# Web Services for the DDSM and Digital Mammography Research

Chris Rose, Daniele Turi, Alan Williams,
Katy Wolstencroft, and Chris Taylor

Imaging Science and Biomedical Engineering,
The University of Manchester,
Manchester, M13 9PT, United Kingdom
http://www.medicine.manchester.ac.uk/isbe/

**Abstract.** The Digital Database for Screening Mammography (DDSM) is an invaluable resource for digital mammography research. However, there are two particular shortcomings that can pose a significant barrier to many of those who may want to use the resource: 1) the actual mammographic image data is encoded using a non-standard lossless variant of the JPEG image format; 2) although detailed metadata is provided, it is not in a form that permits it to be searched, manipulated or reasoned over by standard tools. This paper describes web services that will allow both humans and computers to query for, and obtain, mammograms from the DDSM in a standard and well-supported image file format. Further, this paper describes how these and other services can be used within grid-based workflows, allowing digital mammography researchers to make use of distributed computing facilities.

## 1 Background

The DDSM [6] provides high-resolution digitised mammograms, expert ground-truth and metadata (including the date of study and digitisation, the Breast Imaging Reporting and Data System (BI-RADS) [1] breast density and assessment categories, a subtlety rating, the type of pathology and detailed categorisation of the nature of the perceived abnormality using the BI-RADS lexicon). The DDSM is available free of charge by File Transfer Protocol (FTP).

While the DDSM does provide software to decode their mammograms[1], the default distribution of this software does not build under modern compilers without modification, a step that may prove difficult to those with insufficient experience of C/C++ software development for UNIX-like operating systems. Furthermore, even when properly compiled, the DDSM software outputs the image data as a stream of raw bytes; one then has to normalise these according to the model of digitiser used to image the original films and then create an image file that is readable by one's image analysis software environment. An introduction to web services is given in Section 1.1. Section 2.1 describes a web service that allows digital mammograms from the DDSM to be obtained

[1] In particular the DDSM's `jpeg` program.

Susan M. Astley et al. (Eds.): IWDM 2006, LNCS 4046, pp. 376–383, 2006.

in a standardised and well-supported lossless file format. Section 2.2 describes a service that allows groundtruth images to be obtained in the same file format.

While a web-based query tool is provided by the DDSM, it is useful only to human users or automated tools that have been specifically designed with the DDSM in mind. If the metadata were in a more useful format, it could easily be exposed for both human and computer use. Section 2.3 describes a formal ontology that has been developed to describe the DDSM resource and a web-based user interface to allow users to query the ontology.

Section 2.4 describes how web services can be used together within *workflows* to run full experiments and how a full record of how such experiments were performed can be recorded by capturing provenance events. Section 2.5 details a supporting website for the work described in this paper.

### 1.1 An Introduction to Web Services

The concept of web services may best be explained with a simple example of a hypothetical scientist named Bob who lives in Manchester, UK. Bob has a CAD algorithm that other scientists want to use. Traditionally, Bob would package his CAD algorithm into some form that is easily installed and run by other scientists. He would then deliver it to those scientists via Internet download or on physical media (e.g. CD-ROM). However, Bob might not be able to let other scientists run his software on their computers because:

- Bob may not have planned to share his software and may have made assumptions in its design that limits its portability;
- users might need an expensive license to use a required proprietary library;
- the software may need access to a resource (e.g. a large database) that resides at Bob's lab;
- Bob might frequently update his software, so making each update available to all his users might be troublesome;
- packaging the software for easy installation might be too time-consuming for a busy research scientist.

Bob might decide that it is easier to allow other scientists to run his software on his computer, accessing it via the Internet. This can be achieved by exposing his software as a web service. Using an appropriate piece of client software that implements the Simple Object Access Protocol (SOAP) standard [5], other scientists can run Bob's CAD algorithm on their data. In this way, Bob's CAD algorithm can be used by remote scientists as if it were installed on their computers, or integrated into software as if it is a library containing the required functionality.

The "interface" to a web service—the location of the computer that provides the service and a specification of its inputs and output and their data types—is described using the Web Service Description Language (WSDL) [2]. The URL of a service's WSDL file is all that is needed for a SOAP client to be able to use the service[2].

