VRERI JISC Project Plan – MILARQ (draft)

Project Name: **MILARQ: Improving Search and Discovery Performance over Combined Data**

Project Tag: MILARQ

Project Website: http://code.google.com/p/milarq/

Author(s): Graham Klyne <graham.klyne@zoo.ox.ac.uk>

Date this reports was started: 15-03-2010

Date this report was finished: <!--Please use dd-mm-yyyy to format completion date-->

Total amount of JISC money allocated to this project: **£35112**

Link to Project details on JISC PIMS: https://pims.jisc.ac.uk:443/projects/view/1592

Table of Contents:

Aims and Objectives

This MILARQ Project proposal seeks JISC funds to enhance performance of the CLAROS Explorer VRE, to allow wider public access to a rich set of classical art resources from major European research centres, and incorporation of additional source data.

CLAROS (Classical Art Research Online Services; http://www.clarosnet.org/) provides search and discovery services over datasets that have been integrated, using state-of-the-art Semantic Web technologies, from leading independent European research centres. The data web created from these datasets includes descriptions of classical art objects (pots, gems, statues, etc.) located in museums worldwide.

Project Outputs

The intended outcome of this project is a robust and performant publicly accessible Virtual Research Environment, the *CLAROS Explorer*, to be used by academic researchers, educators, students and members of the public to access information about objects from classical antiquity, and their relationships with classical mythology.

Wider Benefits

The immediate purpose of this project is to create a high performance CLAROS Explorer VRE that we can deploy publicly in order to engage wider participation, both by users and additional data providers.

When the system performance is comparable with the existing Beazley Archive (http://www.beazley.ox.ac.uk/) database system, we will seek to publicize the service through the same channels that have made that site so popular.

With usable performance of the underlying CLAROS data web, we will be able to explore new techniques for accessing and presenting the CLAROS data, such as using Shuffl (http://code.google.com/p/shuffl/) to perform comparative timeline or geographical distribution visualizations of objects from antiquity.

We will also document and publish our findings for others in the Semantic Web and Linked Web Data communities, in the hope that our successes can be replicated in other RDF storage systems (and our failures avoided). Techniques for overcoming performance problems with triple stores are of great current interest within this community.

Finally, this MILARQ work could be a stepping stone on the way to more automated query optimization and planning for RDF data stores, which we know to be of interest to some Jena developers, by providing a platform for manual testing of possible optimizations before committing to the expense of coding a full-blown optimizer.

Risk Analysis

Technical risk. The technical aspects of this project are tightly scoped and well understood. In particular, we will be enhancing the existing LARQ query and Lucene indexing mechanisms rather than creating new technology in this area. It is conceivable, though unlikely, that when we come to deployment and testing we will find that some of the anticipated benefits are not realized: if so, this information will be a valuable lesson for practitioners using this technology, as there are many discussions about building a more automated version of what we are undertaking. By working closely with the original Jena developers now at Epimorphics, we aim to minimize the risk of technical surprises, and to lock the technical advances into the Jena code base.

Staff recruitment and retention. Graham Klyne, the lead developer and project manager, is already in post, so there is no risk of recruitment delay.

Epimorphics Ltd will be a full project partner, paid on a consultancy basis. Their presence in the project mitigates the unlikely loss of the lead developer, which otherwise would be fatal for such a short project. The agile approach and open sourcing of project materials from the outset also mitigates the risk of losing all outputs in such an event.

Failure to engage with Jena project. Sustainability of this work is predicated on it becoming part of Jena and related open source projects. We will ensure that by means of the Epimorphics partnership. We already have good personal and professional working relationships with many of the Jena developers individually, and aim to build upon these though this partnership.

Jena project uncertainties. The Jena project itself, formerly lead and substantially funded by Hewlett Packard Laboratories, has recently transitioned into an independent open-source project. This implies some undeniable risks, but the high level of interest in Jena globally [JenaUsers], and the determination

of the Jena team from HP Labs to maintain the project, are all positive signs. Many of the original Jena team will be continuing to work with Jena, many as part of our partner Epimorphics, and some with other employers (notably Talis).

We have indications that our CLAROS work is providing valued input to the future development of Jena, and we would expect this project to help build relationships that will help other JISC-funded projects that use Jena to take a more proactive role in ensuring its future. Indeed, at least two other projects, one in the same VRERI programme, have expressed interest in our outputs.

Intellectual Property Rights

The MILARQ software enhancements to LARQ and/or SPARQLite will be compatibly licensed and contributed back to their respective open source projects, where they may benefit other systems based on these components and may be maintained by Jena community developers.

Documentation of MILARQ and its usage will be released under a Creative Comments attribution licence.

The specific CLAROS deployment and configuration will be owned by Oxford University. All data remains property of the respective contributing partners, who Oxford and others to redistribute selected data for the purposes of deploying CLAROS.

Project Relationships

The Beazley Archive, in Oxford University's Classics department leads development of the CLAROS system, with assistance from the Oxford e-Research Centre (OeRC), Oxford University Computing Services (OUCS), Engineering Science Robotics Group and Zoology Image Bioinformatics Research Group (IBRG). Of these, only IBRG is in receipt of JISC funds for this project.

The CLAROS project is being developed by a partnership currently of Beazley Archive, Oxford; German Archaeological Institute, Berlin; Lexicon of Greek Personal Names, Oxford; Lexicon Iconographicum Mythologiae Classicae (LIMC), Basel and Paris; and Research Archive for Ancient Sculpture, Cologne. None of these are receiving MILARQ project funds.

The MILARQ development work is being conducted in collaboration with Epimorphics Ltd, who are providing consulting and development services at favourable rates as the work also benefits the Jena open source software framework.

