Building a Semantic Web Image Repository for Biological Research Images

Jun Zhao, Graham Klyne, and David Shotton

Department of Zoology
University of Oxford
South Parks Road, Oxford
OX1 3PS United Kingdom
{jun.zhao,graham.klyne,david.shotton}@zoo.ox.ac.uk

Abstract. Images play a vital role in scientific studies. An image repository would become a costly and meaningless data graveyard without descriptive metadata. We adapted EPrints, a conventional repository software system, to create a biological research image repository for a local research group, in order to publish images with structured metadata with a minimum of development effort. However, in its native installation, this repository cannot easily be linked with information from third parties, and the user interface has limited flexibility. We address these two limitations by providing Semantic Web access to the contents of this image repository, causing the image metadata to become programmatically accessible through a SPARQL endpoint and enabling the images and their metadata to be presented in more flexible faceted browsers, jSpace and Exhibit. We show the feasibility of publishing image metadata on the Semantic Web using existing tools, and examine the inadequacies of the Semantic Web browsers in providing effective user interfaces. We highlight the importance of a loosely coupled software framework that provides a lightweight solution and enables us to switch between alternative components.

1 Introduction

Images are semantic instruments for capturing aspects of the real world, and form a vital part of the scientific record for which words are no substitute. In the digital age, the value of images depends on how easily they can be located, searched for relevance, and retrieved. Images are usually not self-describing. Rich, well-structured descriptive image metadata thus carries high value information, permitting humans and computers to comprehend and retrieve images, and without them an image repository would become little more than a meaningless and costly data graveyard. A public image repository should thus publish its images along with such metadata.

In this paper, we present the *Drosophila* Testis Gene Expression Database, FlyTED [1], which presently contains images of expression patterns of more than

[1] http://www.fly-ted.org/

S. Bechhofer et al.(Eds.): ESWC 2008, LNCS 5021, pp. 154–169, 2008.

five hundred genes that are expressed in the testis of the fruitfly *Drosophila melanogaster*, both in normal wild type and in five meiotic arrest mutant strains, revealed by the technique of *in situ* hybridisation. These images were created as part of an ongoing research effort [1] by Helen White-Cooper and her team, who have also provided us with user requirements. Each image in FlyTED is described by the following *domain-specific* metadata: 1) the gene name, 2) the strain name, and 3) the gene expression pattern, in addition to the metadata about experimental details. This database aims to enable scientists to search for images of particular genes and compare their expression patterns between wild type flies and mutant strains.

To avoid building yet another purpose-built database, and to create a working database rapidly with the minimum development effort, we developed FlyTED by adapting an existing open source repository software system, EPrints[2]. The initial implementation of FlyTED was presented to our biological colleagues and received positive feedback. However, because EPrints is designed for text publications and is not Semantic Web-enabled, it is not effective in making its metadata programmatically accessbile and linkable to other *Drosophila* data resources, such as the Berkeley Drosophila Genome Project (BDGP) database of gene expression images in *Drosophila* embryos[3], or the global database of *Drosophila* genomic information, FlyBase[4]. Furthermore, although the built-in EPrints user interface can be customised to improve data presentation using technologies such as Cascading Style Sheets (CSS), it is not trivial to support some of the advanced image browsing functionalities requested by our researchers, such as filtering images first by gene name, then by mutant name, and then by expression pattern.

These limitations of metadata dissemination and user interface motivated us to enhance FlyTED with Semantic Web technologies, in order to make its images and metadata Semantic Web accessible. This has enabled the FlyTED image metadata to be queryable through the SPARQL protocol, and the images to be accessible through simple HTTP requests, enabling the use of faceted Semantic Web browsing interfaces.

The goals of this paper are to:

- Share our experience of enhancing an image repository developed using a conventional repository software package with Semantic Web technologies, which may be useful to others holding legacy data in conventional databases.
- Show the benefits obtained by making this image repository Semantic Web accessible.
- Discuss lessons we learnt through this experiment, summarised as the following:
  - Available Semantic Web tools can be employed to support Semantic Web access to data held in pre-existing Web applications.
  - The creation of effective user interfaces still remains challenging.

[2] `http://www.eprints.org/`
[3] `http://www.fruitfly.org/`
[4] `http://www.flybase.org/`

- A software framework that permits loose coupling between components shows its advantages, by enabling the reuse of existing toolkits and by facilitating switching between alternative software implementations.

