Enhancing YouShare: the online collaboration research environment for sharing data and services.

Victoria Hodge, Aaron Turner, Martyn Fletcher, Mark Jessop, Michael Weeks, Tom Jackson, Jim Austin,
Advanced Computer Architectures Group, Dept. of Computer Science, University of York,
York, YO10 5GH UK.
Email: {victoria.hodge, aaron.turner, martyn.fletcher, mark.jessop,
michael.weeks, tom.jackson, jim.austin}@york.ac.uk

Abstract.

This paper describes recent enhancements to the YouShare platform, the online collaboration environment, which allows researchers to share data and software applications and perform compute-intensive analysis tasks quickly and securely. The enhancements to the platform are a result of user feedback on the current system and technology advancements. These fall into four groups – better handling of searching, use of synonyms, the addition of a workflow tool and enhancements to the infrastructure. The paper outlines these improvements.

Key / Index word or phrases.
Search, synonym search, workflow.

1 Introduction.

YouShare [1] is a collaboration platform that allows research output, in the form of data, software applications or workflow processes, to be captured and made available for reuse. However, it is not simply a repository service, it offers a live repository of tools and services, derived from research output, which can contribute to more effective use of research software outputs and the opportunity for collaborative working on research data output. The YouShare platform grew out of the CARMEN project which focussed on neuroscience. YouShare is being developed as an independent and generic platform for all academia and also underpins three projects: the current CARMEN project [2], Brain Injury Index (BII) [3] and Condition Monitoring in the Cloud (CMAC) [4], and each of these have their own instantiation of the platform.

The following sections describe recent enhancements to the YouShare system over and above those reported previously [1].

2 Recent enhancements.

2.1 Search systems enhancements

A search capability is vital in a system such as YouShare, particularly as the number of data and services artefacts grows over time making any complete view of the data and services unfeasible. To allow searching against a high number of artefacts, YouShare requires a high performance, multi-faceted search engine. Baeza-Yates and Ribeiro-Neto [5] found Apache Lucene [6] to be the fastest, most compressed and most competitive of the search systems that they evaluated. Lucene is feature-rich, mature and robust with a large development community, wide industry

adoption and is suitable for most full-text search applications. It provides a Java API which integrates seamlessly with the YouShare Java-based architecture. Hence, we chose Lucene to underpin the search in YouShare. In the YouShare search engine, the stored documents comprise fields with name and value pairs. The YouShare metadata is stored as XML and the search engine parses the XML to extract the field and value pairs where the field is the XML tag name and the value is the XML tag text. The pairs are stored in a Lucene index structure in the file system with one index document (set of field name and value pairs) per metadata XML document. We have enabled the full Lucene query syntax: Boolean operators; term modifiers: wildcard searching, fuzzy searching and proximity searching; term boosting; term/phrase grouping; and, fielded searching. YouShare has augmented Lucene with both "Did you mean?" and auto-complete functionality. "Did you mean?" spell checks each of the user's query terms against the stored Lucene index and suggests alternative spellings for any misspelt query terms. Autocomplete analyses the query string as the user types, detects when a specific number of new characters have been typed (currently set to four but configurable) and suggests a list of matching words and phrases. The user may select a match or continue typing and the list of suggestions will be constantly refined as new text is typed.
Lucene now provides the search capability for the YouShare system and its variants: CARMEN, BII and CMAC.

2.2 Synonyms

A requirement that came from CARMEN neuroscience users was the need to suggest synonyms for users' search terms. These can then be used to enhance the users' search query on the system. A synonym facility has been added to the CARMEN system, which retrieves synonyms from the Neuroscience Information Framework (NIF: http://www.neuinfo.org/), a dynamic inventory of Web-based neuroscience resources: data, materials, and tools accessible via any computer connected to the Internet. The user's query term is submitted to the NIF portal which returns a list of synonyms in an XML document, the XML is parsed and the synonyms are extracted and displayed to augment the user's query. Figure 1 shows the CARMEN synonym user interface which has been integrated with the search panel on the left and the results of a synonym query for the word "mouse" on the right.

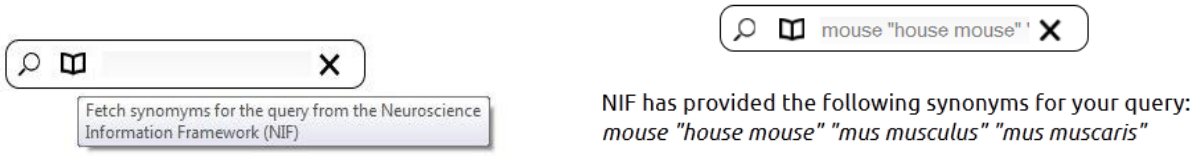

Figure 1 CARMEN synonyms interface (left) and the synonyms for 'mouse' (right).

The use of other synonym resources for other domains on YouShare can easily be substituted for other variants of the system.

2.3 Workflow enhancements

A tool to combine services together in an orchestrated processing pipeline, or workflow will provide: a systematic and automated facility for executing analyses across different dataset; an easy and reproducible way of capturing the processing pipeline so that results can be reproduced and the pipeline can be replayed, shared and adapted; and, a visual interface to generate these pipelines without low-level programming knowledge. YouShare is developing such a workflow

tool based on the CARMEN project's cloud execution model and data format. CARMEN has developed a standard data format that allows heterogeneous data to be specified and encapsulated in an XML wrapper [7].

The workflow tool itself consists of a graphical workflow design environment running in the portal inside the browser window, and a back-end workflow execution engine with access to a library of services and common workflow tasks. The graphical design tool (figure 2) allows users to create, share and execute workflows. Services and data can be located within the system via the search facility. The workflow execution engine coordinates execution of the services within the workflow, and manages the data between connected services. During workflow execution the system's main service execution scheduler is used allowing the workflow engine to make use of YouShare's dynamic service deployment and execution system, achieving scalable heterogeneous distributed processing. Parallel branches within the workflow are executed concurrently ensuring that services execute alongside input data to reduce data transfer.

