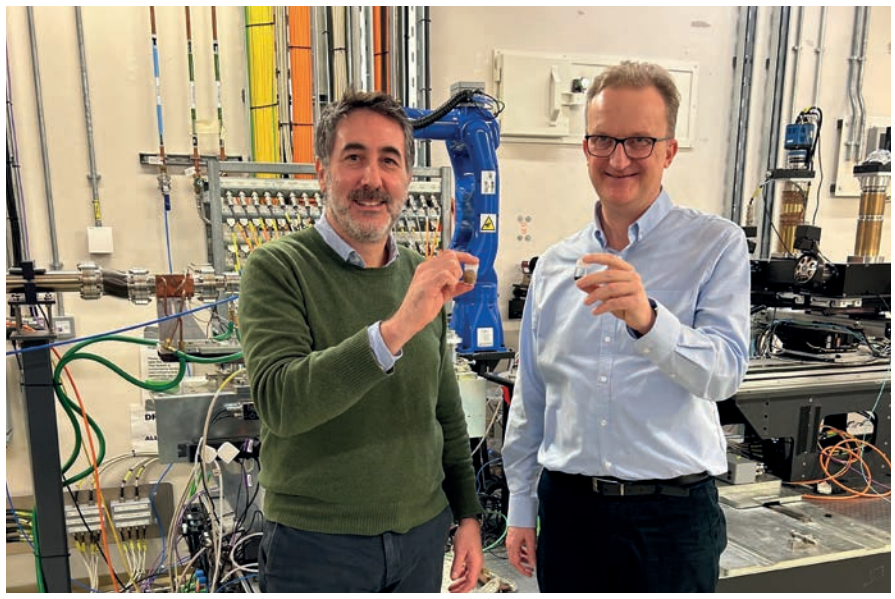
Getting Synchrotron and Neutron Big Data the FAIR Way

Kat Roarty, Diamond Light Source Impact Manager and Communications Coordinator for ExPaNDS

Synchrotrons are research institutes that welcome scientists from all over the world to answer their scientific questions in all scientific domains, from life science to cultural heritage. Their experiments increasingly produce incredible amounts of data every year as synchrotrons are always progressing and developing new techniques to tackle a diverse range of scientific challenges. Faster detectors to allow in-operando analysis, robots to handle more samples and increase the capacity of a beamline. These improvements result in a huge uplift in the amount of data produced; thus data analysis is becoming one of the biggest challenges for synchrotron and neutron facilities.



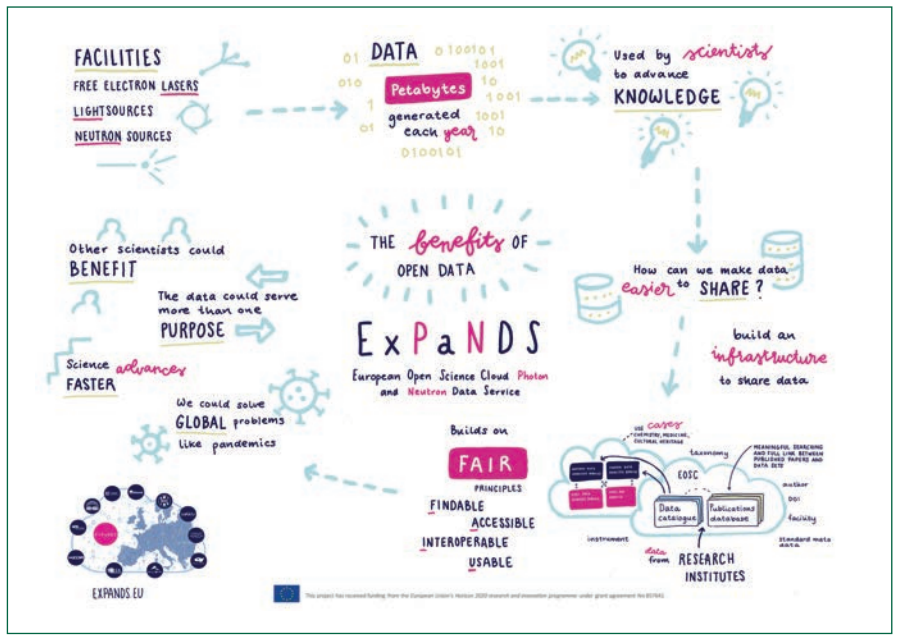
Dr Roberto Volpe and Dr Christoph Rau with Biochar samples inside I-13 beamline at Diamond

