Enabling Scientific Discovery with Microsoft SharePoint

Kenji Takeda, Mark Scott, Simon Coles, Les Carr, Jeremy Frey, Simon Cox, Steven Johnston
Microsoft Institute for High Performance Computing, University of Southampton, Southampton, UK

Abstract

Scientists and engineers facing increasing amounts of data must create, execute and navigate complex workflows, collaborate within and outside their organisations, and need to share their work with others. In this paper we demonstrate how the Microsoft SharePoint platform provides an integrated feature set that can be leveraged in order to significantly improve the productivity of scientists and engineers. We investigate how SharePoint 2010 can be used, and extended, to manage data and workflow in a seamless way, and enable users to publish their data with full access control. We describe, in detail, how we have used SharePoint 2010 as the IT infrastructure for large, multi-user facilities including the UK National Crystallography Service, µ-Vis CT scanning facility, and the Southampton Nano-Fabrication facility. We also demonstrate how SharePoint 2010 can be integrated into the everyday lives of scientists and engineers for managing and publishing their data in our Materials Data Centre[1], which provides an easy-to-use data management system from lab bench to journal publication via EPrints.

Introduction

The world of the scientist and engineers is becoming increasingly challenging as instruments, sensors and computing become more capable, resulting in complex workflows and a deluge of data to manage and comprehend. Increased focus on collaboration and opening up of data and information is also driving researchers to network in a social sense. Over the last decade the eScience research community has explored many issues and created many systems that tackle these issues.

Many of the challenges faced by companies are analogous to those faced by the scientific community, and platforms for managing data, workflows and collaboration have become more capable in recent years. Microsoft SharePoint is one such platform that is designed to manage content, communities, dynamic web sites, collaboration and information reporting[2]. In this paper we describe how we are using Microsoft SharePoint as an integrated platform for managing the scientific discovery process across a number of facilities.

Scientific Application of Microsoft SharePoint

The requirements for a computational platform to support science are wide-ranging, and discipline-specific in many cases. There are, however, classes of application areas that share significant commonality. One such area is large, multi-user experimental facilities. The management and use of these facilities goes beyond the scope that individual users usually have to deal with in terms of the end-to-end process: from applying to use a facility; through authorisation and training; to running experiments and processing results. Such workflows usually involve a number of stages that are both manual and automated.

Here we consider three large multi-user facilities at the University of Southampton. The UK National Crystallography Service[3], funded by EPSRC, comprises a number of diffractometers that are used for crystal structure determination. It also provides access to the national Diamond Light Source Synchrotron when more intense X-rays are required by users. The µ-Vis X-ray Computed Tomography (CT) Centre at Southampton is a multi-disciplinary facility used for imaging samples from a wide-range of users, from engineering, biomedical and environmental science. The system allows us to look inside objects and create

[1] www.materialsdatacentre.com
[2] http://sharepoint.microsoft.com/
[3] http://www.ncs.chem.soton.ac.uk/

3D reconstructions for detailed analysis (for example, Figure 1). The high-resolution imaging requires processing large amounts of data, O(TBytes), using advanced image processing on GPUs. The Southampton Nano-Fabrication Centre[4] is a purpose-built facility that has a variety of optical and electron-beam systems for lithography down to 5nm. Characterisation is carried out on a range of instruments, including a focussed ion beam with integrated SEM.

While these facilities have different scientific outputs, their operation shares many common features, particularly in terms of the workflow that users must follow. We consider the full end-to-end process here, as illustrated in for μ-Vis, from both the user and facility perspectives. We term this as the *business workflow*, to distinguish it from *scientific workflow*, which has been the focus of previous work in the area such as Taverna and Kepler workflow systems. In our description the scientific workflow is just one of the activities within the business workflow, although this can be extremely complex in itself.

Figure 2: User experience and workflow for μ–Vis CT scanning centre

Figure 3: EP2DC architecture, linking EPrints to the Materials Data Centre

A number of key features within SharePoint 2010 are directly relevant for application in scientific and engineering domains. As a platform it integrates workflow, data management and connectivity services within a social computing framework, making it fully personalisable based on user roles.

In order to execute the business workflow, the SharePoint Workflow Engine provides an extensible framework in which developers can create steps and workflow templates. These can then be assembled using the SharePoint Designer package by domain specialists without them having to write code. These can then be published to the user community after being tested. This pipeline for workflow creation and publication divides the tasks to the most appropriate people and provides a reliable management framework; for multi-user facilities this level of rigour is required. An example of such a workflow is for managing health & safety training and clearance before users can access a facility; this is similar in form to the default business expenses claim workflow. Execution of the scientific workflow can be carried out within SharePoint, but can be offloaded to external systems if this is appropriate; for example where bespoke software is required for data processing from a specific machine.

Data management can be handled through SharePoint via its Lists, and Business Connectivity Services (BCS). SharePoint uses Microsoft SQL Server as its dataset engine, with BCS providing full create-read-update-delete access to third-party databases. The inclusion of an XML engine provides additional capability for validating datasets. We have leveraged this functionality in our Materials Data Centre (MDC) implementation, in which we are using SharePoint as a data repository for the materials science community. We have used the Mat-DB schema[5], which is loaded into a schema document list and used to validate datasets uploaded using our EP2DC federated repository framework. EP2DC provides a plugin for EPrints that allows users to submit datasets at the same time as their papers, and then deposits the

[4] http://www.southampton-nanofab.com/

[5] Ojala, J., and Over, H.H., Approaches in using MatML as a common language for materials data exchange, Data Science Journal, Volume 7, 4 November 2008

datasets in a discipline data repository built on Microsoft SharePoint (Figure 3). The MDC provides users with a managed data store, that allows metadata tagging, version control, search and access control for publishing data to colleagues and the wider world when appropriate.