2 Background

In this section, we provide background information about FlyTED and the EPrints 3.0 software system.

2.1 FlyTED

The spatial and temporal expression patterns of genes provide important knowledge for biologists in understanding the development and functioning of organelles, cells, organs and entire organisms. Biological researchers in the Department of Zoology at the University of Oxford are working towards this goal by determining expression data for approximately 1,500 genes involved in spermatogenesis in the testis of the fruitfly *Drosophila melanogaster*, representing ∼10% of the genes in the entire genome [1]. Comparative studies are being carried out in both wild type flies and five meiotic arrest mutant fly strains in which sperm development is defective. It is hoped that such studies can assist in the understanding of and the development of treatments for human male infertility.

In this work, for each gene, at least one *in situ* gene expression image is acquired from a wild type fly, and possibly one or more are taken from each strain. In addition, images are sometimes acquired of expression patterns of mutants of these genes in wild type flies, where such mutants are available. Thousands of images were accumulated by our biological researchers during the initial months of their research project. These images were kept in a file system organised by the date of creation and described using Excel spreadsheets. Each row in the spreadsheets describes one gene expression image by the following metadata terms (columns):

- The `GeneName` associated with this image.
- The `Strain` name of the fly from which the image was acquired.
- The gene `ExpressionPattern` being revealed in the image, as defined by the *Drosophila* anatomy ontology[5].
- Other domain-specific metadata, such as the number of the microscope slide on which the *in situ* hybridisation specimen was mounted for image acquisition.

Without organizing these images using structured metadata, it was extremely difficult for researchers to locate images from the file directories by their domain-specific metadata. A proper image repository was needed to assist researchers in uploading, storing, searching and publishing images with appropriate domain-specific metadata.

[5] `http://obofoundry.org/cgi-bin/detail.cgi?fly_anatomy`

2.2 EPrints

EPrints is an open source software package for building open access digital repositories that are compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [2]. The latest release, EPrints 3.0, was chosen for setting up FlyTED for the following reasons:

- It is one of the well-established repository software systems for archiving digital items, including theses, reports and journal publications, with ∼240 installations worldwide.
- It has built-in support for the OAI-PMH protocol, a simple HTTP-based protocol which allows repository metadata to be harvested by any OAI-PMH compliant parties.
- It has a built-in user interface, which makes it fairly quick to set up the repository and present it to users.
- It has previously been adapted by the Southampton SERPENT project[6] to publish images using domain-specific metadata.

2.3 Image Ingest

EPrints was designed as a digital text archive. It has good support for using Dublin Core (DC) metadata for describing and searching for digital items. However, to use EPrints to store and publish our biological research images along with domain-specific metadata, we needed to take the following steps:

- Preprocess researchers' Excel spreadsheets into a collection of free-text image metadata files, one for each image, using a Python script. Each file contains all the domain-specific metadata terms used by the researchers to describe their gene expression images. These metadata terms from different columns of the original spreadsheets are line separated in the metadata files.
- Modify the underlying EPrints relational database schema to accommodate some of the domain-specific metadata needed for both browsing and searching images, *i.e.*, the gene name, strain name and expression pattern.
- Ingest images and their metadata files into EPrints using a customised script written in Perl. This script achieves two goals: 1) extract from the image metadata files the metadata terms needed for searching and browsing images and store these metadata in the EPrints relational database when uploading images; 2) store the images as well as the individual metadata files for the images as binary objects within the EPrints database. Both the images and their image metadata files become Web-accessible.

The Perl and Python image ingest scripts are available at our SVN repository[7]. In addition to this customisation of the underlying EPrints database, we also needed to customise the user interface in order to permit searching and browsing for images using domain-specific metadata, and to enable the metadata within the search results to be displayed in the most useful manner. The details of this interface customisation can be found in [3].

[6] `http://archive.serpentproject.com/`

[7] `https://milos2.zoo.ox.ac.uk/svn/ImageWeb/FlyTED/Trunk/`

3 Problem Statement

This EPrints FlyTED repository clearly fulfils our initial goal of publishing images and their metadata on the Web. We can provide a working system for scientists and their community quickly and easily. However, it has two limitations:

- The OAI-PMH protocol fails to provide an effective programmatic access to the images themselves and to all their domain-specific metadata, in a form that can be linked with other data sources that are not OAI-PMH compliant.
- It is difficult to configure the user interface to permit researchers to browse, search and retrieve images by their domain-specific metadata for complex tasks.

