Chapter 7
Towards the Interoperable Data Environment for Facilities Science

Vasily Bunakov
Science and Technology Facilities Council, UK

Catherine Jones
Science and Technology Facilities Council, UK

Brian Matthews
Science and Technology Facilities Council, UK

ABSTRACT

The research enabled by national and international photon and neutron source facilities makes a key contribution to the modern scientific community supporting thousands of researchers around the world to explore and understand the structure of materials. In this chapter, the authors describe the characteristics of facilities science with a distinct common research lifecycle which requires the provision of facility-based centralized IT infrastructure and data management platforms. The chapter then considers how the nature of facilities science is changing and what opportunities this brings for a more cohesive approach to data management. It then goes on to consider investigation research objects and to formalize the aggregation of related and contextual information which is important for re-use of research outputs by those who were not directly involved in the original experiment. Finally, there is consideration of what infrastructure can support the use of investigation research objects.

INTRODUCTION

Today’s scientific research is conducted not just by single experiments but rather by sequences of related experiments or projects linked by a common theme that lead to a greater understanding of the structure, properties and behaviour of the physical world. This is particularly true of research carried out on large-scale facilities such as neutron and photon sources, which are used for the detailed investigation of the structure of matter in areas such as pharmaceuticals, chemistry, mate-
rial science or biology. Such large-scale facilities support many individual researchers and small research teams who share access to common, rare resources supported by specialist infrastructure. Facilities science combines the characteristics of “big science” (dedicated infrastructure) with those of “bench science” (many individual experiments).

Experiments at these facilities require complex computing support to convert the information collected at the facility into the knowledge which is interpretable by the scientific community. Additionally, experiments are increasingly not carried out in isolation, but require the combination of sequences of experiments for the full understanding of the sample structure. A further feature of this area is the extent to which the research community is shared across facilities. As a consequence the support required by scientists to manage their data and access the compute resources required for a complete result is getting more complex and large-scale.

Facilities experiments often require sophisticated software for data analysis and visualization, as well as for the experiments’ design through simulation of experimental environment. The advances of computer science and information technology have made it possible to use supercomputers for the simulation of experiments themselves, and use simulation along with experimentation to get new insights into the nature of samples of materials under investigation. Computer simulations produce substantial amount of data that contributes to data management requirements.

Thus in the area of facilities science, there is a growing need for a comprehensive data infrastructure across facilities to enhance the productivity of research. In this chapter, we discuss the issues and approaches to providing a common open data infrastructure to support the users of facilities as they pass through and between different facilities. The issues which we consider include:

- The special nature of scientific collaboration and support within facilities science, and the requirements this nature poses for infrastructure support for the scientific community. This is built around a particular scientific workflow which provides the basis for a common data infrastructure.
- The changing landscape for facilities science as data rates from instruments increase, automation increases in the data infrastructure, and co-investigations undertaken across facilities mean that sharing and combination of data is required. These factors are affecting the established scientific practice. Further, common changes in the research policy that encourage open access to data for validation, sharing and reuse are also influencing facilities science, leading to a need for more effective data curation, publication and sharing.
- Curating and publishing information about experiments in context beyond the publication of data alone, using a notion of Investigation Research Object as a sharable persistent unit of discourse for facilities science. This also allows the publication and sharing of data within a Linked Open Data framework. This can point the way towards a framework for data curation and data sharing in facilities science which is required, in addition to technological considerations, to make the collaborative IT solutions realistically designed and sustainable through time.

We conclude with a future vision of an Open Data Infrastructure for facilities science across facilities interoperating with different user communities which use the large-scale facilities.

BACKGROUND

The class of large-scale facilities considered in this chapter supply intense beams of particles (such as photons or neutrons) generated from sources