SharePoint is an integrated portal solution that presents a familiar web interface that most users are comfortable with. The extensive customisability and extensibility of the platform provides flexibility for use across different domains. In our experience, there is significant commonality between disciplines for their core business workflows, with the development environment allowing customisation to meet the specific needs of groups of users.

Conclusions

In this paper we have described how Microsoft SharePoint can be applied to the management of scientific facilities to increase efficiency and usability for both end-users and facility owners. A key capability is the ability to create, publish and execute workflows to manage the end-to-end business process of accessing major facilities. By integrating this with data management, it provides a single platform that can increase productivity by providing a single access and execution portal for scientists. An additional benefit is where Microsoft SharePoint is used in the research organisation for other functions. In this case, deploying it for scientific use means that another platform, with requisite infrastructure and support, need not be separately deployed. This not only provides potential costs savings, but also allows provides users with a more seamless experience for their day-to-day lives.

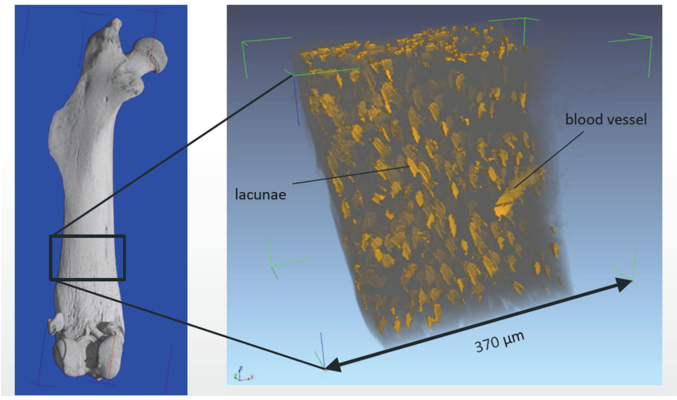

Figure 1. 3D rendering of CT scan of a whole femur of a OPN knockout mouse (left) and inspection of micro porosity of cortical bone (right) with bone being semi-transparent and lacunae and blood vessels coloured orange

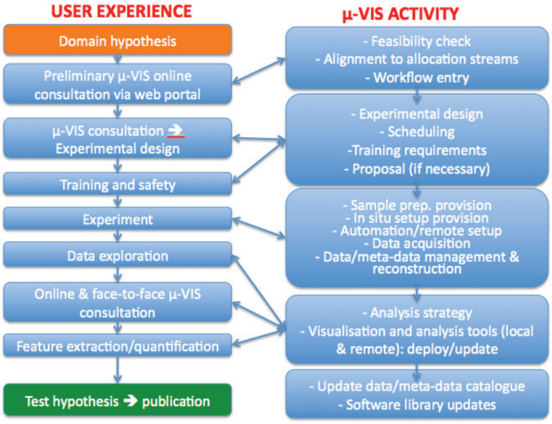
USER EXPERIENCE
Domain hypothesis
Preliminary µ-VIS online consultation via web portal
µ-VIS consultation ➔ Experimental design
Training and safety
Experiment
Data exploration
Online & face-to-face µ-VIS consultation
Feature extraction/quantification
Test hypothesis ➔ publication
µ-VIS ACTIVITY
- Feasibility check
- Alignment to allocation streams
- Workflow entry
- Experimental design
- Scheduling
-Training requirements
- Proposal (if necessary)
- Sample prep. provision
- In situ setup provision
- Automation/remote setup
- Data acquisition
- Data/meta-data management & reconstruction
- Analysis strategy
- Visualisation and analysis tools (local & remote): deploy/update
- Update data/meta-data catalogue
- Software library updates

Figure 2: User experience and workflow for µ–Vis CT scanning centre

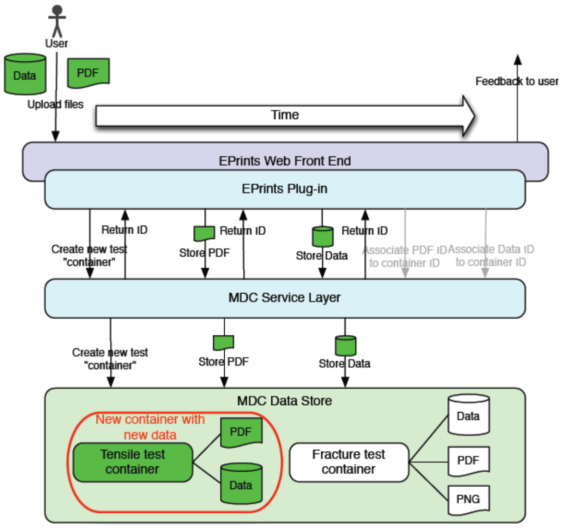

Figure 3: EP2DC architecture, linking EPrints to the Materials Data Centre