[2] The WSDL files for the services described in this paper can be obtained from the website described in Section 2.5.

The web services proposed in this paper open up the possibility of performing digital mammography research using grid-based workflows. In this context, a *grid* is an *ad hoc* collection of computing services offered by a number of (typically) different computers via a network (typically the Internet) and a *workflow* specifies how the services provided by the grid are orchestrated in order for some task to be performed. Section 2.4 describes workflows in more detail.

## 2 Method

### 2.1 Digital Mammogram Web Service

To create a web service that makes available mammograms from the DDSM, we developed a command-line program that, given the name of a particular DDSM mammogram (e.g. D_0160_1.RIGHT_MLO), downloads the associated `.LJPEG` file from the DDSM's FTP server, decodes the raw image data, normalises it according to the digtiser used and finally converts it to a PNG file [4]. This format is ideal for encoding mammograms as it is standardised, guarantees lossless compression and is widely supported by common software tools and libraries[3]. (In future, other lossless image formats may be more suitable—such as JPEG-LS and JPEG 2000—but as of this writing these formats are not widely supported.)

Downloading and converting the images takes a few minutes on a desktop computer with a fast connection to the Internet and so the program caches some of the converted mammograms locally so that future requests can be efficiently serviced. Our program is exposed as a service via SOAP [5]. DDSM mammograms can be obtained from this service using any SOAP client.

### 2.2 Groundtruth Web Service

To enable the evaluation of algorithms run on the DDSM mammograms delivered by the service described in Section 2.1, there is a need to be able to obtain the corresponding groundtruth images. We have developed a web service that allows DDSM groundtruth images to be generated and delivered in a suitable image format. Our approach to developing this service was the same as that described in Section 2.1. We first developed a command-line program that, given the name of a particular DDSM mammogram (e.g. D_0160_1.RIGHT_MLO) and an abnormality number[4], downloads the corresponding `.OVERLAY` file (which contains a description of the shape of the radiologist-annotated abnormalities for the image using a chain code). The DDSM groundtruth metadata defines two possible types of region: a *'boundary'* and a *'core'*, though *'core'* regions may be absent (see [6] for details). Our program then creates a image with the same number of rows and columns as the corresponding digital mammogram—i.e. there is a one-one correspondence between every pixel in a digital mammogram and the

[3] We encode DDSM mammograms as 16 bits/pixel grey-level images.

[4] Mammograms may contain more than one abnormality. The abnormality number is captured by the ontology that is described in Section 2.3.

pixel at the same location in its groundtruth image—and this image is populated with pixel values that indicate the class of each pixel. A value of zero represents normal tissue or the non-breast region, a value of 128 represents a pixel within a *'boundary'* region of the abnormality and a value of 255 represents a pixel within a *'core'* region of the abnormality. The resulting groundtruth image is saved as a PNG image. While downloading the `.OVERLAY` file and creating a groundtruth image takes approximately a minute, the resulting PNG files are very small (approximately 14 kb) due to the high correlation between successive scan lines in the images, and we therefore keep all generated images to efficiently service future requests. This program is then exposed as a web service via SOAP.

### 2.3 DDSM Ontology and Metadata Query Service

Being able to obtain mammograms and groundtruth is not particularly useful without knowing which mammograms have what characteristics. To this end, we have developed a formal *ontology* of the mammograms, groundtruth and metadata (e.g. abnormalities, patient information). An ontology is a description of the concepts and relationships that exist in some knowledge domain. The formal representation of metadata within ontologies (using technologies such as the Web Ontology Language (OWL) [9]) allows domain knowledge to be used alongside explicit labelling to infer implicit relationships and hence deliver more useful results. As a simple example, if an ontology were to state that a stellate lesion is a type of mass, then a query for masses could return—in addition to items explicitly labelled as masses—items that were labelled as stellate lesions; i.e. the domain knowledge captured in the formal ontology allows it to be inferred that items labelled as stellate lesions must also be returned.

