Future Generation Computer Systems 29 (2013) 549–559

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Towards a virtual research environment for language and literature researchers

Muhammad S. Sarwar [a,*], T. Doherty [b], J. Watt [b], Richard O. Sinnott [c]

[a] Room 246C, Kelvin Building, National e-Science Centre, University of Glasgow, Glasgow, UK

[b] National e-Science Centre, University of Glasgow, Glasgow, UK

[c] Department of Computing and Information Systems, University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 16 March 2011
Received in revised form
1 March 2012
Accepted 22 March 2012
Available online 1 April 2012

Keywords:
Humanities
Language and literature
MapReduce
HPC
Grid
ENROLLER

ABSTRACT

Language and literature researchers often use a variety of data resources in order to conduct their day-to-day research. Such resources include dictionaries, thesauri, corpora, images, audio and video collections. These resources are typically distributed, and comprise non-interoperable repositories of data that are often licence protected. In this context, researchers typically conduct their research through direct access to individual web-based resources. This form of research is non-scalable, time consuming and often frustrating to the researchers. The JISC funded project Enhancing Repositories for Language and Literature Researchers (ENROLLER, http://www.gla.ac.uk/enroller/) aims to address this by provision of an interactive, research infrastructure providing seamless access to a range of major language and literature repositories. This paper describes this infrastructure and the services that have been developed to overcome the issues in access and use of digital resources in humanities. In particular, we describe how high performance computing facilities including the UK e-Science National Grid Service (NGS, http://www.ngs.ac.uk) have been exploited to support advanced, bulk search capabilities, implemented using Google's MapReduce algorithm. We also describe our experiences in the use of the resource brokering Workload Management System (WMS) and the Virtual Organization Membership Service (VOMS) solutions in this space. Finally we outline the experiences from the arts and humanities community on the usage of this infrastructure.

1. Introduction

Consider a scenario where a humanist wishes to search for a word, say '*canny*', in the dictionary to find its meaning; in a thesaurus to look up associated concepts and categories it is found and used in, and in a corpus of work to find the documents containing it. Researchers may also want to see the concordances (context where the term was used) and determine the frequency of occurrence of the word in each found document as a basis for further analysis. The ability to save the different result sets and analysis of those results for later comparison between many different resultant data sets and with different researchers is compelling to the humanities community (and indeed is a challenge faced by many other research domains). This scenario becomes especially interesting and challenging when multiple dictionaries, thesauri and text corpora need to be cross-searched or differences between the textual resources exist. For example, searching for the word 'canny' in the Oxford English Dictionary (OED) [1], Scottish National Dictionary (SND) [2] and Dictionary of Older Scottish Tongue (DOST) [3] at the same time will have slightly different results on the definition of the term. When compared with other resources such as the Historical Thesaurus of English (HTE) [4] to look up the related concepts and categories and/or the Scottish Corpus of Text and Speech (SCOTS) [5] and/or the Newcastle Electronic Corpus of Tyneside English (NECTE) [6] the multitude of definitions and their historical context is especially challenging to establish. The problem scales further if the researcher decides to search for multiple, possibly hundreds, of words at once and do all of the mentioned tasks. Currently, language and literature scholars use multiple independent web-based resources to achieve these tasks. Licencing access to multiple resources is commonly required and the end user researchers are left with traditional Internet hopping based research. An interactive research infrastructure that brings together all of the different provider's data sets together in a seamless and secure environment is thus highly desirable and has been the focus of the JISC funded ENROLLER project [7].

The ENROLLER project began in April 2009 and is tasked with providing such a capability through the establishment and support of a targeted Virtual Research Environment (VRE) implementing secure and seamless data integration and information retrieval system for language and literature scholars.

* Corresponding author. Tel.: +44 141 330 2958.
E-mail addresses: Muhammad.Sarwar@glasgow.ac.uk, muhammad.sarwar@unimelb.edu.au, sulmansarwar@gmail.com (M.S. Sarwar), rsinnott@unimelb.edu.au (R.O. Sinnott).

0167-739X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2012.03.015

1.1. Related work

VRE-SDM [8], TextGrid [9], TEXTvre [10] and gMan [11] are some of VRE systems that exist in the arts and humanities domain. VRE-SDM focused on the development of services for sharing and annotating manuscripts [8]. TextGrid involves providing tools and services for analysis of textual data and support for data curation over the Grid [9]. TEXTvre builds upon the success of TextGrid and provides tools for TEI-based resource creation [10]. gMan VRE is targeted at Classics and Ancient History researchers and aims to be a general purpose VRE for arts and humanities researchers [11]. While all of these mentioned projects are aimed at arts and humanities researchers, they are not particularly targeted at language and literature researchers. Furthermore their focus is not on supporting federated data access models where data providers are autonomous, e.g. as is the case with the Oxford English dictionary, but rather on the amalgamation of tools used for data processing and analysis associated with humanities research and/or the establishment of data warehouses where data sets are imported from archives as is the case with gMan.