Petabytes of data are produced by Photon and Neutron (PaN) facilities every year, so the need for collaboration and a coordinated approach is an issue facing most scientists. A single tomography experiment produces several terabytes of data in a couple of hours and all of it needs to be thoroughly annotated to be analysed by the team of researchers who produced it. But what if data could be analysed again by different teams and thereby maximise their value? This is one of the main goals of the ExPaNDS (European Open Science Cloud (EOSC) Photon and Neutron Data Service) European project, a collaboration of 10 national Photon and Neutron Research Infrastructures (PaN RIs) from across Europe. ExPaNDS partners share a diverse user community of at least 25,000 researchers. Its users undertake experiments involving imaging capabilities and other innovative techniques as well as a huge diversity in data management techniques. PaN facilities are notorious for generating huge volumes of data and massive data files making harmonisation a challenge. The ambitious ExPaNDS project has been working with users to review how the value of data can be increased by more efficient management to support sharing and reuse, advocating for better management.

PaN facilities are moving forward, preparing machine upgrades that will allow new possibilities for science. For example, at Diamond Light Source, its planned upgrade to Diamond-II will not only increase brightness and coherence by a factor of 70, but also enhance beam quality and stability through new X-ray optics and instrumentation, state-of-the-art sample delivery, and manipulation through the development of optimised sample environments. As a consequence, the huge gains in throughput for many experiments will necessitate a transformation in Diamond's ability to gather, manage and analyse the vast data volumes and data rates that will be generated.

Approaching harmonisation of data

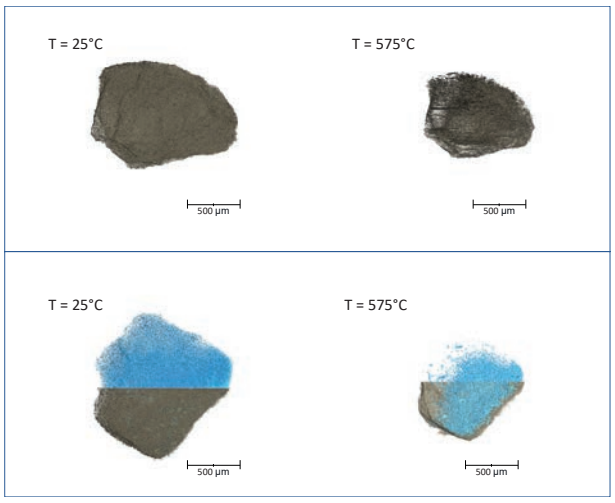
The ExPaNDS project, along with its sister project PaNOSC (panosc.eu), involving EU RIs, enabled the assessment of how all the PaN facilities managed their data and to understand the synchrotron users expectations regarding data. One of the agreed goals was the concept of FAIR data i.e. meeting the following principles: The data must be Findable, Accessible, Interoperable, and Reusable (FAIR).



ExPaNDS Infographics 2

The first step in (re)using data is to be able to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the 'FAIRification' process. Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation. The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing. The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Dr Roberto Volpe with samples and imaging data in I-13 hutch



Untreated biomass particles (on the left) change their dimension with temperature (on the right) and showing their evolution when undergoing pyrolysis.