3.1 Metadata Accessibility

The first limitation is caused by the nature of the OAI-PMH protocol, through which metadata in FlyTED can be programmatically accessed. OAI-PMH [2] is developed by the Open Archives Initiative for harvesting the metadata descriptions of resources in a digital archive. It is widely supported by repository software systems, including EPrints, DSpace[8], and Fedora[9].

It is compulsory for OAI-PMH compliant repositories to return XML format DC metadata for OAI-PMH requests. However, exposing domain-specific metadata in a semantic-rich data format is not natively supported by repository software systems, and would require a significant amount of work from repository administrators to implement. Furthermore, as a harvesting protocol, OAI-PMH does not allow *querying* of resources based on their (domain-specific) metadata values. To overcome these shortcomings of OAI-PMH, we chose to deploy a SPARQL endpoint over FlyTED.

SPARQL [4], a W3C-recommended RDF query language and protocol, has two distinct advantages over OAI-PMH for providing accessible metadata:

- SPARQL permits query selection by domain-specific metadata values. For example, a SPARQL query "`?image hasPatternIn Mid_elongation-stage_spermatid`" would query for all entities, locally designated by `?image`, that have an `ExpressionPattern` in `Mid_elongation-stage_spermatid`, which is a type of `Elongation-stage_spermatid`. OAI-PMH cannot do this.
- SPARQL query results are in RDF, which can be linked to the wider Semantic Web of data, including the metadata resources of BDGP that have also been exposed via a SPARQL endpoint[10].

3.2 User Interface Issues

We customised the EPrints user interface using CSS to improve the presentation of images. However, this interface was still unable to fulfil all the requirements

[8] `http://www.dspace.org/`
[9] `http://www.fedora.info/`
[10] `http://spade.lbl.gov:2021/sparql`

of our users, and we found it difficult to perform further customisations without significant modification of the software source code.

Each image in FlyTED is described by at least three key metadata properties, the `GeneName`, one of several `Strain` names, and one or more `ExpressionPattern` keyword(s). Scientists need to use these domain-specific metadata to examine images from different perspectives, for example, comparing expression patterns among all the images of the same gene in different strains. Furthermore, scientists need groups of images to be presented as thumbnails, so that they can obtain an overview of them and make comparisons between them.

The following list of requirements were articulated by our biologist colleagues for browsing and searching for images, ordered by their priorities:

- **R1:** Browse images by the gene name, strain name, or expression pattern keyword, preferably with images presented as thumbnails.
- **R2:** Search for all images of a gene that show the same expression pattern, *e.g.*, searching for all images of gene CG10396 with expression in `Mid_elongation-stage_spermatid`.
- **R3:** Search for all images of a gene that show the same *set* of expression patterns, *e.g.*, searching for all images of gene CG10396 with expression in both `Mid_primary_spermatocyte` and `Mid_elongation-stage_spermatid`.
- **R4:** Browse images of a single gene, first by the different strains in which this gene was imaged, and then by the expression patterns exhibited in those images.
- **R5:** Browse images from a particular strain by the (set of) expression pattern(s) shown in the images.
- **R6:** Find all the images of a particular strain NOT showing a certain set of expression patterns.

EPrints supports requirements R1-3, but not the others. It allows users to construct conjunctive conditions (such as both `Mid_primary_spermatocyte` and `Mid_elongation-stage_spermatid`) in the search interface, but not in the browsing interface.

EPrints provides *views* for users to browse repository records that are grouped by the value of their metadata properties. These views are dynamic web pages automatically generated by EPrints' built-in Perl scripts. They are capable of sorting and presenting repository resources by each different value of a property, but not by *a set of* values. For example, EPrints can provide a "Strain View" which groups images by each of the six different strain names, but it cannot provide a view grouping images by different sets of expression patterns, as required by R4-5. R6 is supported neither in the search interface nor in the browsing interface. To support requirements R4-6 using EPrints alone, we would have needed to put substantial effort into modifying EPrints' Perl source code, which would have consumed valuable human resources and led to a less sustainable software package.