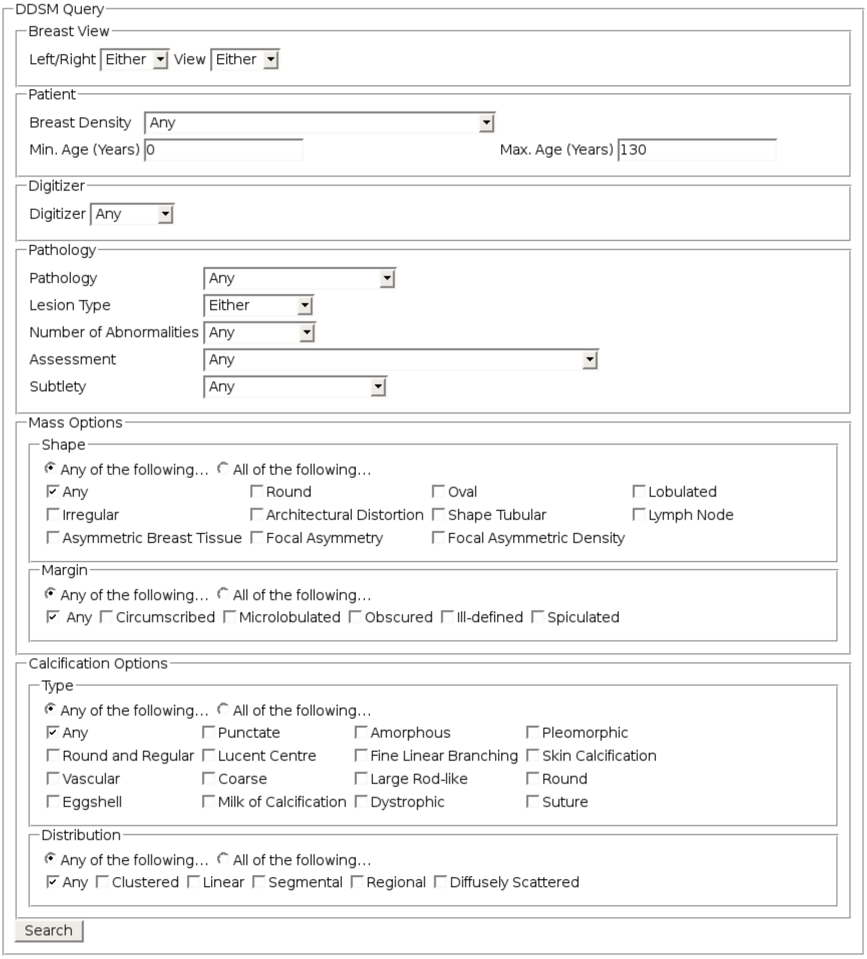
**Previous Work.** The most relevant work on ontologies for digital mammography was done by the Medical Imaging with Advanced Knowledge Technologies (MIAKT) project, which developed a fairly complete ontology for breast cancer imaging studies called the Breast Cancer Imaging Ontology. This multi-level ontology incorporated classes for both X-ray and MRI breast imaging, for abnormal findings and medical assessments [3]. The project also used the DDSM images as exemplar breast X-ray image data [7]. In contrast, the ontology that we have developed is more narrowly confined to the DDSM database, as justified below.

**Our DDSM Ontology.** Within the DDSM, information about the mammograms is specified in an `.ICS` file and, for each image that contains abnormalities, an `.OVERLAY` file. The `.ICS` file contains information common to the case e.g. the patient's age, and also information necessary to interpret the four mammograms e.g. the number of pixels per scan line. The `.OVERLAY` file contains information particular to the abnormality, or abnormalities, that have been interpreted within a particular mammogram e.g. the left CC mammogram.

The ontology that has been developed for the DDSM allows the description of the information within the `.ICS` and the `.OVERLAY` files. The ontology is written in OWL [9]. A decision was made that the ontology would only describe the information specified within the DDSM, in particular it would not attempt to be a general model of mammograms, mammogram interpretation or breast cancer—as the MIAKT project developed—as we are interested only in making the DDSM database easily available.

