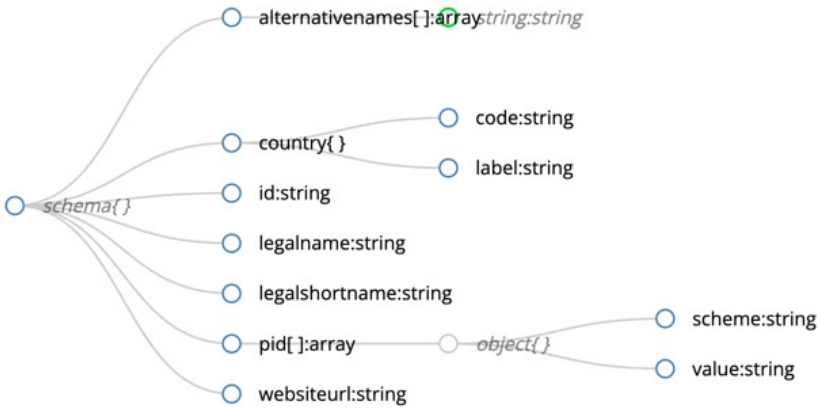

```
"projects": [
    {
        "code": "776681",
        "title": "PHUSICOS: 'According to nature' - solutions to reduce risk in mountain landscapes",
        "acronym": "Phusicos",
        "provenance": {"trust": "0.900","provenance": "Harvested"},
        "funder": {
            "shortName": "EC",
            "jurisdiction": "EU",
            "name": "European Commission",
            "fundingStream": "H2020"
        },
        "validated": {"validationDate": "2021-10-20","validatedByFunder": true},
        "id": "40lcorda__h2020::16a8d0f2527d083af2c361529f982e96"
    }
],
```

Json schema for organisations

RISIS

- Organisation schema: 10.5281/zenodo.5799514
- See organisation-example.json

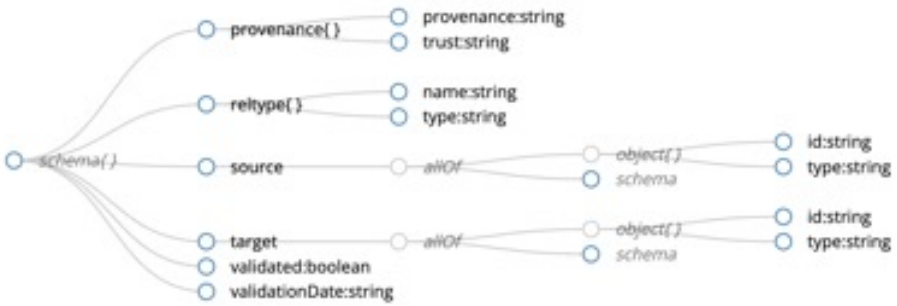

```
{
  "legalshortname": "IFW",
  "country": {"code": "DE","label": "Germany"},
  "pid": [
    {"scheme": "ISNI","value": "0000 0000 9972 3583"},
    { "scheme": "OrgRef","value": "25588007"},
    {"scheme": "GRID","value": "grid.14841.38"},
    {"scheme": "ROR","value": "https://ror.org/04zb59n70"},
    {"scheme": "Wikidata","value": "Q835883"},
    {"scheme": "PIC","value": "999544746"},
    {"scheme": "OrgReg","value": "DE1163"}
  ],
  "websiteurl": "http://www.ifw-dresden.de/",
  "legalname": "Leibniz Institute for Solid State and Materials Research",
  "alternativenames": [
    "IFW Dresden",
    "Leibniz Institute for Solid State and Materials Research",
    "Leibniz-Institut für Festkörper- und Werkstoffforschung Dresden",
    "IFW"
  ],
  "id": "20|openorgs____::b26c5d36199137435bc577f28a3fef5e"
}
```

Json schema for relations

RISIS

- Relation schema: 10.5281/zenodo.6372977
- Examples: Supplement-example.json, Participation-example.json, Affiliation-example.json