Although R4-6 are less frequently required than R1-3, they are nevertheless essential to enable researchers to organise and integrate these images effectively, for example, by comparing the expression patterns of different mutants of the same

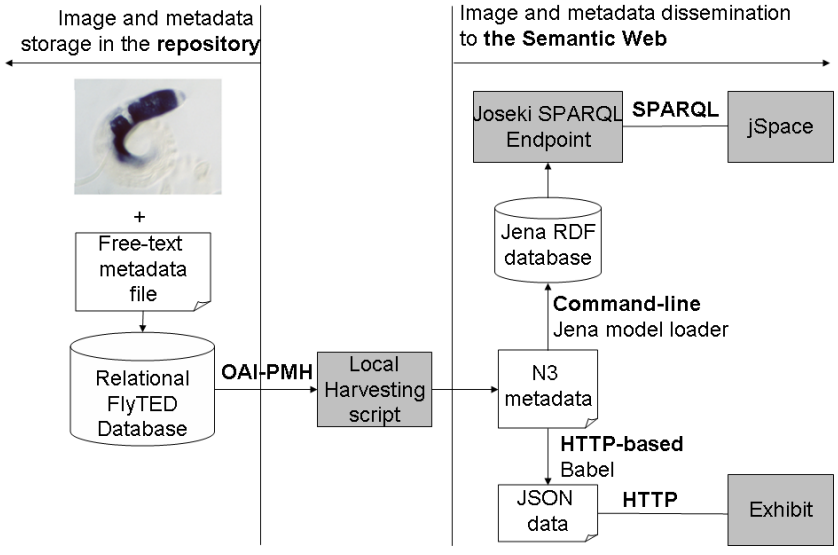


Fig. 1. The software framework of building a Semantic Web accessible image repository

gene or reviewing the group of genes expressed at different locations throughout spermatogenesis.

Requirements R4-5 closely correlate to the vision proposed by existing Semantic Web faceted browsing tools [5,6,7]. These tools provide user interfaces either for RDF data (such as Exhibit [6]) or for resources exposed through SPARQL endpoints (such as jSpace [7]). This means that by making our image metadata Semantic Web accessible, as either RDF metadata or a SPARQL endpoint, we would obtain an opportunity of exploring alternative and more flexible user interfaces for our images and associated metadata.

4 Publishing Semantic Web Accessible Metadata

Figure 1 shows the software framework that enables our images and their domain-specific metadata to become accessible to the Semantic Web. The gap between the EPrints FlyTED repository and the Semantic Web is bridged by a local harvesting script, which harvests domain-specific image metadata from FlyTED through OAI-PMH, and writes this metadata in Notation 3 (N3) format to a local disk. We then disseminate this RDF metadata to the Semantic Web, directly or through a SPARQL endpoint, enabling our image metadata to be accessible to any Semantic Web data resources and applications.

4.1 Metadata Harvesting

The local harvesting script extracted image metadata from the FlyTED repository and constructed semantic metadata in three steps:

- Harvesting of DC metadata of all the repository records using OAI-PMH.
- Analysis of the DC metadata of each record to extract the value of `dc:identifier` for two URLs: one pointing to the image and the other to its metadata file.
- Retrieval of the image metadata file from the FlyTED repository through the HTTP protocol using the metadata file URL, and construction of statements about this image in RDF.

As said in Sect. 2.3, both the images in FlyTED and their image metadata files are made Web-accessible. The meaning of each metadata term in the image metadata files had been well understood when we first constructed these files from our researchers' Excel spreadsheets during the image ingest. This understanding guided us in the creation of classes and properties that make RDF statements about each image using these metadata terms.

The harvesting script was written in Java, and built upon the open source OAIHarvester2 APIs from the Online Computer Library Center (OCLS)[11]. This harvesting script, and also the image ingest scripts for EPrints (written in Perl and Python), are available at our SVN repository[12].

4.2 Creating the SPARQL Endpoint

We built a SPARQL endpoint over FlyTED[13] using the Jena/Joseki[14] toolkit. This approach shows the following two advantages:

- The ability to use a lightweight HTTP-based protocol: the local harvesting script, described above, achieved harvesting metadata from the relational repository database through the simple HTTP-based OAI-PMH protocol.
- The ability to use a lightweight toolkit: the Jena model loader, a command-line tool from Jena, supports loading any RDF data into a Jena database without requiring any code writing, and its generic software interface means that it can be executed in any programming language. This makes our framework more sustainable.