In grey we see the biomass particles and in light blue the network of pores contained within.

a. Untreated biomass particles (on the left) change their dimension with temperature (on the right) and showing their evolution when undergoing pyrolysis

b. In grey we see the biomass particles and in light blue the network of pores contained within

ExPaNDS Provides Guidelines

PaN facilities operate in so many different techniques, it's a real challenge to harmonise the format of data and all the associated metadata that goes along with it. ExPaNDS helped with this part by developing guidelines and recommendations providing a toolkit that facilities can use to ensure that data generated from experiments is FAIR, and thus suitable for sharing and reuse, as well as easier for the experimenters themselves to use.

Firstly, the organisational context is considered in guidance on setting a Data Policy that supports data sharing, with a commitment to support the publication of FAIR data, while protecting the user's priority in conducting science. A second recommendation discusses establishing a FAIR experimental process, including coordinating the tools and information systems of the facility to support the collection of rich metadata about data, so researchers can discover and understand the data sufficiently to allow reuse. A third recommendation discusses Persistent Identifiers (PIDs), which uniquely label resources such as data, papers and even people, so that they can be unambiguously found and used, while allowing credit to be given to the experimental team. Data Management Plans (DMPs) are discussed in the fourth recommendation, bringing 'FAIR-ness' to a particular experiment, specifying additional metadata describing that instance. DMPs are time-consuming for users to produce, and the guidance considers how this burden can be significantly reduced by integrating DMPs into the experimental process.

A consultation was carried out, reaching over 14,000 researchers calling for their needs for better data management. The consultation, which received feedback from just under 200 respondents, revealed that:

• 82% of respondents declared making at least some of their data open and 71% declared making at least some of their data FAIR;

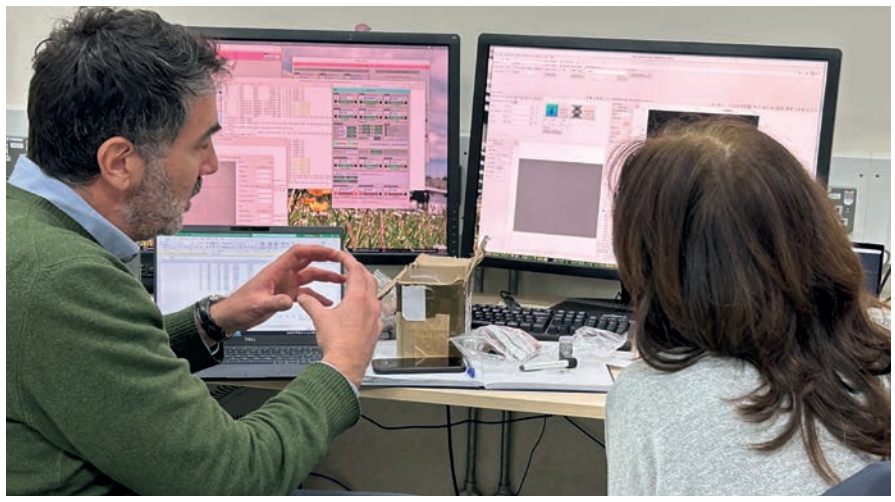
• Almost 70% of respondents declared that "Documenting the datasets (auxiliary and main) so that the results can be replicated and understood" was a challenge to make their Data FAIR & Open;

• Almost 50% of respondents declared that "Data are too big to share" was a challenge to make their Data FAIR & Open.

Exploitation of FAIR principles (especially Findability) can be hampered by the lack of consistency in metadata used to annotate data records and search databases. Another key part of the project was the development of several small ontologies to facilitate consistent semantics for terms within the PaN domain all gathered under our umbrella PaNET, the Photon and Neutron Experimental Techniques Ontology. This simple ontology provides a taxonomy of PaN techniques, with new techniques being defined as subclasses of multiple, more elementary, technique classes.