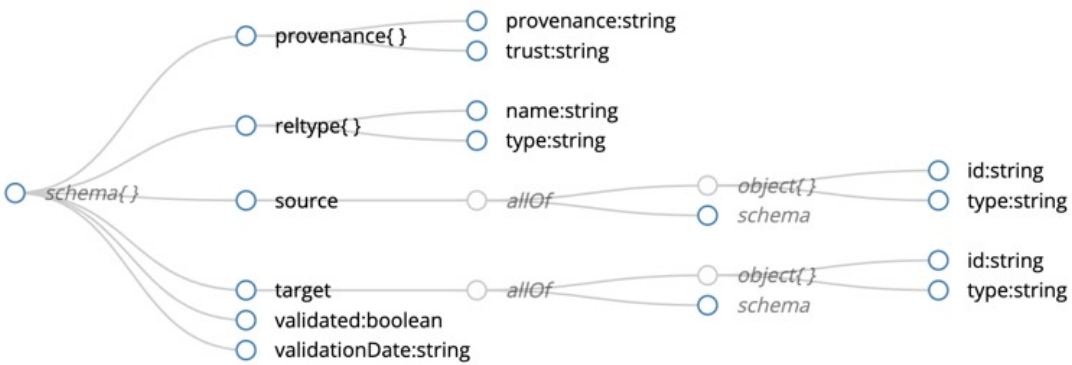



```
{
    "source": {
        "type": "result",
        "id": "50ldoi_________::39af87f0d1d69f135e7e5f3a41a77220"
    },
    "reltype": {"type": "supplement","name": "IsSupplementedBy"},
    "provenance": {"provenance": "Harvested","trust": "0.9"
    },
    "validated": false,
    "target": {
        "type": "result",
        "id": "50ldoi_dedup___::17d8bd9c2095e56b9a0017b71cb0505f"
    }
}
```

Json schema for relations

RISIS

- Relation schema: 10.5281/zenodo.6372977
- A relation may come from a source (Harvested), from a user (Linked by user), or inferred by OpenAIRE.
- Source and target tells you the type of object (result, organization, project) and its OpenAIRE id
- reltype.type tells you the “type” of relation (e.g. supplement). Specific semantics is in reltype.name (e.g. IsSupplementTo)