This FlyTED SPARQL endpoint not only allows us to expose our image metadata through a programmatic interface that permits querying of metadata but also provides an interface for developers to execute SPARQL queries that are not included in the list of our users' requirements.

To summarize, this software framework for enabling our images and their metadata Semantic Web accessible (Figure 1) shows the following advantages:

- It avoids defining tightly constrained interfaces between components that can be inflexible and fragile when faced with evolving needs.
- It creates a lightweight software environment using simple RESTful (Representational State Transfer) [8] interfaces and HTTP-based protocols.

[11] `http://www.oclc.org/research/software/oai/harvester2.htm`
[12] `https://milos2.zoo.ox.ac.uk/svn/ImageWeb/FlyTED/Trunk/`
[13] `http://www.fly-ted.org/sparql/`
[14] `http://www.joseki.org/`

- It minimises the cost of development effort by adopting or adapting existing tools and services.
- It maximises the opportunity of replacing or updating any element of the technology used.

The images and their metadata, thus made available on the Semantic Web, can now be explored using Semantic Web faceted browsers, which present data in more flexible ways than does EPrints, without requiring any software installation.

5 Faceted Image Browsing

A faceted browser [5,6,7] presents categories of a knowledge domain to a user, which assists the user in filtering information by selecting or combining categories. This type of interaction enables users to query and manipulate information in an intuitive manner without having to construct logically sophisticated queries, which requires specialised knowledge about query languages and the underlying data model. This freedom is what is greatly valued by biological researchers: provision of a flexible interface for exploring their datasets without imposing additional cognitive demands. Two faceted browsers, Exhibit [6] and jSpace [7], were evaluated for accessing our image data.

5.1 The Exhibit Approach

Exhibit is a lightweight server-side data publishing framework that allows people with basic HTML knowledge to create Web pages with rich, dynamic visualisation of structured data, and with faceted browsing [6]. To 'exhibit' FlyTED, we needed to create the dataset that could be consumed by Exhibit, and the HTML page to present these data.

Our local harvesting script exports image metadata from the FlyTED repository into N3 format, which we then translate into JSON format[15] using Babel[16]. The initial HTML code was copied from Exhibit's tutorial and then customised based on the feedbacks from our users. Figure 2 shows the first 10 thumbnails of the 38 wild type gene expression images that reveal expressions in both `Mid_primary_spermatocyte` and `Mid_elongation-stage_spermatid`. Clicking on an image caption will take users to a pop-up bubble that displays metadata details about this image, including a live URL of this image record in the FlyTED repository.

5.2 The jSpace Approach

jSpace [7] is a faceted browser that has been inspired by mSpace [5] (which was not available for evaluation at the time of writing). It is a client-side Java Web Start application, and it supports browsing of RDF data presented by SPARQL

[15] `http://www.json.org/`
[16] `http://simile.mit.edu/babel/`

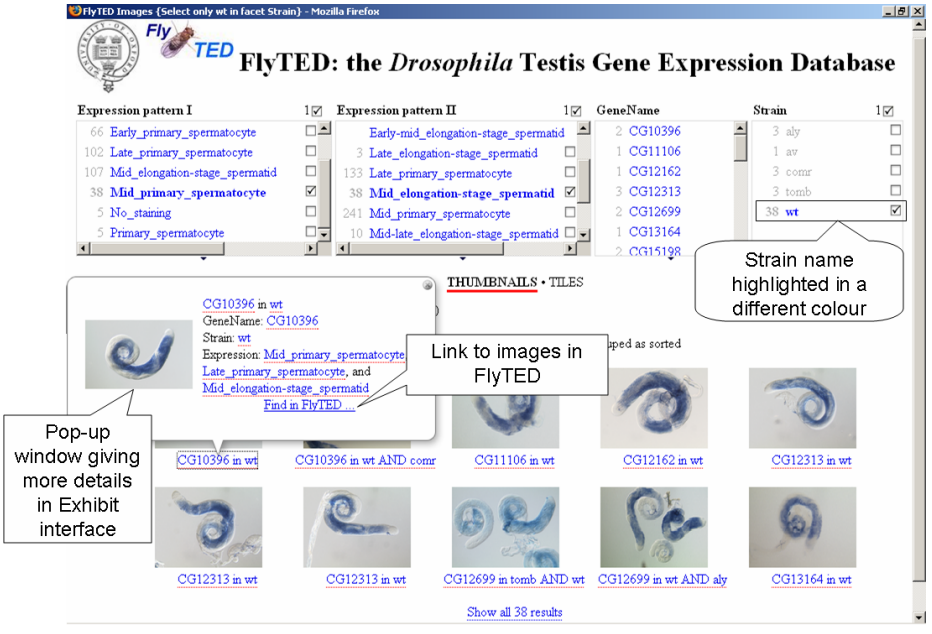


Fig. 2. Presenting FlyTED in Exhibit

endpoints, or kept in local files or Sesame [9] RDF databases. We used jSpace to access images and their domain metadata through the FlyTED Joseki SPARQL endpoint. To present FlyTED in jSpace we needed to:

