

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



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

About

RISIS LAB

An environment for collaborative analysis and discussion of RISIS contributions to OpenAIRE and to deliver a computational environment for exploration and validation of datasets dedicated to OpenAIRE communities.

Request access (if needed)



<https://risi2.d4science.org/explore>

RISIS2 a series of services made available by the D4Science and the federation and integration of the resources provided by the RISIS2 consortium.

A screenshot of the RISIS2 services interface. It features two main service cards. The left card is for 'RISIS LAB' and has a red rectangular box highlighting the 'Request Access' button and the 'Info' button. The right card is for 'RISIS Open Data VRE' and has a blue rectangular box highlighting the 'Info' button. Below the logos, the text 'RISIS2Lab' and 'RISIS2OpenData' is visible, along with 'Private' and 'Info' buttons for the second service.

RISIS LAB

RISIS Open Data VRE

RISIS2Lab

Request Access Info

RISIS2OpenData

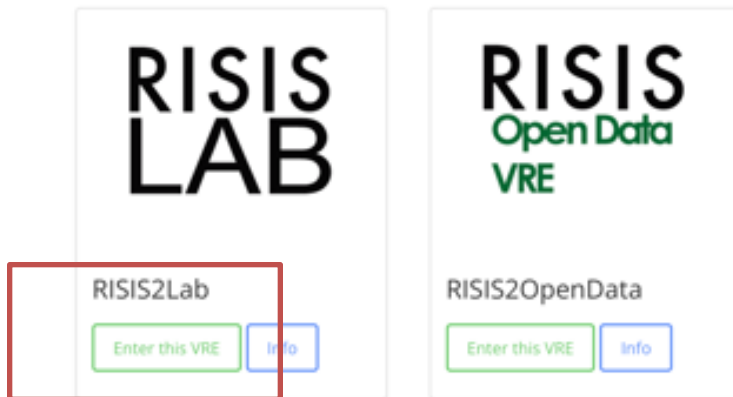
Private Info

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

Get into the VRE

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

RISIS2 a series of services made available by the D4Science and the federation and integration of the resources provided by the RISIS2 consortium.