References to controlled vocabulary

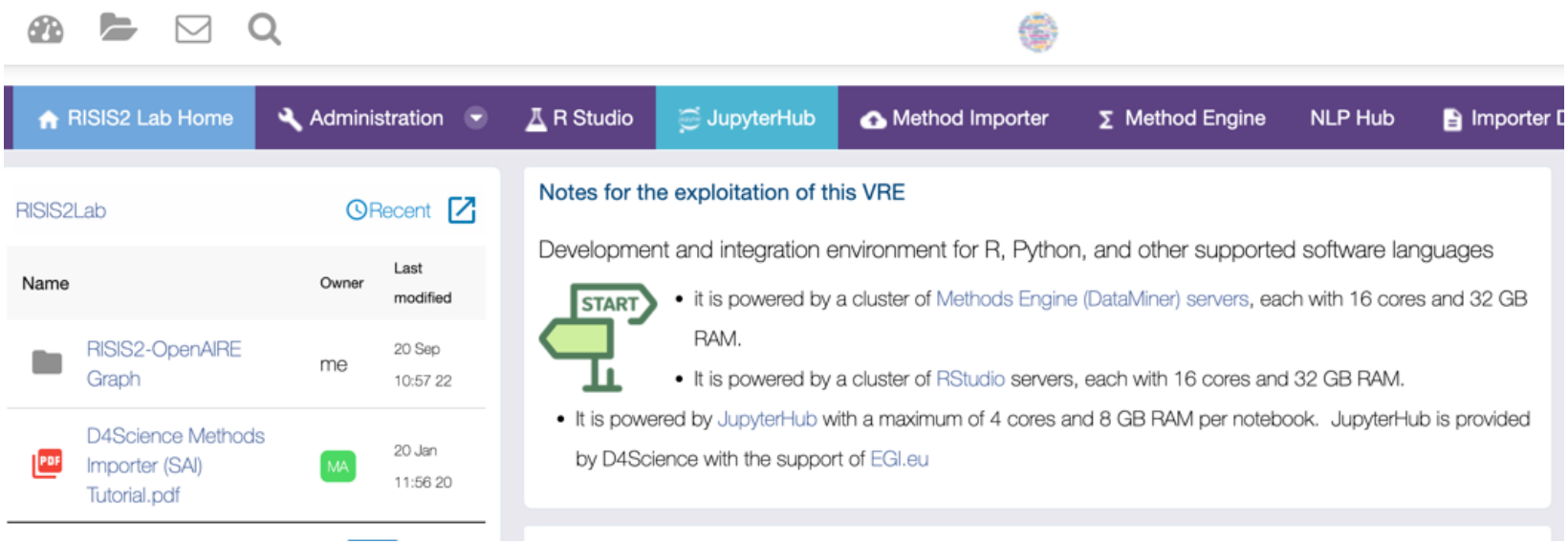
RISIS

- https://api.openaire.eu/vocabularies/
- In the dump you'll find the terms, not the codes (that are useful mostly for internal use)
- Of particular interest:
  - dnet:provenanceActions: details about provenance of the records, properties and relationships
  - dnet:result_typologies: types of research results
  - dnet:subject_classification_typologies: subject classification schemes
  - dnet:review_levels: peer reviewed or not?
  - dnet:pid_types: types of PIDs you may find
  - dnet:access_modes: access rights

Ready for coding?