- Publish our metadata as a SPARQL endpoint, which could then be accessed by jSpace using the SPARQL protocol.
- Create a jSpace model file, which defines the facets used for filtering and presenting images.
- Create a *Web view builder* to load images from the FlyTED repository Web site.

Out of the box, jSpace does not support browsing multimedia data content, but it can load and display Web pages which contain multimedia content using its Web view builder API. By default, this builder takes the string value of a currently selected resource, searches for that term on Google restricted to `site:wikipedia.org`, and navigates to the first hit. For our experiment, we created an alternative customised Web view builder, which takes the URL of a currently selected image and then navigates to the FlyTED repository site using that image's URL. Figure 3 shows jSpace being used to browse gene expression images recorded from wild type flies that reveal expressions in both `Mid_primary_spermatocyte` and `Mid_elongation-stage_spermatid`. Selecting an image URI in the image column brings, in this example, the expression image of gene CG10396 in wild type flies retrieved from FlyTED. Compared to Exhibit, the major limitation of jSpace is that links displayed in the retrieved

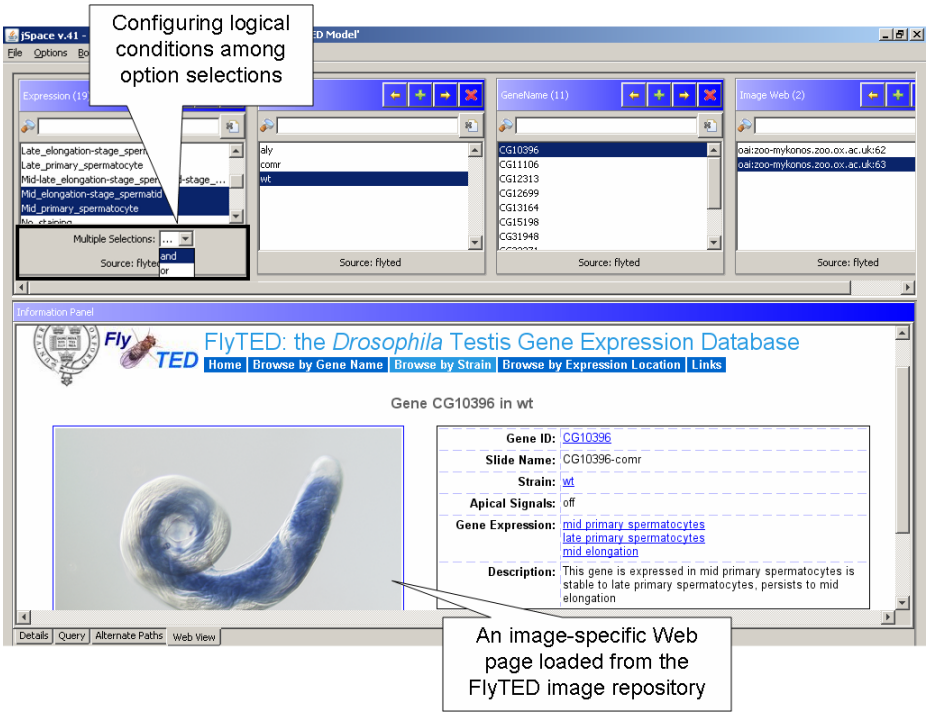


Fig. 3. Presenting FlyTED in jSpace

FlyTED web page are not active and that a view showing thumbnails of all retrieved images side by side is not available. But jSpace has better support for constructing logical conditions between option selections (see Figure 3).

