Integrating eSAD (The Image, Text, Interpretation: e-Science, Technology and Documents) and VRE-SDM (VRE for the Study of Documents and Manuscripts) projects

David Wallom, Segolene Tarte, Tiejun Ma, Pin Hu and Kang Tang

University of Oxford

The aim of the project was to enable the use of eSAD developed image-processing algorithms within the framework developed within the VRE-SDM. This means bridging the gap between the use of the NGS for image processing and the web based access mechanisms used by this research community.

The image processing algorithms, simple ones such as brightness and contrast adjustment, illumination correction, woodgrain removal, as well as more complex ones such as stroke detection, are offered as functionalities wrapped in one or several web-services and presented to the user in a portlet in the VRE-SDM application. Similarly, access and search through the knowledge base currently being constructed as part of the ISS would also be offered either within the same portlet as the image processing tools or in a different portlet in the VRE-SDM application. The underlying algorithms are wrapped in an application that has been installed on the UK NGS using the Uniform Execution Environment tools developed to ensure common locates for different software packages where they are installed on a number of different physical resources.

Before this project the algorithms and tools developed within eSAD were difficult for a non-computationally aware researcher to access as well as the main highly complex algorithms also taking significant processing to run on a single system, which made them significantly cumbersome for a researcher to try multiple methods on an image and experimentally find the best results. The interface developed within eSAD was a test system and as such not conducive to the long term storage and maintenance of the raw or processed images and so integration with the data model within VRE-SDM was also necessary. This allows metadata to be associated with the input and processed images and a record to be kept of the operations that were performed on the image.

Within the system there were in separate strands of development, each with a separate developer working directly on it. Firstly, an installation system for the algorithm had to be developed to fit into the NGS Uniform Execution Environment. Since the algorithms have been developed within the MATLAB toolkit they cannot be installed onto remote systems easily in their native format. The project devised a mechanism by which the MATLAB compiler could be used to provide distributable binaries that were then called by the user. This also meant that we have to distribute the MATLAB runtime libraries with the binaries. The initial building of the binary files was done on an architecturally compatible system within the OeRC.

Once compiled the resulting files and libraries were packaged into a format that could be easily distributed with a developed installer script. This not only copies all the executable files into areas on the system which are shared between head and execution nodes but also creates the soft links etc that have to be made to allow the software to operate within the NGS uniform execution environment. The system also includes verification script to ensure correct installation at the remote site.

To allow the user community easy access a user interface had to be developed for the system. This had to fit easily into the already developed system that came from VRE-SDM but be extended to deliver additional functions to facilitate secure and seamless distributed image analysis to users. Furthermore, the portlet transforms most server-side (Java) logic used in the VRE-SDM system to “Asynchronous JavaScript and XML” (AJAX) based dynamic client-side (JavaScript) components in order to make the system much easier to integrate into different Web application or Portal frameworks; enhance the interactivity between server and client; and eliminate unnecessary page reloads at the user-end. The finished interface is shown in figure 1. Overall, the portal part of the system delivers a set of functions in both frontend and backend, including:

Frontend

- Search original images or processed images obtained from analysis;
- Display thumbnails and image metadata;
- Display single or tiled images;

- View and edit annotations on images;
- Select algorithms for analysis;
- Select region of interest for analysis;
- Run analysis on individual images or a group of images belonging to the same object view;
- View real-time GridSAM job status;
- Dually display single images.

(Backend)

- Access metadata from TripleStore;
- Manage and locate images;
- Generate JSDL job descriptions;
- Asynchronous GridSAM job submission and monitoring;
- Make tiles for processed images;
- Generate and add metadata for processed images.

The final component of development work is the interface between the portal and NGS installed GridSAM instances. A key benefit we found during this project when using GridSAM is that, it also helps to keep a clean boundary between normal web application development and Grid development. The latter is usually considered as complicated and error-prone. By using the GridSAM client side API together with the simple job manager wrapper we developed during the project, the portal server can easily submit jobs and query job status without knowing any details of grid resources in the back end. At the end of the process chain on the portal, each job is represented as a JDSL description file that contains following information:

- Executable and argument
- Standard output and error files
- Data Staging in/out information
- Access information

GridSAM supports various protocols for data staging, such as FTP, GridFTP and SFTP, etc. though we decided to use SFTP due to security and simplicity. However, at the moment when this project was carried out, the latest version of GridSAM (v2.1.14) supports only user name and password authentication when using SFTP for data staging, which means we will have to embed access information for file servers into JDSL description. The security risks can be mitigated using SSL. When setting up the GridSAM service in Oxford, we have identified this as a potential security risk and used SSL to encrypt all the plain SOAP messages between portal and GridSAM server. The full data flow within the project is shown in figure 2.

The project outcomes are both technical and sociological. Firstly we have created a build and installation configuration toolkit that can be used by researchers from the eSAD project to distribute the underlying tools that they create onto the resources of the NGS. This includes all of the scripts to create the binary products that can then be redistributed whilst protecting their IP. We have proved that utilizing the NGS UEE scripts to produce a simple common location for application installations is useful and enables researchers to work with code installed at a number of sites.

The researcher is now provided with the necessary tools and knowledge to add new functionality to the portlet, without having to worry about how the connection to the NGS happens. The implication of this for the community are that users now have access to an easy to use tool, which will facilitate the dissemination of the outcomes of the eSAD project. It also proves the usability of the VRE.

We have shown that it is possible to not overwhelm researchers with the technical details of how the NGS works and communicates with a web-based application! All the researcher needs to know, is that it is possible to use the NGS, and that it doesn't mean that they have to worry excessively about the details of the communication protocols between the NGS and a web-based application.

The authors would like to thanks OMII and NGS who funded this work through the ENGAGE project as well as the partner resources of the NGS on which this work is executed.

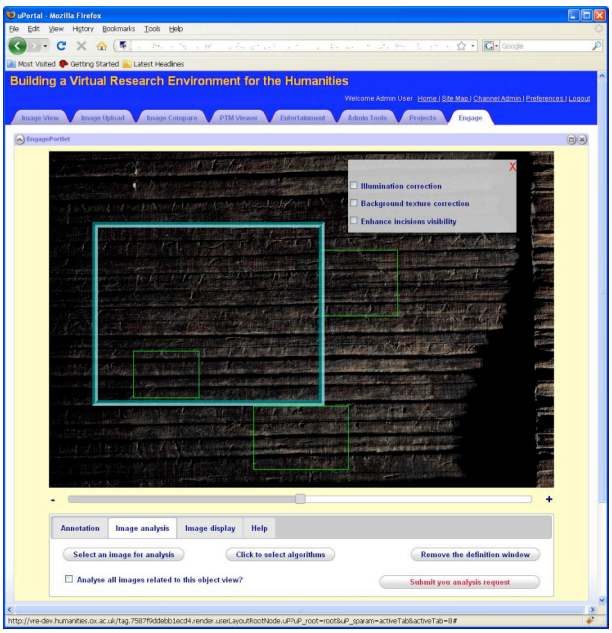

Figure 1 The user interface portal showing enhancements and selections of regions within an image

Figure 2 the components showing connections between the portal, jobmanager, GridSAM and the NGS