ExPaNDS

European Open Science Cloud Photon and Neutron Data Services

How to get data automatically from a DOI.

Paul Millar

2022-06-14
PaNOSC & ExPANDS face-to-face meeting
Prague, Czech Republic

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

What is a DOI?

- DOI: a “Data Object” Identifier.
- In this context, a DOI is a
  - Unique and persistence identifier
  - Represents (provides info about) a dataset
- As a human: view information about a DOI
- As a computer: can fetch metadata about DOI

This project receives funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 857641

How do we use a DOI?

- To find out information about a dataset.
- As a link between the dataset and other objects:

  papers, people, samples, instruments, software, …
- To download the data (files) in a dataset.
- DOIs look like:

  `doi:10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7`

  but equivalently like:

  `https://doi.org/10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7`

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

So, what does a DOI look like?

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Example from PSI

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PSI Public Data Repository

Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space

Karin Tran-Lundmark; PSI (2021)

Abstract

The aim of this proposal is to image the complex vascular lesions seen in human pulmonary hypertension (PH). High resolution 3D image data can by itself, or when combined with subsequent sectioning and immunohistochemical analyses of the same specimen, create a detailed map of the distribution of pathology in 3D space. This project has the potential to advance our understanding of PH pathophysiology and increase our chances of finding relevant targets for drug development.

Publication details

DOI	https://doi.org/10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7
Resource Type	raw
Related Publications	Westoo et al. American Journal of Physiology 2021

Datasets

Data Description human lung with plexiform lessions

20.500.11935/4a1a945b-a6aa-476c-a833-1754100355c8

Actions

To access the data associated with this DOI click below and follow the instructions

Access Data

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PSI Public Data Repository

Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space

Karin Tran-Lundmark; PSI (2021)

Abstract

The aim of this proposal is to image the complex vascular lesions seen in human pulmonary hypertension (PH). High resolution 3D image data can by itself, or when combined with subsequent sectioning and immunohistochemical analyses of the same specimen, create a detailed map of the distribution of pathology in 3D space. This project has the potential to advance our understanding of PH pathophysiology and increase our chances of finding relevant targets for drug development.

Publication details

DOI	https://doi.org/10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7
Resource Type	raw
Related Publications	Westoo et al. American Journal of Physiology 2021

Datasets

Data Description human lung with plexiform lessions

20.500.11935/4a1a945b-a6aa-476c-a833-1754100355c8

Actions

To access the data associated with this DOI click below and follow the instructions

Access Data

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PAUL SCHERRER INSTITUT

Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space : Download Page

Install the wget command for your platform if not yet available. Then type the following commands in your destination folder, which has enough capacity

```
cd destinationFolder
```

And then the transfer commands:

- `wget -m -np` https://doi2.psi.ch/datasets/sls/X02DA/Data10/e17068/disk1/h11913_4_3_/tif `( size: 16243212118 , nFiles: 2033 )`

You can simply repeat the wget command in case the connection is interrupted. In this case only files not yet downloaded will be fetched.

Cite as DOI: 10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PAUL SCHERRER INSTITUT

Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space : Download Page

Install the wget command for your platform if not yet available. Then type the following commands in your destination folder, which has enough capacity

```
cd destinationFolder
```

And then the transfer commands:

- `wget -m -np` https://doi2.psi.ch/datasets/sls/X02DA/Data10/e17068/disk1/h11913_4_3_/tif (size: 16243212118 , nFiles: 2033)

You can simply repeat the wget command in case the connection is interrupted. In this case only files not yet downloaded will be fetched.

Cite as DOI: 10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PAUL SCHERRER INSTITUT

Synchrotron Imaging of Complex Vascular Lesions in Human Pulmonary Hypertension: Pathology Distribution in 3D Space : Download Page

Install the wget command for your platform if not yet available. Then type the following commands in your destination folder, which has enough capacity

```
cd destinationFolder
```

And then the transfer commands

```
wget -nc https://doi2.psi.ch/datasets/sls/X02DA/Data10/e17068/disk1/h11913_4_3_/tif ( size: 16243212118 , nFiles: 2033 )
```

You can simply repeat the wget command in case the connection is interrupted. In this case only files not yet downloaded will be fetched.