Agenda

12:35 – 13:00 OpenAIRE Research Graph

- Concept and data model
- The json dumps

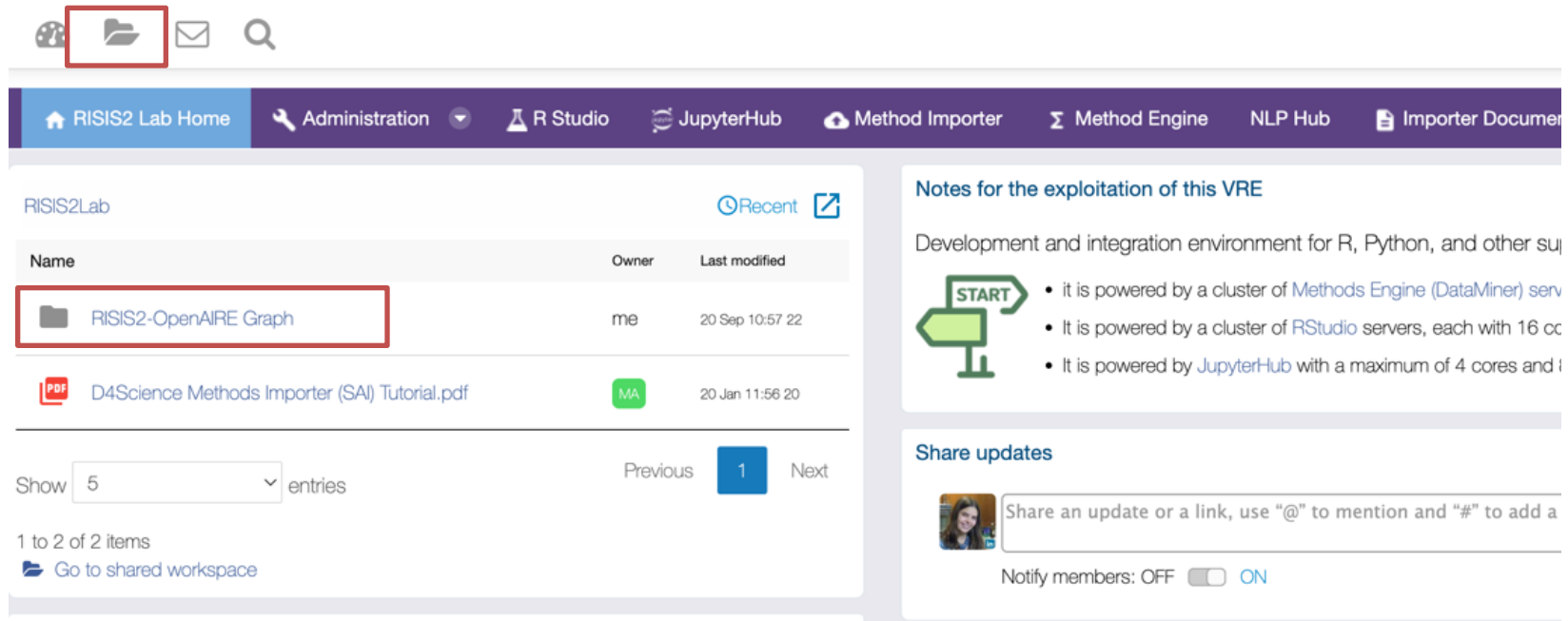
13:00 – 14:15 Practical session

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



14:15 – 14:30 Final discussion

Supporting Material

- You can find everything in the VRE workspace, under the folder [RISIS2-OpenAIRE Graph](#):
 - Slides
 - Examples



The screenshot shows the RISIS2 Lab interface. At the top, there is a navigation bar with links for RISIS2 Lab Home, Administration, R Studio, JupyterHub, Method Importer, Method Engine, NLP Hub, and Importer Document. Below the navigation bar, the main content area displays a file browser for 'RISIS2Lab'. A table lists the files and folders:

Name	Owner	Last modified
 RISIS2-OpenAIRE Graph	me	20 Sep 10:57 22
 D4Science Methods Importer (SAI) Tutorial.pdf	MA	20 Jan 11:56 20

Below the table, there is a 'Show 5 entries' dropdown, 'Previous 1 Next' navigation, and '1 to 2 of 2 items' with a 'Go to shared workspace' link. On the right side, there is a 'Notes for the exploitation of this VRE' section with a 'START' button and a list of bullet points. Below that is a 'Share updates' section with a text input field and a 'Notify members' toggle switch.

Subset of the dump

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

June 10, 2022 Dataset Open Access

OpenAIRE Research Graph Dump

● Manghi, Paolo; ● Atzori, Claudio; ● Bard, Alessia; ● Baglioni, Miriam; Schimwegen, Jochen; Dimitropoulos, Harry; ● La Bruzzo, Sandro; Fofoulas, Ioannis; ● Mannocci, Andrea; Horst, Marek; ● Czerniak, Andreas; Kiatropoulou, Katerina; ● Kokogiannaki, Argiro; ● De Bonis, Michele; Arini, Michele; Ottonello, Enrico; ● Lempessa, Antonio; Ioannidis, Alexandros; Manola, Natalia; ● Principe, Pedro

The OpenAIRE Research Graph is exported as several dumps, so you can download the parts you are interested into.

- publication_part.tar**: metadata records about research literature (includes types of publications listed [here](#))
- dataset.tar**: metadata records about research data (includes the subtypes listed [here](#))
- software.tar**: metadata records about research software (includes the subtypes listed [here](#))
- otherresearchproduct.tar**: metadata records about research products that cannot be classified as research literature, data or software (includes types of products listed [here](#))
- organization.tar**: metadata records about organizations involved in the research life-cycle, such as universities, research organizations, funders, **datasource.tar**: metadata records about providers whose content is available in the OpenAIRE Research Graph. They include institutional and thematic repositories, journals, aggregators, funders' databases.
- project.tar**: metadata records about projects funded by a given funder.
- relation_part.tar**: metadata records about relations between entities in the graph
- communities_infrastructures.tar**: metadata records about research communities and research infrastructures

Each file is a tar archive containing gz files, each with one json per line. Each json is compliant to the schema available at <http://dx.doi.org/10.5281/zenodo.5799514>.