RISIS

https://risis2.d4science.org/group/risis2lab and go to the JupyterHub

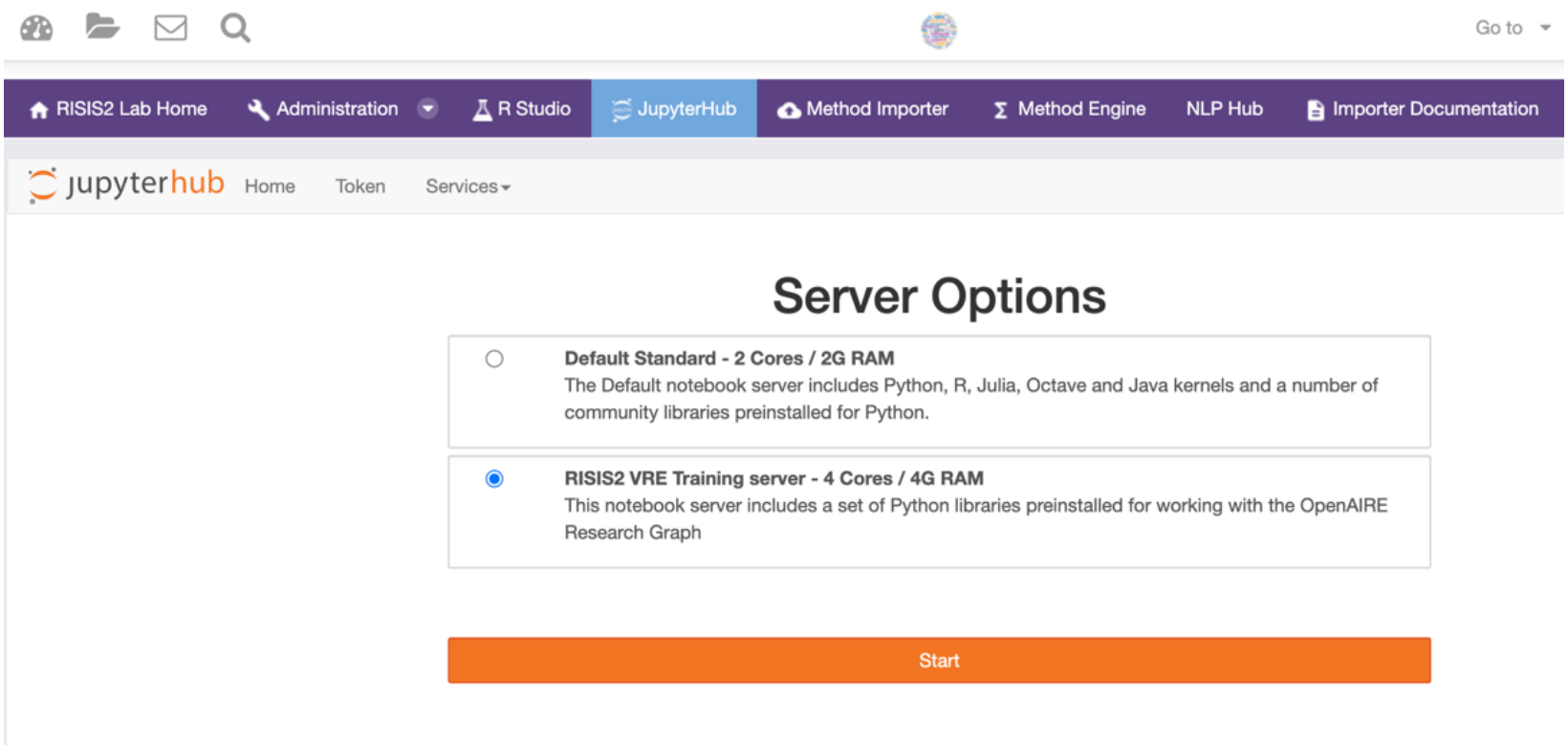


D4SCIENCE INFRASTRUCTURE

Get your server

RISIS

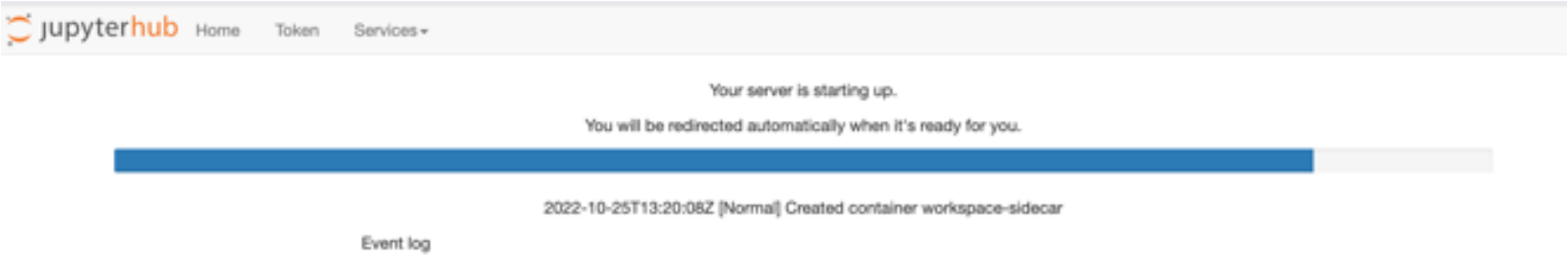
Start the configuration for the RISIS2 VRE Training Server



Server ready?

RISIS

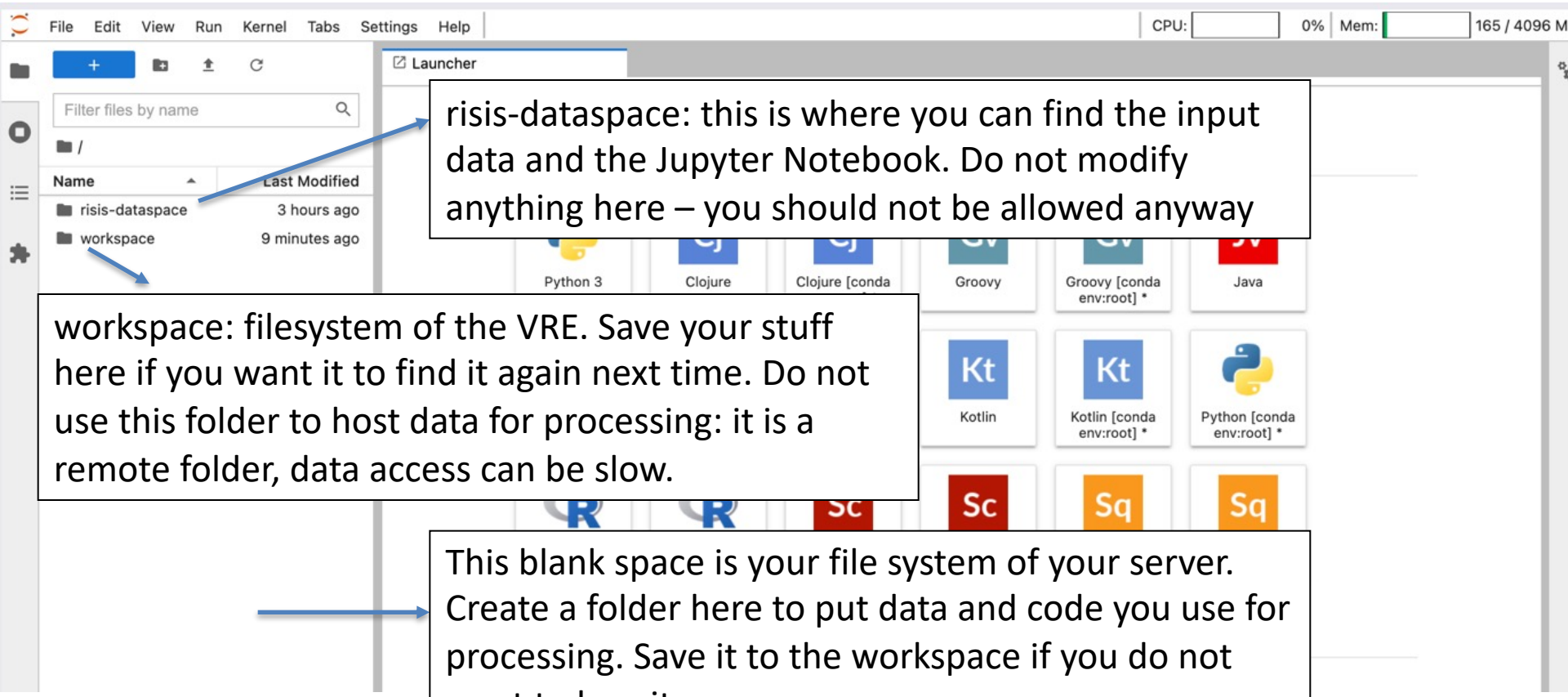
Give it few seconds to set everything up