Cite as DOI: 10.16907/d699e1f7-e822-4396-8c64-34ed405f07b7

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

Index of /datasets/sls/X02DA/Data10/e17068/disk1/h11913_4_3_/tif/

../		
__checksum_filename_0__	01-Jan-1970 00:00	132302
h11913_4_3_.log	08-Feb-2018 20:16	6817
h11913_4_3_.xml	08-Feb-2018 02:59	2523
h11913_4_3_0001.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0002.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0003.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0004.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0005.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0006.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0007.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0008.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0009.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0010.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0011.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0012.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0013.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0014.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0015.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0016.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0017.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0018.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0019.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0020.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0021.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0022.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0023.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0024.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0025.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0026.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0027.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0028.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0029.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0030.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0031.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0032.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0033.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0034.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0035.tif	08-Feb-2018 02:55	7997638

This page is auto-generated by ...

NGINX

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Index of /datasets/sls/X02DA/Data10/e17068/disk1/h11913_4_3_/tif/

../		
__checksum_filename_0__	01-Jan-1970 00:00	132302
h11913_4_3_.log	08-Feb-2018 20:16	6817
h11913_4_3_.xml	08-Feb-2018 02:59	2523
h11913_4_3_0001.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0002.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0003.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0004.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0005.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0006.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0007.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0008.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0009.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0010.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0011.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0012.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0013.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0014.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0015.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0016.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0017.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0018.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0019.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0020.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0021.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0022.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0023.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0024.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0025.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0026.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0027.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0028.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0029.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0030.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0031.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0032.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0033.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0034.tif	08-Feb-2018 02:55	7997638
h11913_4_3_0035.tif	08-Feb-2018 02:55	7997638

This page is auto-generated by ...

NGINX

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

File: __checksum_filename_0__

```
# Used python version for creating this checksum file: 3.6.[...]
h11913_4_3_1618.tif     2944f07af73e2cb3b0f78afbf0b9b325f593051e
h11913_4_3_0604.tif     c6f6e99bdbeb28ec77104edc9fb9a3b6d1960d1f
h11913_4_3_1710.tif     4d652e9f2a7954f5f5eb20aae2928bb6749284bc
h11913_4_3_1364.tif     c823b5fcff4383c04bc599c489fa51ee535d1f5f
h11913_4_3_1551.tif     4f128642f65b8eb44434f2d350077fb52c869386
h11913_4_3_1018.tif     45c5eca1d6cd7ef3c90b5e9846b7cddbf9e96b9e
h11913_4_3_0298.tif     cbdf82dcf0d4dd5242cad94390f654ba2cff1f60
h11913_4_3_1582.tif     c5d76cd240598e6dddbe92fd94d659895896fcd8
h11913_4_3_0209.tif     a87788a7ec4199a685fbadc51f2a3463d36b690e
h11913_4_3_1433.tif     dd35b064697201c4f9dd2d15a9dfa8e0f2d98f15
h11913_4_3_1240.tif     773f0bf1f28397ab174c72c74bea385583b1e87b
h11913_4_3_0344.tif     56377dc133d9af1e094834978803fc172bb62abb
[...]
```

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Example from ESRF

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

ESRF

DOI > 10.15151/ESRF-DC-818569841

Data collection

Dataset Open access

SERIAL DATA EXAMPLE

Gianluca Santoni ; Sylvain Aumonier.

DOI
DOI 10.15151/ESRF-DC-818569841

Licence (for files)
Creative Commons Attribution 4.0

Abstract
A fluorescent protein dataset to test EIGER-4M processing with crystfel

Proposals	Beamlines
MX-1887	ID30A-3

Publication year
2022

Experimental report

There is currently no experimental report.

Experimental data

The data can be accessed by clicking on the link below

Access data

Reference
Below is the recommended format for citing this work in a research publication.

Santoni G., Aumonier S. (2022). Serial data example. European Synchrotron Radiation Facility (ESRF). doi:10.15151/ESRF-DC-818569841

European Synchrotron Radiation Facility

Access to data is governed by the ESRF data policy .

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

ESRF

DOI > 10.15151/ESRF-DC-818569841

Data collection

Dataset Open access

SERIAL DATA EXAMPLE

Gianluca Santoni ; Sylvain Aumonier.