Learn more about the OpenAIRE Research Graph at <https://graph.openaire.eu>.

Discover the content of the graph on [OpenAIRE EXPLORE](#) and our [API for developers](#).

7,549 views 5,641 downloads

[See more details...](#)

	All versions	This version
Views	7,549	1,352
Downloads	5,641	749
Data volume	94.1 TB	4.4 TB

zenodo Search Upload Communities

June 13, 2022 Dataset Open Access

OpenAIRE Research Graph: Dump of funded products

● Manghi, Paolo; ● Atzori, Claudio; ● Bard, Alessia; ● Baglioni, Miriam; Schimwegen, Jochen; Dimitropoulos, Harry; ● La Bruzzo, Sandro; Fofoulas, Ioannis; ● Czerniak, Andreas; Horst, Marek; Kiatropoulou, Katerina; ● Kokogiannaki, Argiro; De Bonis, Michele; Arini, Michele; Ottonello, Enrico; Lempessa, Antonio; ● Mannocci, Andrea; Ioannidis, Alexandros

This dataset contains the metadata records about research products (research literature, data, software, other types of research products) with funding information available in the OpenAIRE Research Graph produced on May 2022. Records are grouped by funder in a dedicated archive file (=funder acronym*.tar).

Funder acronym	Funder name
AKA	Academy of Finland
ANR	French National Research Agency
ARC	Australian Research Council
CHISTERA	CHISTERA
CIHR	Canadian Institute of Health Research
EC-EP7	European Commission EP7 contracts

1,420 views 138,252 downloads

[See more details...](#)

	All versions	This version
Views	1,420	278
Downloads	138,252	761
Data volume	46.8 TB	104.2 GB
Unique views	1,208	251
Unique downloads	1,703	97

[More info on how stats are collected](#)

Inferred on

A new version of this dataset is published every 6 months. The content available on the OpenAIRE EXPLORE and CONNECT portals might be more up-to-date with respect to the data you find here.

Files (151.4 GB)

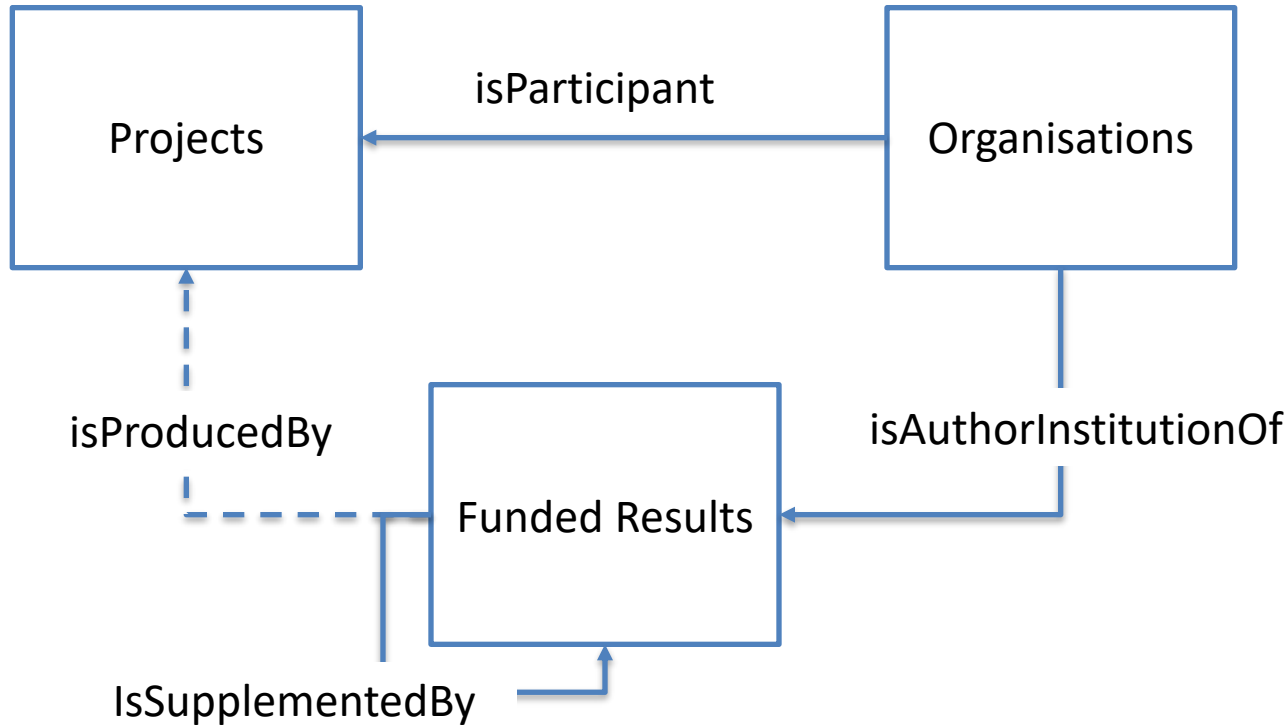
Name	Size	Download
communities_infrastructures.tar	13.3 kB	Download
md5:850c0da1140f4c3c03703d8934a7135		
dataset_1.tar	10.2 GB	Download