Work Plan

The project will be run along agile lines, as a series of "sprint"s or iterations, which will aim to achieve the progress outlined above. Each sprint will be planned in greater detail when it starts, adjusting the goals to take account of progress made and user feedback to date.

The target sprint duration is 4 weeks (which is longer than often desired, but reflects (a) the "part-time" nature of the project, and (b) the overhead of having too-frequent meetings between distributed projects). Sprints 3 and 4 are planned to be shorter as this is anticipated to be where the bulk of

development work is progressed. The final sprint is extended as this is expected to mainly involve "mopping up" activities.

The sprint reviews should be supplemented by weekly progress summaries posted to the project blog, which ''(experimentally)'' may stand in lieu of formal project retrospectives.

Outline of plan (months):

1-2: project setup, implementation plan, outline specification and test plan;

3-4: initial implementation;

5-6: evaluation and refinement;

7-8: CLAROS public deployment, documentation, hand-over to Jena community.

Month	Activity	Deliverables
1	Initial planning. Project setup, and discussion with Epimorphics about engagement with the open source Jena project. Analyze existing Jena/LARQ code and plan changes. Outline test framework for changes.	Project plan. Publicly hosted project tracking facility and source control (assuming no committer status with Jena project at this stage). Reviewed implementation plan documents.
2	Agree initial functionality with Epimorphics. Outline test cases for multiple indexes. Code initial test cases.	Initial test cases run-able (not expected to pass).
3	Implement Jena and/or SPARQLite extensions to satisfy test cases.	Jena/SPARQLite code passes initial test cases.
4	Continue coding test cases and implementation. Start documenting new APIs and/or configuration.	Initial functionality complete and unit-tested. Initial documentation.
5	Create CLAROS test deployment. Update documentation in light of experiences. Develop load tests. Run load tests, compare with current CLAROS.	Initial CLAROS deployment with new indexes. Initial performance results. Updated documentation.
6	Continue deployment, load testing and evaluation. Document findings. Package software for more widespread deployment. Seek code and documentation reviews from Jena community.	Documented evaluation. Refined code, documentation. Deployable software package for updated query service. CLAROS service ready to go public.
7	Evaluation by CLAROS team. Write up approach and results for general publication.	Public deployment of CLAROS. Documented approach and lessons.
8	Finalize packaging for deployment. Create patches of contributions to existing open source projects.	Code tested, documented evaluated, lodged with existing open source projects.

Week-by-week progress reporting will be primarily via the project announcements blog, with less transient project information recorded in the project wiki. Documentation about general enhancements to the Jena and SPARQLite software will become part of their respective projects.

Success Plan

MILARQ is conceived as a small number of incremental improvements to existing projects, and the standards and practices of the existing projects will be adopted from the outset. By focusing as much as possible on improving the existing Jena code, we aim to become part of an existing vibrant developer- and user-community.

The collaboration with Epimorphics is intended to ensure that any modifications to the Jena code base are applied by developers who are familiar with the Jena code, Jena decvelopment community practices and the supporting technologies.

Testing and evaluation will be test-led, based on the existing CLAROS test suite, which will ensure that required functionality is maintaned and also provide a baseline for measurring performance improvements.

Success criteria are simple: the existing CLAROS VRE provides the existing capabilities with greatly improved performance for certain key queries. Each of the queries in the test framework, suitably modified to use the additional indexes, should produce the same results with sub-second response times. A secondary, non-essential, success criterion for CLAROS is that the timeline data can be retrieved using a single query rather than a series of queries as at present.

Additional success criteria will include adoption of the MILARQ enhancements to Jena by other projects.

Budget

Table 2: MILARQ PROJECT BUDGET	01/03/10	01/08/10	01/01/10
	- 31/07/10	- 31/10/10	- 30/06/10
Oxford University: Directly incurred staff			
Research officer: 25%, 8 months (A)	£6,115.00	£3,781.00	£9,896.00
Oxford University: Non-staff			
Travel and expenses	£1,100.00	£900.00	£2,000.00
Equipment	£2,000.00	£0.00	£2,000.00
Total Oxford University directly incurred non-staff (B)	£3,100.00	£900.00	£4,000.00
Total directly total (C = A+B)	£9,215.00	£4,681.00	£13,896.00
Oxford University: Directly allocated			
Staff: Principal investigator, 2.5%	£567.00	£347.00	£914.00
Estates costs	£1,452.00	£922.00	£2,374.00

Infrastructure technician (centrally levied for laboratory-based departments)	£141.00	£90.00	£231.00
Oxford University directly allocated total (D)	£2,160.00	£1,359.00	£3,519.00
Oxford University indirect costs (E)	£5,412.00	£3,438.00	£8,850.00
Total Oxford University FEC (=C+D+E)	£16,787.00	£9,478.00	£26,265.00
Epimorphics Ltd Consultancy (F)	£7,050.00	£7,050.00	£14,100.00
Epimorphics Ltd Contribution in Kind (G)	£5,875.00	£5,875.00	£11,750.00
Total project full economic costs (H = C+D+E+F+G)	£29,712.00	£22,403.00	£52,115.00
AMOUNT REQUESTED FROM JISC (80% of Oxford FEC; 55% of consultancy fee)	£20,480.00	£14,632.00	£35,112.00
OXFORD UNIVERSITY CONTRIBUTION (20% of Oxford FEC)	£3,357.00	£1,896.00	£5,253.00
EPIMORPHICS CONTRIBUTION (45% of consultancy fee)	£5,875.00	£5,875.00	£11,750.00
JISC CONTRIBUTION AS PERCENTAGE OF TOTAL	69.00%	65.00%	67.00%