DOI
DOI 10.15151/ESRF-DC-818569841

Licence (for files)
Creative Commons Attribution 4.0

Abstract
A fluorescent protein dataset to test EIGER-4M processing with crystfel

Proposals	Beamlines
MX-1887	ID30A-3

Publication year
2022

Experimental report

There is currently no experimental report.

Experimental data

The data can be accessed by clicking on the link below

Access data

Reference
Below is the recommended format for citing this work in a research publication.

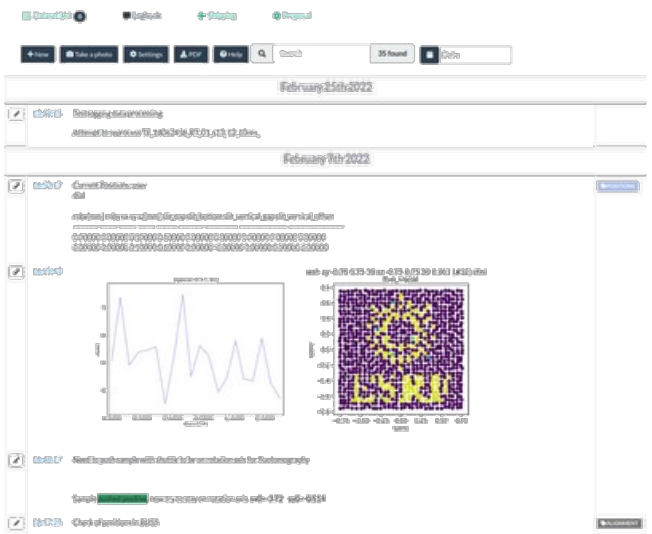
Santoni G., Aumonier S. (2022). Serial data example. European Synchrotron Radiation Facility (ESRF). doi:10.15151/ESRF-DC-818569841

European Synchrotron Radiation Facility

Access to data is governed by the ESRF data policy .

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641



Electronic logbook

Keep track of the experiment, so the data and metadata can be better understood and reused

SSO | Database

Sign in with ESRF SSO

or sign in as anonymous

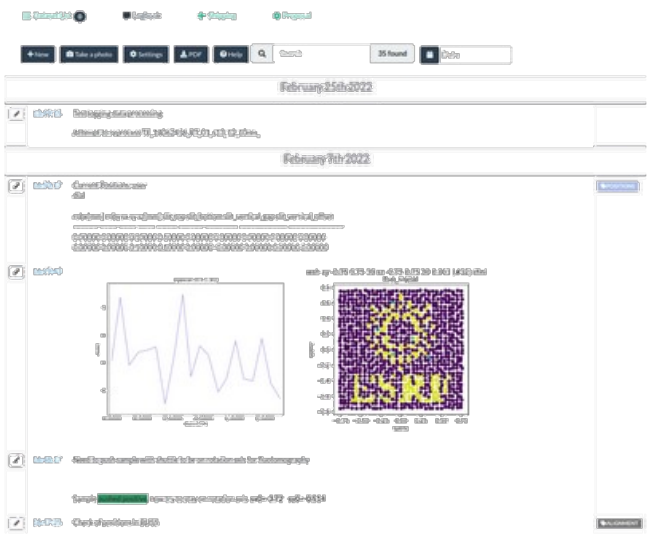
Don't have an account yet? Register now

I need further assistance

Important note

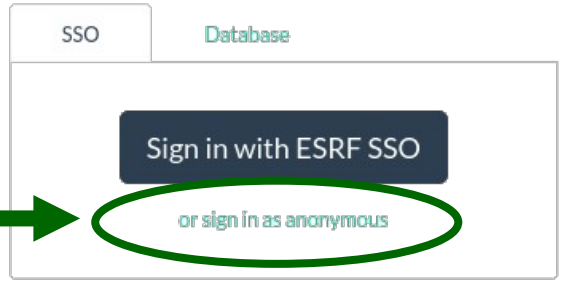
During 2019 and according to the General Data Protection Regulation, all portal users who did not consent to the User Portal Privacy Statement have had their account deactivated. Please contact the User Office if you wish to reactivate it.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641



Electronic logbook

Keep track of the experiment, so the data and metadata can be better understood and reused