Selection of the subset



- 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



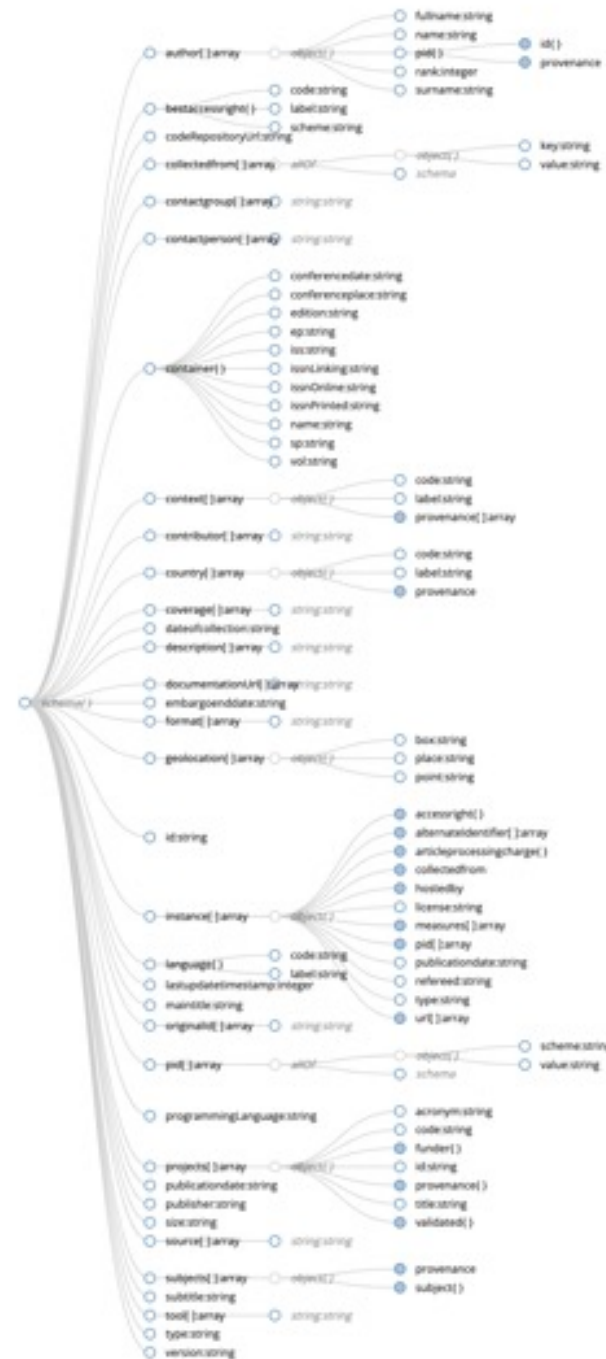
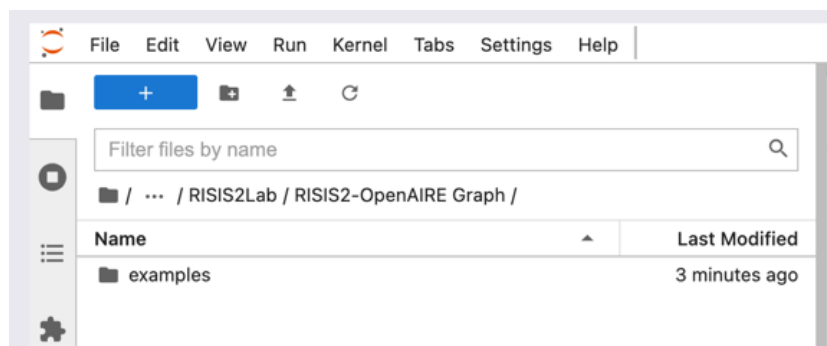
Links in relation dataset



