Photon and Neutron Facilities in EOSC: challenges and opportunities

17th april, 2019 – EOSC-hub week

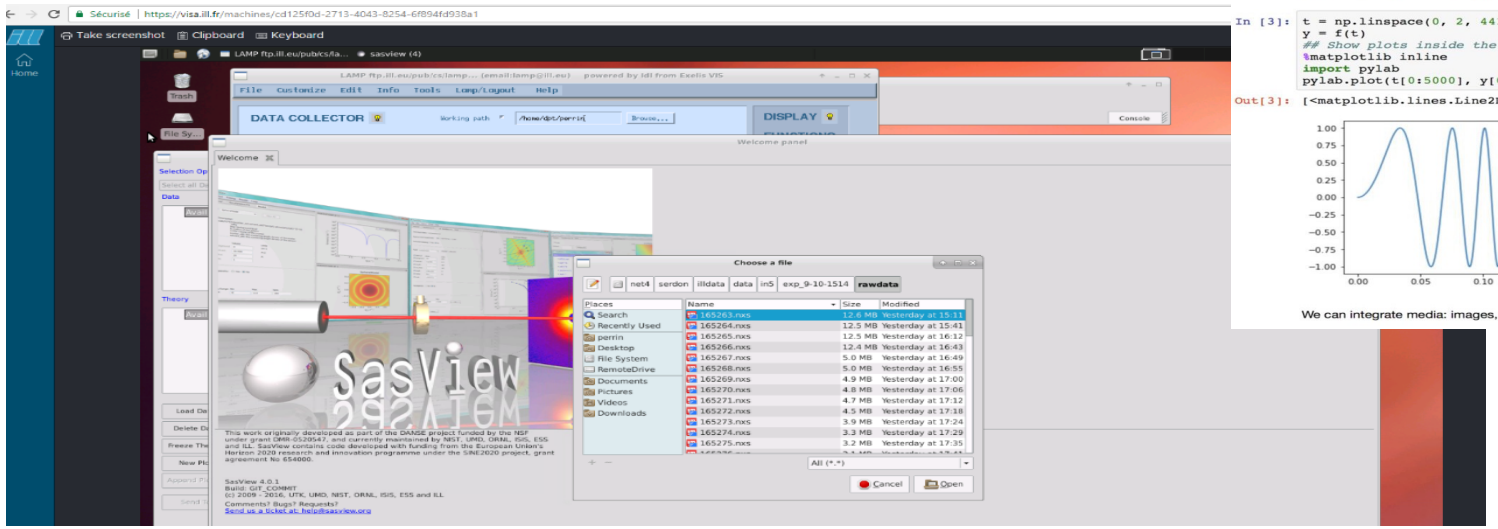
Author: J.-F. Perrin on behalf of the PaNOSC cluster

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

The proposal

- **50,000 users** – Biology, Medicine, Materials, Chemistry, Nuclear Physics, Particle Physics, Cultural heritage, Geology … and **industrial** applications.
- **State of the art Large Scale Facilities** – 5 ESFRI + 25 national RIs (**PaNs**)
- **Data policies** implementing FAIR principles – PaNdata data policy
- **10s of Petabytes** of scientific data, curated and archived for 5-10+ years
- **PaNs** manage and provide access to data from experiments across Europe
- **Working together** in past and ongoing projects – PaNData, SINE2020, CALIPSOPlus, EUCALL, LEAPS, …

What we want to provide



In [1]:
```
1 + 2
```
Out[1]: 3

Cells can contain text and latex equations such as $f(x) = \sin(2\pi\omega t^2)$ and $\omega = 220$ Hz.
We can use code to define the corresponding functions:

In [2]:
```
import numpy as np
def f(t):
    omega = 220
    return np.sin(2 * np.pi * omega * t**2)
```

Let's compute the data and plot the beginning of it:

In [3]:
```
t = np.linspace(0, 2, 44100)
y = f(t)
## Show plots inside the notebook
%matplotlib inline
import pylab
pylab.plot(t[0:5000], y[0:5000])
```
Out[3]: [<matplotlib.lines.Line2D at 0x1047ada58>]

We can integrate media: images, videos, interactive elements and sound:

- Curated Open Data and metadata of the highest quality
- Reliable services dedicated to understanding and to further exploiting these data
- Technical and scientific support on these data and data services
- Our experience on FAIR data policies and FAIR implementation guidelines for Photon and Neutron science
- Our knowledge and understanding of our scientific community
- Our ability to promote FAIR culture amongst our community

Data Policies

2008
- PaNData policy framework
- Adoption by several facilities
- Similar experience for our users

- PaNOSC: updated Data Policy framework:
  - Lessons learned from the community
  - Better understanding of FAIR
  - Align with other EOSC projects/clusters work on DP
- Common guidelines
  - Minting DOIs, long term archiving, downloading and citing data, ….
- Implementing DMP template

Metadata

2015

- Nexus data format https://doi.org/10.1107/S1600576714027575
- Nexus is the de facto standard
- Adopted by most synchrotrons
- Not perfect needs to be continuously improved

- PaNOSC: adopted Nexus as standard:
  - Lessons learned from the community
  - Work closely with Nexus Committee (NIAC)
  - Collect rich metadata on all experiments
  - FAIR principles require rich metadata
- Common guidelines
  - Define new metadata standards where missing, share metadata catalogues or definitions based on Nexus, ….
- Store data in HDF5 following Nexus conventions

Data Catalogues

https://icat.esrf.fr https://data.ill.eu …

Standard metadata available via OAI-PMH (DataCite service)

- Development of an API to allow federation and exposure of metadata relevant for the area.
- Provisioning Federated Search
  - Linked with EOSC-Hub data catalogue(s)
  - Searching all PaNOSC partner sites for available datasets

Data Analysis Services

Enabling transition from measurements to insight and new science

Providing Raw data is not enough … we also need to provide specific services for extracting scientific knowledge.

- Web remote desktop based analysis services
  - Provide generic solution for analysis software
- Web notebook based based analysis services
  - Focus on reproductible and publishable data analysis
- Integration into EOSC service portfolio
- Moving from single facility to EOSC scope

Simulation Services

Simulations of the various parts and processes involved in complex experiments play an increasingly important role in the entire lifecycle of scientific data generated at RIs.

- Expose source, beamlines optics and scattering simulations as cloud services
- Expose simulation data services in data analysis frameworks accessed via Jupyter notebooks or remote desktop solutions.
- Integration into EOSC service portfolio

Support & Training

Expanding from facilities' users specific support and training to PaN cluster and EOSC users.

- Integrated technical and scientific Helpdesk that will give support to data scientists
- E-learning platform
- Staff training in data stewardship
- Participate in scientific schools to promote FAIR principles and introduce the use of EOSC services