5.3 Functionality Measurement

Table 1 compares the support for our users' image browsing requirements given by EPrints, Exhibit and jSpace. The result shows that the faceted browsers provide more flexible interfaces than the conventional repository interface, which can only support R1-3. However, a list of inadequacies remain to be addressed by the faceted browsers, which include:

- Support for users in the construction of complex conditions between multiple selections, including AND, OR (possible in jSpace) and NOT operators.
- Support for users' customisation of their sorting conditions, for example, presenting the *wild type* strain at the top of the strain categorisation, rather than providing only the *ad hoc* string index in alphabetical order.
- Support for the organisation of facet conditions in a tree structure, so that users can zoom in/out within one category.
- Support for image browsing by presenting a group of images as thumbnails (possible in Exhibit), or presenting an enlarged view of the image when the user hovers the mouse over the thumbnail image.

Table 1. Functionality comparison between EPrints, Exhibit and jSpace

Requirements	EPrints	Exhibit	jSpace
R1: Browse images by the gene name, strain name, or expression pattern keyword, preferably with images presented as thumbnails	Yes	Yes	Partial
R2: Search for all images of a gene that show the same expression pattern	Yes	Yes	Yes
R3: Search for all images of a gene that show the same *set* of expression patterns	Yes	Yes	Yes
R4: Browse images of a single gene, first by the different strains in which this gene was imaged, and then by the expression patterns exhibited in those images	**No**	Partial	Yes
R5: Browse images from a particular strain by the (set of) expression patterns shown in the images	**No**	Partial	Yes
R6: Find all the images of a particular strain NOT showing a certain set of expression patterns	**No**	**No**	**No**

5.4 Performance

Figure 4 shows the performance achieved for loading and browsing images and their metadata in datasets of varying sizes in Exhibit and jSpace respectively. The test was performed using a laptop of 1GB memory. Our goal is to compare how these two tools perform for browsing an image dataset of moderate size on an average personal desktop where most scientists would work on.

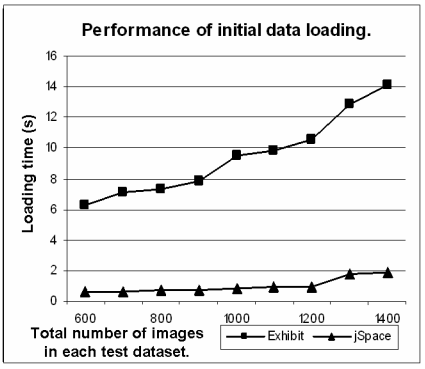
There are currently around 26,000 RDF triples for ∼1500 images in our Jena RDF repository and in the JSON metadata file. The size of our dataset is likely to grow at least three times by the end of the biological image gathering.

The evaluation aimed to discover how these two tools performed for typical image browsing tasks shown in Figure 2 and 3 as the size of data grows. The results show that:1) Exhibit takes on average 10 times longer to load the dataset into a Web browser than jSpace (Figure 4(a)) sending the query and receiving the SPARQL response from the local FlyTED endpoint, and 2) Exhibit takes 2-4 times less time to browse images and their metadata than jSpace (Figure 4(b)).

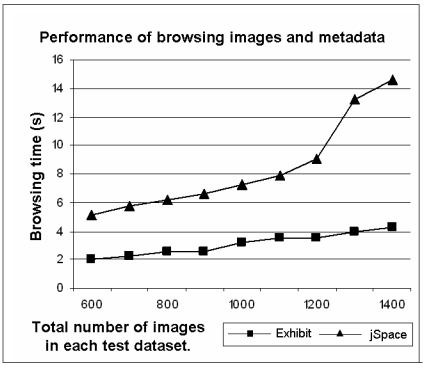
Exhibit loads all the metadata into memory during initialisation. This makes its subsequent image browsing more responsive, but means that coping with large datasets becomes more difficult. The loading time in Exhibit grows linearly with the growth of data and its scalability problem is known to its developers. In jSpace, each selection of an option leads to a HTTP call between the client and the SPARQL endpoint. The scalability of jSpace thus relates to how many triples are returned from the client for each SPARQL query.

5.5 Development Cost

Both tools required no installation and were easy to start up. The existing examples on both tools' web sites provided sufficient information that could be copied and pasted to create either the model files required for jSpace or the



(a) Loading time (LT) of Exhibit and jSpace. LT_Exhibit=Load the JSON data into Web browser; LT_jSpace =Send the query and receive the SPARQL response.