For an individual case, an OWL ontology is populated with RDF (Resource Description Framework [8]) instances. The instance ontology combines the information within the `.ICS` and `.OVERLAY` files into a single semantic structure. This allows the easy searching of the instance ontology when it is loaded into an RDF repository.

Within the DDSM ontology, the *'case'* class specifies information that applies to a patient's visit and their four images. It has four relationships to *'views'* corresponding to the four mammograms. *'Views'* are subclassed into either abnormal or normal views. The information about the image is held within the



DDSM Query

Breast View

Left/Right: Either ▾ View: Either ▾

Patient

Breast Density: Any ▾

Min. Age (Years): 0 Max. Age (Years): 130

Digitizer

Digitizer: Any ▾

Pathology

| | |
|---|---|
| Pathology | Any ▾ |
| Lesion Type | Either ▾ |
| Number of Abnormalities | Any ▾ |
| Assessment | Any ▾ |
| Subtlety | Any ▾ |

Mass Options

Shape

● Any of the following... ○ All of the following...

☑ Any ☐ Round ☐ Oval ☐ Lobulated
☐ Irregular ☐ Architectural Distortion ☐ Shape Tubular ☐ Lymph Node
☐ Asymmetric Breast Tissue ☐ Focal Asymmetry ☐ Focal Asymmetric Density

Margin

● Any of the following... ○ All of the following...

☑ Any ☐ Circumscribed ☐ Microlobulated ☐ Obscured ☐ Ill-defined ☐ Spiculated

Calcification Options

Type

● Any of the following... ○ All of the following...

☑ Any ☐ Punctate ☐ Amorphous ☐ Pleomorphic
☐ Round and Regular ☐ Lucent Centre ☐ Fine Linear Branching ☐ Skin Calcification
☐ Vascular ☐ Coarse ☐ Large Rod-like ☐ Round
☐ Eggshell ☐ Milk of Calcification ☐ Dystrophic ☐ Suture

Distribution

● Any of the following... ○ All of the following...

☑ Any ☐ Clustered ☐ Linear ☐ Segmental ☐ Regional ☐ Diffusely Scattered

**Fig. 1.** The DDSM metadata query form

*'view'* superclass. If the *'view'* is an *'abnormal view'* then it has relationships to one or more *'abnormalities'*.

An *'abnormality'* contains information such as the assessment and subtlety. It also has specific information about the calcification or mass intrepretation of the abnormality. In addition, the bounding curves of the abnormality and any cores within it are specified.

We have automatically populated an RDF store with instances of the classes in our ontology by processing the DDSM metadata files. We are currently developing a web-based user interface that will allow users to query the RDF store for images and groundtruth in a user-friendly manner. Figure 1 shows the form used to create queries.

### 2.4 Workflow Enactment and Provenance Capture

A workflow describes how a number of services can be combined to perform some useful task. The Taverna workbench program—a Java application that originated in the bioinformatics research community—allows users to create and run workflows within a graphical user interface [10]. Taverna displays workflows as directed graphs, where nodes represent inputs, services or outputs and arcs represent the flow of data and control (see Figure 2 for an example). Workflows can be saved and easily exchanged between colleagues. Taverna allows users to run, pause, monitor and debug workflows in a manner similar to modern software development environments. Workflow results can be directly displayed within Taverna.

Aside from allowing researchers to make use of distributed computing resources, Taverna can capture *provenance events*—e.g. when a particular workflow was started and with which inputs—allowing the workbench to operate as an automated laboratory log book. This also allows researchers to obtain

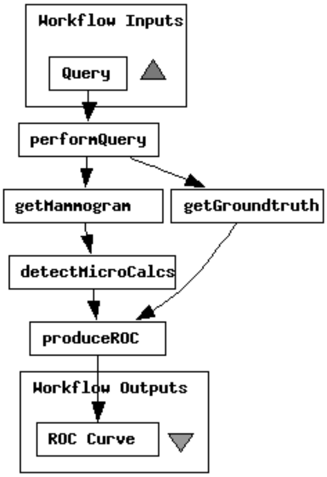


**Fig. 2.** A simple Taverna workflow

answers to questions like *'which images were used as inputs to the workflow I ran on 24 January 2006?'*.