Working with EOSC-hub

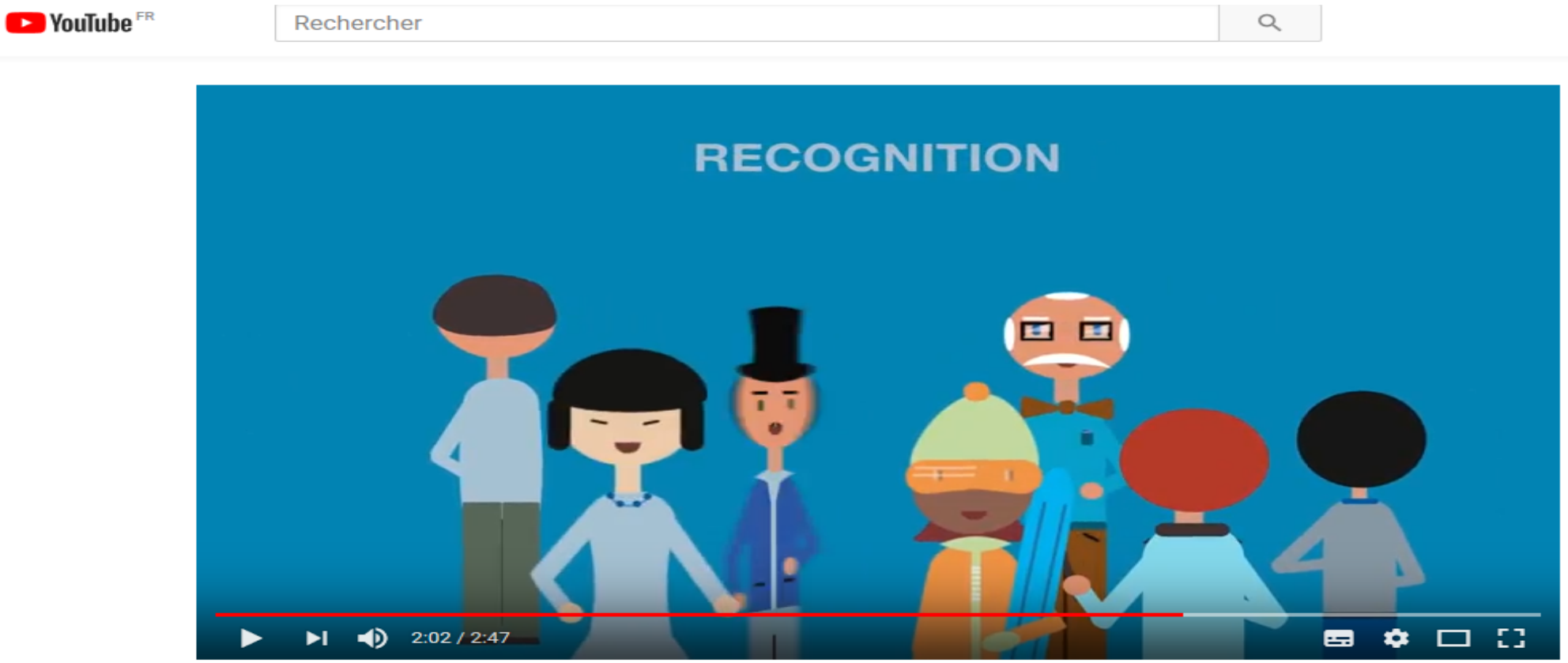
- Active participation in governance.
- Active Participation in open policies activities.
- Integration of our data catalogues into the EOSC data catalogue.
- Use of E-Infra IT services to deploy more specific services targeted at Photon and Neutron data type and users.
- Provisioning of models and solutions to bring small datasets to the compute resources and vice versa for very large datasets.
- Commonly defined service quality levels (Service Level Agreements) and if necessary upgrade the services to reach and maintain reliably this level of quality.
- Commonly defined usage metrics and the adoption of the necessary tools to collect and publish them.
- Harmonization of solutions for federated identity provisioning, authentication and authorization.
- Set up a technical and scientific support structure for handling data scientist (not necessarily facility users) requests
- Promoting FAIR data culture.

Key challenges with e-infra

- Integrating our services into EOSC Hub Service Catalogue
  - SLA Standards
  - Metrics and monitoring
- Providing compute capacity for the services
- Data availability for the services
  - Moving data to the services
  - Moving the computing capacity to the data
- PaN Software catalogue integration into EOSC database catalogue
- Models for AAI integration
- Data archiving pilot

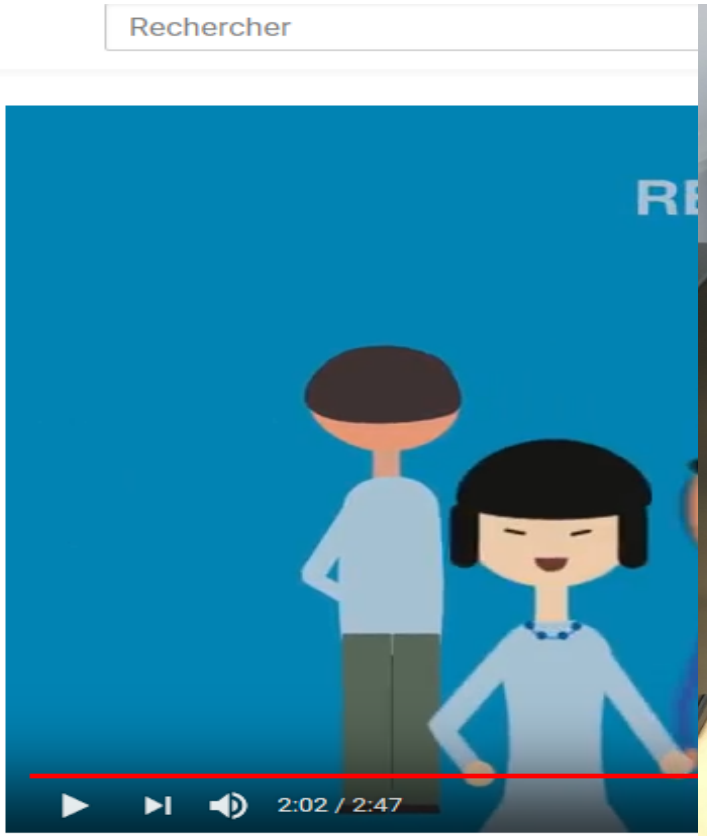
Open Data and FAIR principles stewardship …all together