Diamond Light Source Experimental Hall

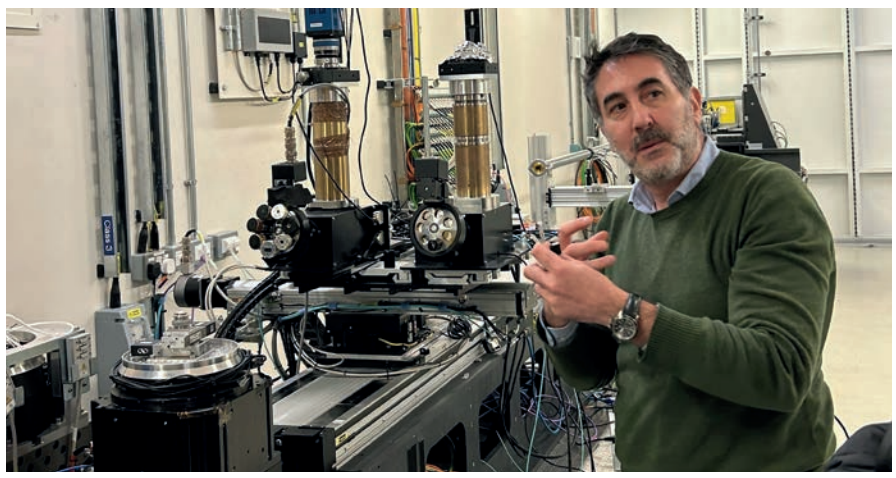
Dissemination leader for the ExPaNDS grant, Isabelle Boscaro-Clarke from Diamond comments; "The ongoing work of ExPaNDS has been very important to our PaN community. It has delved deep into what the users need, and has laid the foundations for data catalogues to connect to EOSC platforms, allowing them to be shared in a uniform way. The ExPaNDS grant has made progress towards more efficient ways to share and manage data, which will make it easier to find and share research, help prevent repetition of experiments and spur scientific progress. The project has also advocated to nations facilities the importance of delivering standardised, interoperable, and integrated data sources and data analysis services for Photon and Neutron facilities".



Dr Roberto Volpe explaining some of imaging results from I-13 beamtime

Tackling complex data mining sets

ExPaNDs was the opportunity to follow users from the beginning of the experiment to the analysis of the data obtained at various facilities. One of the case studies for Tomography/Imaging features Dr Roberto Volpe and his team at Queen Mary University of London and University College London. They have for the first time imaged the porosity of biochars via unprecedented operando experiments measured at Diamond. Dr Volpe's work to overcome existing knowledge gaps in the thermochemical decomposition of biomass could enable production of tailor-made bio-chars for high priority environmental applications. With the support of ExPaNDS - Diamond has worked with Dr Volpe on understanding the analysis barriers facing users like him.



Dr Roberto Volpe in I-13 beamline at Diamond

Dr Volpe said that the helping hand he received analysing his data from the ExPaNDS and Diamond team fastened his research. Commenting he said that data mining of these huge datasets is a new discipline and requires extensive collaboration. Sharing of such large and complex sets of information is challenging and the ExPaNDS grant helped identify better ways to deliver data management which is really useful to speed up results and transparency."

Isabelle Boscaro-Clarke, Head of Impact, Communications & Engagement at Diamond

Professor Dr. Helmut Dosch, Chairman of the DESY Board of Directors

Chairman of the Board of Directors for DESY, one of the largest PaN facilities and the leading partner in the ExPaNDs grant – Professor Dr Helmut Dosch, concluded: "We can now create solutions these days and in the future, even more so - atom by atom, you know materials which can be used for fighting climate change and diseases. But this data, this information is coming with a huge avalanche of data to us, and we need concepts how to turn this data in to useful information and to knowledge. It needs the right people; it needs the right infrastructure, and it needs financial resources. But I only can say now that knowledge is expensive, but ignorance we cannot afford."

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

Further information is available from Kat Roarty, Diamond Impact Manager and Communications Coordinator for ExPaNDS email: k.roarty@diamond.ac.uk

More information online: https://www.diamond.ac.uk

Picture credits: Diamond Light Source