What do you see?

RISIS

JupyterLab GUI. General documentation in “Help”



risis-dataspace: this is where you can find the input data and the Jupyter Notebook. Do not modify anything here – you should not be allowed anyway

workspace: filesystem of the VRE. Save your stuff here if you want it to find it again next time. Do not use this folder to host data for processing: it is a remote folder, data access can be slow.

This blank space is your file system of your server. Create a folder here to put data and code you use for processing. Save it to the workspace if you do not want to lose it.

Get the data and the notebook

RISIS

- Open risis-dataspace
- Copy to your local file system (i.e. at the same level of workspace and risis-dataspace:
  - risis-dataset.zip
  - risis.ipynb
- Before leaving the JupyterLab remember to save everything into the workspace

risis_dataset.zip

RISIS

```
jovyan@jupyter-alessia-2ebardi:~/risis_dataset$ ls -l
total 260
drwxr-xr-x 2 jovyan users 77824 Oct 25 13:30 organizations
```
→ metadata about orgs

```
drwxr-xr-x 2 jovyan users 40960 Oct 25 13:31 rel_affiliation
```
→ affiliation rels between results and orgs

```
drwxr-xr-x 2 jovyan users 40960 Oct 25 13:31 rel_participant
```
→ rels between orgs and projects

```
drwxr-xr-x 2 jovyan users    55 Oct 25 13:30 rel_supplement
```
→ rels between research results

```
drwxr-xr-x 2 jovyan users 12288 Oct 25 13:31 results
```
→ metadata about results

Useful documentation

RISIS

- PySpark: https://spark.apache.org/docs/latest/api/python/index.html
- D4Science infrastructure: https://www.d4science.org/
- JupyterLab: https://docs.jupyter.org/en/latest/ (also integrated into the JupyterHub)

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

THANK YOU !

www.risis2.eu

CONTACT@RISIS2.EU

@RISIS_EU

FACEBOOK.COM/RISIS.EU

RISIS2 EU PROJECT

UPEM Université Paris-Est Marne-la-Vallée · AIT Austrian Institute of Technology · Politecnico Milano 1863 · Consiglio Nazionale delle Ricerche · Universiteit Leiden · The University of Sheffield. · Fraunhofer ISI · JOANNEUM RESEARCH · University of Sussex · DZHW · University of Strathclyde Glasgow · Sapienza Università di Roma · NIFU Nordisk institutt for studier av innovasjon, forskning og utdanning · CSIC · Samuel Neaman Institute for National Policy Research · Technology Centre ASCR · ATHENA Research & Innovation Information Technologies