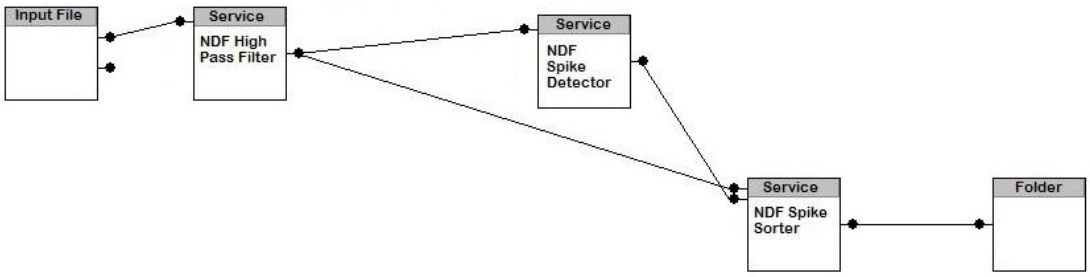

Figure 2 Workflow graphical design tool

2.4 Infrastructure enhancements

The current hardware platform is used to provide a core system and also for local investigation of options. It comprises a series of blade compute elements (24 core, 96GB RAM) and 60TB of storage. It employs extensive virtualisation of these resources to provide the agility to move beyond the boundaries of this hardware. A number of cloud computing platforms have been considered, but with the rapidly-changing nature of cloud offerings it is felt that the correct solution should be vendor-agnostic so as to allow maximum flexibility in terms of choice of resource. So an Open Cloud Computing Interface [8] model should be used to control provisioning of compute elements on resources required to run YouShare workloads. Internally, OpenStack [9] is being employed to provide an internal cloud. The advantages of OpenStack are its ease-of-use, industry support, academic support, and licensing model in an academic context.

The original data model used Storage Resource Broker [10] to achieve geographical data distribution, but this presented a lock-in to a particular software set, and also an overhead where this was not needed. In the interim the file storage paradigm has been replaced by a POSIX-file system [11] based view, which allows distribution to be handled by any data management infrastructure which can provide a POSIX-files system view. Whilst this prevents lock-in to any single vendor, it does restrict the system to using a single system across all nodes on which the data may be distributed. In future it is anticipated that bulk storage will be underpinned by the Cloud Data Management Interface [12] standard.

We are currently evaluating the Open Archival Information System [13] - compliant archival model for full lifecycle management of data items based on Content Data Objects [14] generated

from research outputs, and to link this research information systems.

The project is tracking developments in security models to assess the appropriate model which will maximise ease-of-use and can use well-supported services for longevity, for example the core National Grid Service [15].

3 Conclusions and Future Work.

With recent enhancements users can now use an enhanced search facility and will soon be able to use workflows via a new workflow editor and neuroscience users can search the search the Neuroscience Information Framework for neuroscience synonyms. Other enhancements which have been documented elsewhere include visualisation of times series data [4].

4 Acknowledgements.

The funding for YouShare is provided by the HEFCE University Modernisation Fund. CARMEN was developed under funding provided by the UK EPSRC on contract number EP/E002331/1 and is now supported by the UK BBSRC under contract BB/I000984/1. The CMAC project is funded by the UK Technology Strategy Board.

5 References.

[1] Austin, J., Fletcher, M., Jackson, T., Jessop, M., Turner, A. & Weeks, M. 2011. YouShare, an online collaboration research environment for sharing data and services. UK e-Science All Hands Meeting 26th-29th September 2011, York, United Kingdom.

[2] Watson, P., et al. 2007. The CARMEN Neuroscience Server. UK e-Science All Hands Meeting, September 2007, Nottingham, United Kingdom.

[3] Real-time detection of the onset of secondary brain injury in the intensive care unit http://www.wellcome.ac.uk/Funding/Technology-transfer/Funded-projects/Health-Innovation-Challenge-Fund/index.htm - retrieved 1st July 2012.

[4] Hickinbotham, S. et al. 2012 Interactive graphics on large datasets drives remote condition monitoring on a cloud. J. Phys.: Conf. Ser. 364 012056.

[5] Baeza-Yates, R. and Ribeiro-Neto, B. 2011. *Modern Information Retrieval: The Concepts and Technology Behind Search,* Pearson Higher Education, ISBN: 9780321416919.

[6] McCandless, M., Hatcher, E. and Gospodnetic O. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA.

[7] Weeks, M. et al. 2012. The CARMEN Software as a Service Infrastructure. To be published in the Royal Society Philosophical Transactions A.

[8] Open Cloud Computing Interface http://occi-wg.org/ - retrieved 3rd August 2012.

[9] OpenStack Cloud Software http://openstack.org/ - retrieved 3rd August 2012.

[10] Storage Resource Broker http://www.sdsc.edu/srb/index.php/Main_Page - retrieved 3rd August 2012.

[11] POSIX http://pubs.opengroup.org/onlinepubs/9699919799/ - retrieved 3rd August 2012.

[12] Cloud Data Management Interface http://www.snia.org/cdmi - retrieved 3rd August 2012.

[13] Open Archival Information System http://www.oclc.org/research/publications/archive/2000/lavoie/ - retrieved 3rd August 2012.

[14] Content Data Object (CDO) https://www.archivematica.org/wiki/Content_Data_Object - retrieved 3rd August 2012.

[15] National Grid Service http://www.ngs.ac.uk/ - retrieved 3rd August 2012.