Directly in result metadata

Json schema for funded results

- Funded result schema: [10.5281/zenodo.6372977](https://doi.org/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



- 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



- 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



- 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

- 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



- 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



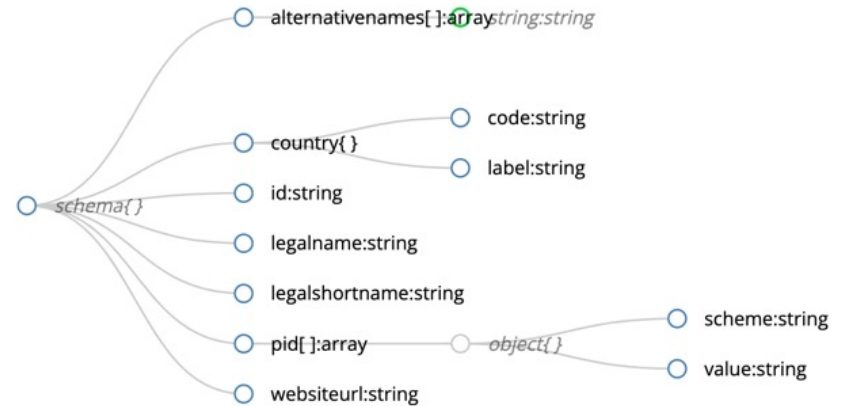
- Project grant in the context of which the result was produced. The association to a project may come from a source (Harvested), from a user (Linked by user), or inferred by OpenAIRE. If validated, it means that it was confirmed by the funder (SyGMA portal for EC projects)

```
"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": "40l corda__h2020::16a8d0f2527d083af2c361529f982e96"  
  },  
],
```


Json schema for organisations

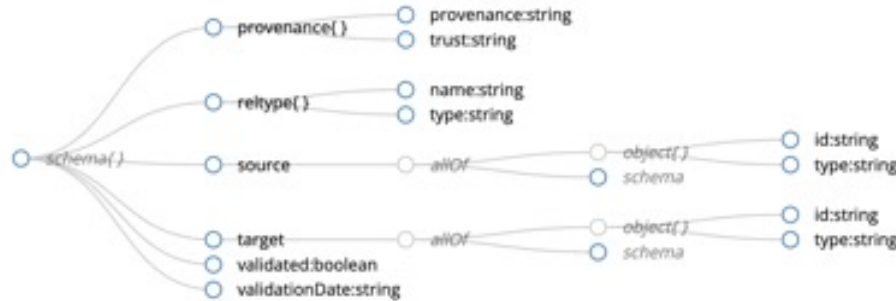
- Organisation schema: [10.5281/zenodo.5799514](https://zenodo.org/record/5799514/files/10.5281/zenodo.5799514)
- See `organisation-example.json`

```
{  
  "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 Werkstofforschung Dresden",  
    "IFW"  
  ],  
  "id": "20lopenorgs_____:b26c5d36199137435bc577f28a3fef5e"  
}
```



Json schema for relations

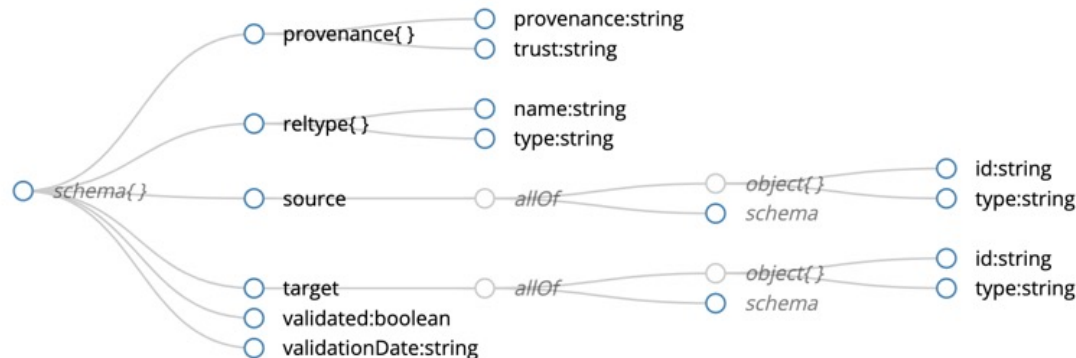
- Relation schema: [10.5281/zenodo.6372977](https://zenodo.org/doi/10.5281/zenodo.6372977)
- Examples: Supplement-example.json, Participation-example.json, Affiliation-example.json