https://www.ill.eu/DOIs

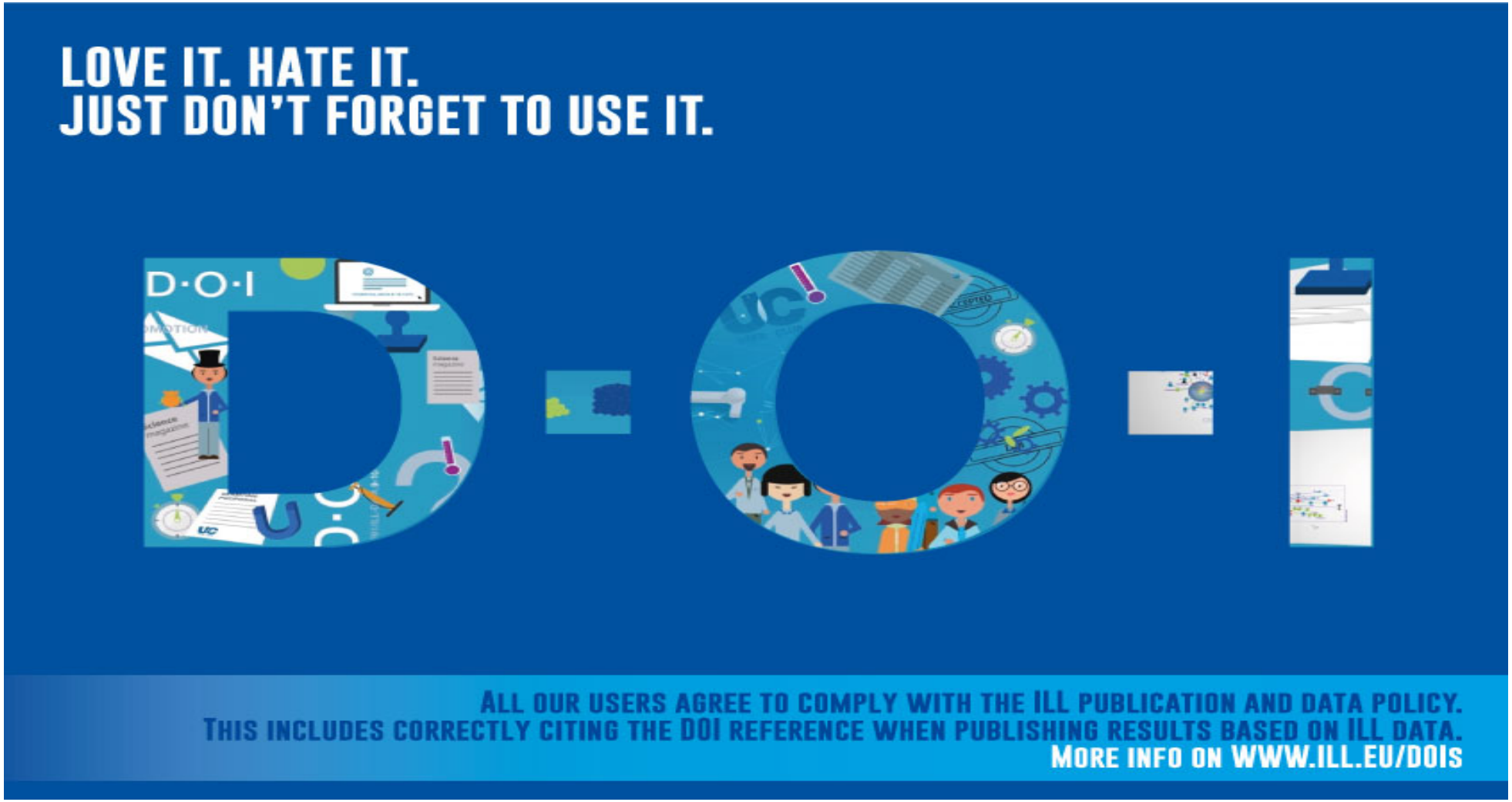


Open Data and FAIR principles stewardship …all together

https://www.ill.eu/DOIs

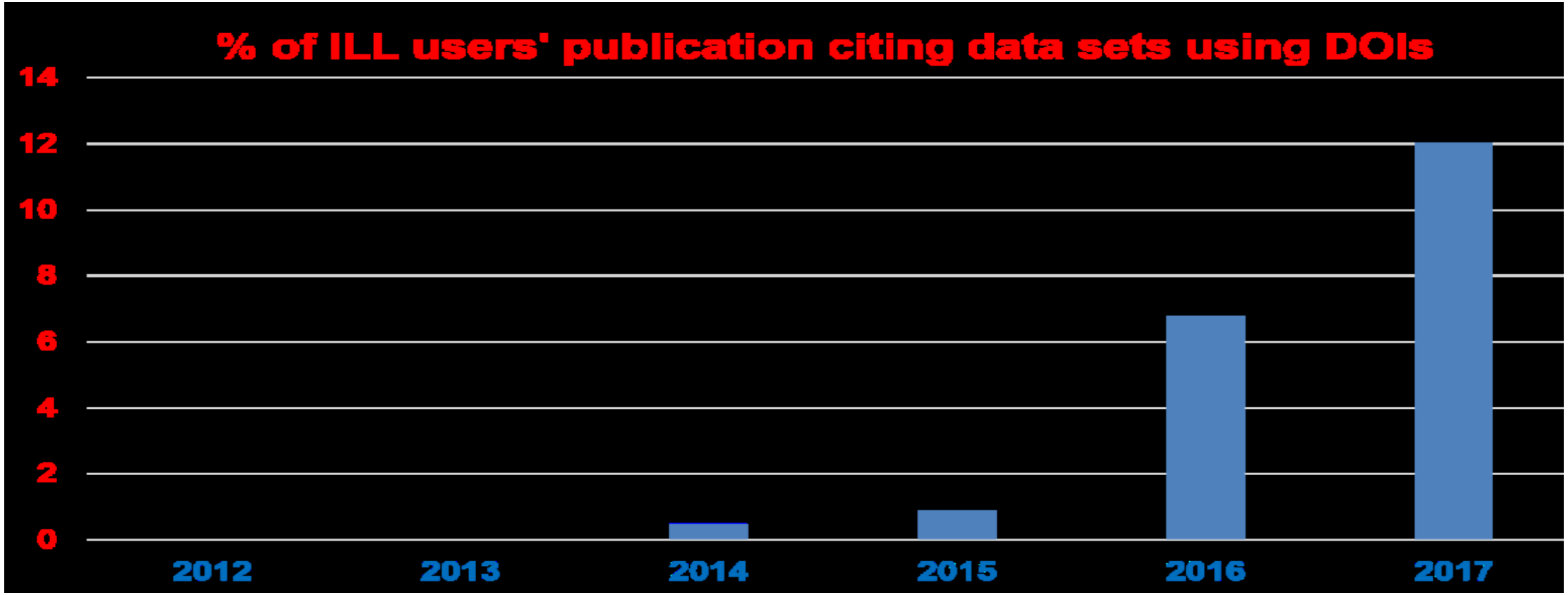


Open Data and FAIR principles stewardship …all together



Alone is not enough

We need complete and focused participation from all parties: Universities, Publishers, RIs, e-Infra, EC, …



We are Open to Collaboration

We need to **work together** with **National RIs** on **Open Data** and **Data Analysis as a Service !**

andy.gotz@esrf.fr

Don Quichotte and Sancho Panza, drawn by Honoré Daumier