Don't have an account yet? Register now

I need further assistance

Important note

During 2019 and according to the General Data Protection Regulation, all portal users who did not consent to the User Portal Privacy Statement have had their account deactivated. Please contact the User Office if you wish to reactivate it.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Data Portal My Data Open Data Closed Data Feedback Log out Anonymous

Open Data / 10.15151/ESRF-DC-818569841

Dataset List 1

Search

		Date	Sample	Dataset	Definition	Files	Size	Download
		14:38 25 Oct 2017	ceru_NaBr6	mesh-ceru_NaBr6_1_2134515		17	1.8 GB	Restore

10 Showing rows 1 to 1 of 1

1

ESRF European Synchrotron Radiation Facility

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Data Portal My Data Open Data Closed Data Feedback Log out Anonymous

Open Data / 10.15151/ESRF-DC-818569841

Dataset List 1

Search

	Date	Sample	Dataset	Definition	Files	Size	Download
	14:38 25 Oct 2017	ceru_NaBr6	mesh-ceru_NaBr6_1_2134515		17	1.8 GB	Restore

Summary Crystallography Instrument Files 17 Metadata List

Search

	Preview	Location	Size
		mesh-ceru_NaBr6_1_1_data_000001.h5	111.3 MB
		mesh-ceru_NaBr6_1_1_data_000002.h5	117.0 MB
		mesh-ceru_NaBr6_1_1_data_000003.h5	124.0 MB
		mesh-ceru_NaBr6_1_1_data_000004.h5	134.8 MB
		mesh-ceru_NaBr6_1_1_data_000005.h5	138.3 MB
		mesh-ceru_NaBr6_1_1_data_000006.h5	134.6 MB
		mesh-ceru_NaBr6_1_1_data_000007.h5	135.1

This project r ... ame under grant agreement No 857641

ExPaNDS European Open Science Cloud Photon and Neutron Data Services

OK, so why this is a problem?

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

OK, so why this is a problem?

- Lots of clicking just to get at the data.
- Assumes the user has a web-browser (with JavaScript).
- Makes automation almost impossible.

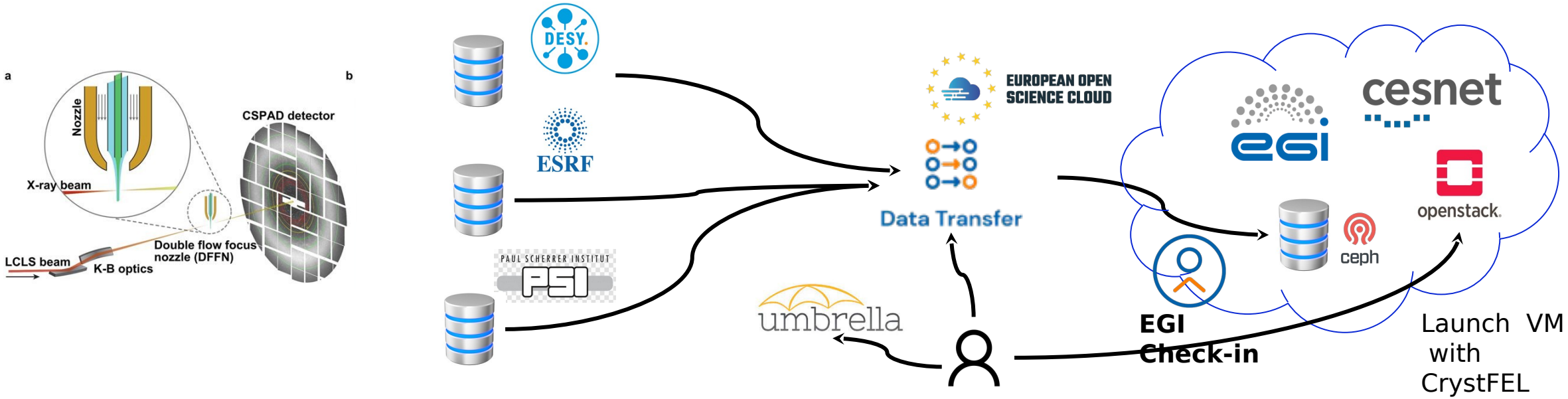
This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

OK, so why this is a problem?