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

Json schema for relations

- Relation schema: [10.5281/zenodo.6372977](https://zenodo.org/doi/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



- <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?



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

RISIS2Lab Recent

Name	Owner	Last modified
RISIS2-OpenAIRE Graph	me	20 Sep 10:57 22
D4Science Methods Importer (SAI) Tutorial.pdf	MA	20 Jan 11:56 20

Notes for the exploitation of this VRE

Development and integration environment for R, Python, and other supported software languages

- it is powered by a cluster of **Methods Engine (DataMiner)** servers, each with 16 cores and 32 GB RAM.
- It is powered by a cluster of **RStudio** servers, each with 16 cores and 32 GB RAM.
- It is powered by **JupyterHub** with a maximum of 4 cores and 8 GB RAM per notebook. JupyterHub is provided by D4Science with the support of EGI.eu

RISIS Tool Demo Event #1 26 Oct |
Open Access Week 2022



Get your server



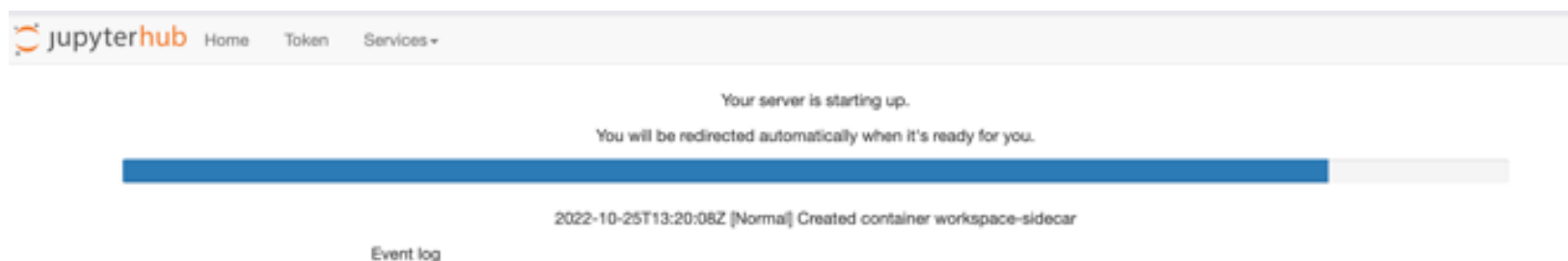
Start the configuration for the RISIS2 VRE Training Server

A screenshot of the JupyterHub web interface. At the top, there is a navigation bar with icons for home, folder, mail, and search, followed by a globe icon and a 'Go to' dropdown. Below this is a dark purple navigation menu with links for 'RISIS2 Lab Home', 'Administration', 'R Studio', 'JupyterHub' (highlighted), 'Method Importer', 'Method Engine', 'NLP Hub', and 'Importer Documentation'. Underneath is a lighter purple header with the 'jupyterhub' logo and links for 'Home', 'Token', and 'Services'. The main content area is titled 'Server Options' and contains two selectable options. The first option is 'Default Standard - 2 Cores / 2G RAM' with a radio button. The second option is 'RISIS2 VRE Training server - 4 Cores / 4G RAM' with a selected radio button. A large orange 'Start' button is positioned at the bottom of the options section.

RISIS Tool Demo Event #1 26 Oct |
Open Access Week 2022

Server ready?