ENROLLER aims to build a sustainable e-infrastructure for language and literature researchers. Through the ENROLLER work, researchers in the language and literature domain will have access to large amounts of language and literature data from a single, easy-to-use portal; membership of an international network of scholars; increased knowledge of digital resources, and direct access to a portfolio of analysis tools. ENROLLER will also raise awareness and understanding of e-Science and establish a focal point for research for the humanities community. It will allow a community of researchers with related aims to collaborate more easily, and already funded data sets to be used in new combinations that could result in heuristic discoveries. The wider humanities community will benefit directly from the models developed here. The resulting knowledge transfer will be of benefit to both the humanities and the e-Science communities as well as to the wider community such as publishers, dictionary creators and national services.

This rest of this paper describes the challenges in implementing the VRE for language and literature researchers and the solutions put together thus far. Section 2 describes the background and data collections involved in the project. Section 3 describes the VRE and its overarching requirements. Section 4 describes the design of the various components that make up the ENROLLER VRE. Section 5 explains the implementation details and outlines the problems faced and solutions implemented during the course of the work. Section 6 presents typical use cases in accessing and using the system. Section 7 highlights the feedback of the work collected from the language and literature community. Finally Section 8 draws conclusions on the work as a whole and areas of future work.

2. Data sets and formats

The ENROLLER project is currently working with numerous major data sets from a variety of data providers. These include:

2.1. The Historical Thesaurus of English (HTE, http://libra.englang.arts.gla.ac.uk/historicalthesaurus/aboutproject.html)

The HTE contains more than 750,000 words from Old English (c700 A.D.) to the present. HTE has been published by the Oxford University Press since 2009 and offers a new and significant development in the historical language studies. HTE data is currently available in XML format.

2.2. Scottish Corpus of Text and Speech (SCOTS—www.scottishcorpus.ac.uk)

The Engineering and Physical Sciences Research Council (EPSRC, www.epsrc.ac.uk) and the Arts and Humanities Research Council (AHRC, www.ahrc.ac.uk) funded SCOTS resource offers a collection of text and audio files covering a period from 1945 to present. The SCOTS corpus is currently available in a Text Encoding Initiative (TEI, www.tei-c.org)—compliant XML format. Data can also be made available through a PostgreSql relational database. SCOTS corpus contains over 4 million words of running text.

2.3. Dictionary of Scots Language (DSL—www.dsl.ac.uk/dsl)

The AHRC funded DSL resource encompasses two major Scottish language dictionaries The Scottish National Dictionary (SND) and The Dictionary of Older Scottish Tongue (DOST). DSL data is currently available in XML format. Scottish Language Dictionaries (SLD) hosts the data on their servers in Edinburgh.

2.4. Newcastle Electronic Corpus of Tyneside English (NECTE—www.ncl.ac.uk/necte)

The AHRC funded NECTE is a corpus of dialect speech from Tyneside in Northeast England. This corpus aggregates the work of two existing corpora: the Tyneside Linguistic Survey (TLS) created in late 1960s and the Phonological Variation and Change in Contemporary Spoken English (PVC) created in 1994. The NECTE corpus is encoded in TEI-compliant XML format. The encoded data is available in four different formats: audio, orthogonal text, phonetic and parts of speech tagged. NECTE corpus contains over 500 k (five hundred thousands) words of running text.

2.5. Corpus of Modern Scottish Writing (CMSW—www.scottishcorpus.ac.uk/cmsw/)

The EPSRC and AHRC funded CMSW is a collection of letters (mostly texts and images) from the period 1700 A.D to 1945 A.D. (This is regarded as 'modern' writing to the language and literature community.)

2.6. Oxford English Dictionary (OED—www.oed.com)

The Oxford English Dictionary (OED—www.oed.com) is a commercial resource published by Oxford University Press and is widely regarded as the primary authority on the current and historic version of the English language vocabulary.

2.7. The Hansard Collection

The Hansard Collection is a collection of transcribed texts of UK's parliamentary speeches from the period 1803 to 2005. The Hansard data is available in XML format. Hansard Collection contains over 7.5 million XML documents.

All of these data resources collectively represent significant independent investments and efforts in capturing and cataloguing the history of the English and Scots languages. It is to be noted that ENROLLER project does not maintain any of the data sets provided by the project collaborators. The Oxford University Press (OUP) maintains OED and DSL is maintained by SLD. Access to OED is made though an OED SRU service (http://www.oed.com/public/sruservice) while DSL is accessed using a secure web service.