- Lots of clicking just to get at the data.
- Assumes the user has a web-browser (with JavaScript).
- **Makes automation almost impossible.**

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Serial crystallography portal



- Project from **Gianluca Santoni** (ESRF) to build an analysis platform with EOSC Marketplace.
- Provide simplified platform (in EOSC) to lower the entry barrier for processing large amounts of data.
- Demonstrated at the EOSC-Future (M12) review.

My thanks go to Gianluca for providing the material on this slide.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

What do we want?

- **Try to avoid special cases.**
- Support **third-party transfers**:
  - Support third-party (storage-to-storage) data transfer (e.g., Globus, FTS, …).
  - Vendor neutral: not preferring one service over another.
- Support **direct downloads**:

  Easy for scientist to “just download” the data from the DOI.
- Follow **existing standards**, wherever possible:

  Work with open-source tools.
- **Minimal** requirements:

  Ideally, with commonly deployed software.
- A **low cost** to implement.

This project receives funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 857641

Process for building solution

- Document a **straw-man proposal**.
- Work to achieve **consensus with key stakeholders**:
  - Data catalogues (ICAT, SciCat, Zenodo, B2SHARE, …),
  - Transfer technologies (Globus, OneData, FTS, Rucio?, …),
  - Other interested parties (e.g., open-source development teams).
- Build a **priority list** (may not be able to do everything in one go).
- Reach out to **EOSC-Future** interoperability framework WG.
- Once consensus is reached:
  - Document it.
  - Work with data catalogues and transfer technologies to add support.
  - Engage in interoperability testing (compliance testing vs one-on-one testing vs connectathons / hackathons).

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

What I need from you…

- You agree this is something **worth pursuing**.
  - This is a genuine problem, something holding back research.
- You agree this approach is **reasonable**.
  - No obvious show-stopper problems.
  - Have the correct groups been identified?
- You agree to **help convince** your facility's data-catalogue team to engage with this process, to help make it a success.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Thanks for listening!

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Bonus material

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

So, is automation *really* impossible right now?

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

So, is automation *really* impossible right now?

- When querying the DOI, the client shouldn't know which system is handling this DOI

  DOIs are supposed to be opaque IDs.

- Therefore, the first operation should be to resolve the DOI.
- The following slides show an experimentally deduced procedure for obtaining data from three dataset catalogues.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PSI

1) Resolve DOI to PSI landing page URL.
2) Build jsonInfo URL:
   - DOI → **`https://doi.psi.ch/oaipmh/Publication/detail/`**`DOI`
   - Double-encode the DOI!
3) Query jsonInfo via HTTP GET.
4) Parse JSON and extract **`downloadLink`** item.
5) Fetch downloadLink page via HTTP GET; result is HTML.
6) Obtain downloadBaseURL by scraping HTML for the “wget …” command.
7) Build checksumInfo URL by resolving **`__checksum_filename_0__`** against downloadBaseURL.
8) Download checksumInfo.
9) Parse JSON. For each file resolve filenames against downloadBaseURL.

This project receives funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 857641

ESRF

1) Obtain a login session (HTTP POST)
   - Hard-coded: `https://icatplus.esrf.fr/session`
   - Use hard-coded credentials ("`reader`", "`reader`") for anonymous access.
   - Extract **`sessionID`** item from response JSON.
2) Build datasetInfo URL from `sessionID` and DOI:

   (`doi`, `sessionID`) → `https://icatplus.esrf.fr/doi/%s/datasets?sessionId=%s`
3) Query datasetInfo using HTTP GET
4) Parse JSON and extract **`datasetID`** item.
5) Build datafile URL from `datasetID` and `sessionID`:

   (`sessionID`, `datasetId`) → `https://icatplus.esrf.fr/catalogue/%s/dataset/id/%s/datafile`
6) Query datafile information.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Zenodo

1) Resolve DOI to Zenodo landing page URL:

   `doi:10.5281/zenodo.6618186` → `https://zenodo.org/record/6618186`

2) Extract **`recordID`** as last path item in landing page URL:

   `https://zenodo.org/record/6618186` → `6618186`

3) Build REST API record URL from **`recordID`**:

   `6618186` → `https://zenodo.org/api/records/6618186`

4) Query REST API record (HTTP GET)

5) Parse JSON, extracting information from **`files`** item.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641