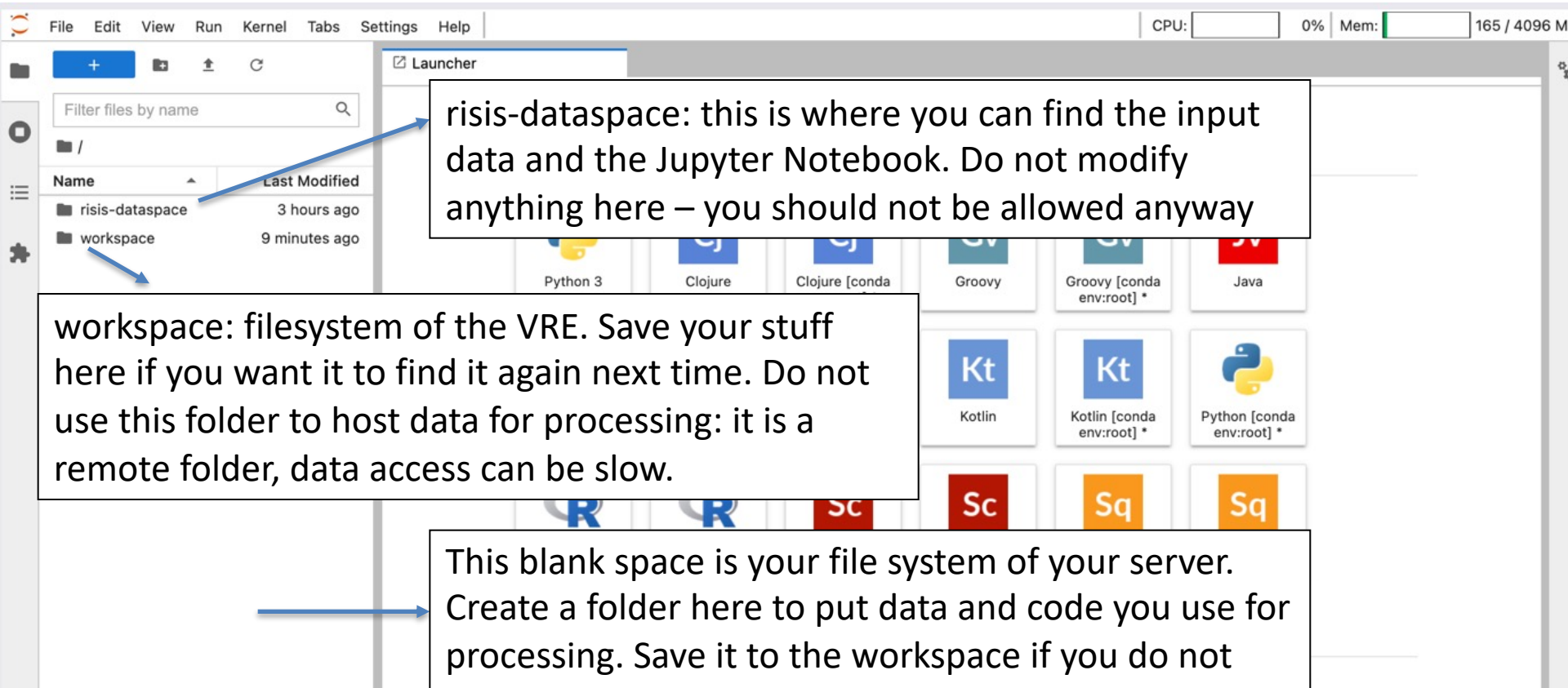
Give it few seconds to set everything up



The screenshot shows the JupyterHub user interface. At the top left, there is a navigation bar with the JupyterHub logo, a 'Home' link, a 'Token' link, and a 'Services' dropdown menu. The main content area displays a status message: 'Your server is starting up. You will be redirected automatically when it's ready for you.' Below this message is a blue progress bar that is approximately 80% full. Underneath the progress bar, there is an 'Event log' section with a single entry: '2022-10-25T13:20:08Z [Normal] Created container workspace-sidecar'.

What do you see?

JupyterLab GUI. General documentation in “Help”



Get the data and the notebook



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

risis_dataset.zip



```
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

- 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](https://twitter.com/@RISIS_EU)

FACEBOOK.COM/RISIS.EU



[RISIS2 EU PROJECT](https://www.youtube.com/RISIS2_EU_PROJECT)