### 2.5 Supporting Website

Those who wish to use the services described in this paper are directed to `http://www.digital-mammography-services.net`. This website will provide the most up-to-date documentation of the available services and provide links to the WSDL files that specify the web services described in this paper. We invite the community to make useful software and data available via services and are keen to document these at the above website.

## 3 Results

Our work is at a relatively early stage, but we already have useful services and infrastructure. The most significant contributions are the digital mammogram and groundtruth "getter" services, the DDSM ontology, the web-based query facility and the supporting website. We hope these will be useful to the digital mammography research community. While we are using the Taverna workbench software to integrate our services into simple grid-based workflows, our general approach—publishing our software as SOAP services—does not require clients to use Taverna; any SOAP client can be used.

## 4 Discussion

We have described three web services which allow both humans and computers to query a formal ontology of the DDSM data and obtain digital mammograms and groundtruth from the DDSM in a well-supported standard image format. We have also described how these services could be used within grid-based workflows. As Section 1 described, obtaining mammograms from the DDSM is currently non-trivial and it is hoped that the web services described in this paper will make using this important resource more convenient.

It is difficult to quantitatively evaluate the type of work that is described in this paper. While we could measure the speed with which requests can be processed, or the number that can be handled concurrently, such measurements do little to tell us if we have achieved our aims of developing and deploying infrastructure that is useful to the community. This will become apparent in time as the resources described in this paper are used (or otherwise) and if other researchers contribute their software and data in the form of web services for use by the community. In this spirit we welcome criticism and suggestions and are able to offer advice on an informal basis to those interested in developing their own web services.

Future work will focus on exposing other useful algorithms as web services (e.g. a CAD task such as microcalcification detection, a receiver operating characteristic (ROC) analysis service) and on maintaining the website described in

Section 2.5. Given these services, it will be possible to run a simple but typical digital mammography CAD experiment using web services (i.e. obtain mammograms → process each image using the CAD algorithm → obtain ground-truth → produce an ROC curve). By publishing the workflow, others in the community would be able to replicate the experiment exactly or swap one service (e.g. the CAD task) for their own to be able to fairly compare algorithms.

## References

1. The ACR Breast Imaging Reporting and Data System (BI-RADS). American College Radiology, 1998. Third Edition.
2. Erik Christensen, Francisco Curbera, Greg Meredith, and Sanjiva Weerawarana. Web Services Description Language (WSDL) 1.1. W3C Note, World Wide Web Consortium, March 2001. (A W3C Recommendation for WSDL 2.0 is currently pending.).
3. S. Dasmahapatra, B. Hu, P. Lewis, and N. Shadbolt. Ontology-Based Medical Image Annotation with Description Logics. In *15th IEEE International Conference on Tools with Artificial Intelligence*, Sacramento, CA, USA, 2003.
4. David Duce. Portable Network Graphics (PNG) Specification (Second Edition). W3C Recommendation, World Wide Web Consortium, November 2004.
5. Martin Gudgin, Marc Hadley, Noah Mendelsohn, Jean-Jaques Moreau, and Henrik Frystyk Nielsen. SOAP Version 1.2 Parts 1–2. W3C Recommendation, World Wide Web Consortium, June 2003.
6. M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer Jr. The Digital Database for Screening Mammography. In M. J. Yaffe, editor, *Digital Mammography: IWDM 2000, 5th International Workshop*, pages 212–218, Madison, Wisconsin, USA, December 2001. Medical Physics Publishing.
7. B. Hu, S. Dasmahapatra, D. Dupplaw, P. Lewis, and N. Shadbolt. Managing Patient Record Instances Using DL-Enabled Formal Concept Analysis. In *14th International Conference, EKAW 2004*, Whittlebury Hall, UK,, October 2004.
8. Frank Manola and Eric Miller. RDF Primer. W3C Recommendation, World Wide Web Consortium, February 2004.
9. Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium, February 2004.
10. Tom Oinn, Tom Greenwood, Matthew Addis, M. Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew R. Pocock, Martin Senger, Robert Stevens, Anil Wipat, and Chris Wroe. Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience Grid Workflow Special Issue (Accepted)*, 2005.