(b) Browsing time (BT) of Exhibit and jSpace. BT_Exhibit=$\sum$response time of the Web Browser for each selection of an option; BT_jSpace =$\sum$ SPARQL response time of each selection.

Fig. 4. The performance comparison between Exhibit and jSpace

HTML pages and JSON data needed for Exhibit. jSpace required moderate extra effort to build the customised Web view builder, but its facet widgets provide more flexibility for users. Total effort required: 3 working days for jSpace and 2 working days for Exhibit.

5.6 Summary

We compared the costs and benefits of two Semantic Web faceted browsers, Exhibit and jSpace, for accessing images in the FlyTED image repository. The experiments showed that these tools provided more flexible user interfaces than does the conventional EPrints repository software. However, it also revealed gaps in the functionality of these tools, which fell short both in satisfying user requirements and in tool performance.

6 Related Work

Semantic digital library systems [10] describe repository resources with rich semantics so that they can be integrated with other resources and consumed by machines. This is closely in line with our goal of building a Semantic Web accessible image repository. Some existing digital libraries systems (such as Jerome DL [11]) also provide a "social semantic web library", which builds a social network among library users to promote sharing of resources and knowledge. However, these systems, compared to existing digital repository systems, are still in their early stage. Their stability and scalability are yet to be tested, and

their components are often tightly coupled. The ability of extending existing digital repository software to provide the extra functionalities of semantic digital libraries is yet to be verified.

Researchers in Southampton have experimented with using mSpace to browse a knowledge repository of heterogenous digital collections, including publication information from their university EPrints repositories and information about researchers from their web sites or funding bodies [12]. Their previous experience encouraged this work.

7 Conclusions

This paper reports our experience in building a Semantic Web accessible image repository by combining an existing repository software package with Semantic Web tools. This approach bridged the gap between conventional repositories and the Semantic Web. The latter provides facilities for disseminating repository resources so that they can be processed along with information from third parties, and for visualising these resources in more flexible user interfaces.

The contributions of this paper are threefold:

- It shows the feasibility of building a Semantic Web accessible image repository using existing tools and simple HTTP-based protocols. This saved us from having to build a repository software system from scratch to achieve the desired functionalities.
- It demonstrates that although existing Semantic Web faceted browsers do provide more flexible user interfaces, they have limitations in supporting a real-world scientific usage. Some of the missing functionalities are likely to be required in different application contexts, such as supporting logical combinations of conditions in one facet; while others are required by the challenges of presenting image data, such as loading multiple images and presenting them side by side.
- It illustrates the significant advantages of employing a lightweight software framework: saving development effort by reusing existing toolkits; and providing the flexibility of replacing components with alternative tools, which increases the sustainability of this framework and enables us to experiment with different approaches.

The provision of the FlyTED repository information in a Semantic Web-accessible form is our first step. We now plan to integrate this information with other related data resources in a data web [13,14], in order to provide a unified platform for scientists to access and integrate these distributed and heterogeneous resources. We also anticipate continuing our experiment with faceted browsers, and thereby to contribute requirements to the browser developers from real scientific studies that involve frequent interactions with image data.

Acknowledgements

The FlyTED Database was developed with funding from the UK's BBSRC (Grant BB/C503903/1, Gene Expression in the *Drosophila* Testis, to Drs Helen White-Cooper and David Shotton). This work was supported by funding from the JISC (Defining Image Access Project to Dr David Shotton; `http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess`), and from BBSRC (Grant `BB/E018068/1`, The FlyData Project: Decision Support and Semantic Organisation of Laboratory Data in *Drosophila* Gene Expression Experiments, to Drs David Shotton and Helen White-Cooper). Help from Michael Grove of jSpace and from the developers of Exhibit is gratefully acknowledged.

References

1. Benson, E., Gudmannsdottir, E., Klyne, G., Shotton, D., White-Cooper, H.: FlyTED - The *Drosophila* Testis Gene Expression Database. In: Proc. of the 20th European Drosophila Research Conference, Vienna, Austria (2007)
2. Lagoze, C., de Sompel, H.V., Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting - Version 2 (June 2002), `http://www.openarchives.org/OAI/openarchivesprotocol.html`
3. Shotton, D., Zhao, J., Klyne, G.: JISC Defining Image Access Project Final Report - Images and Repositories: Present status and future possibilities. Section 12: Project software developments. Project report, University of Oxford (August (2007), `http://imageweb.zoo.ox.ac.uk/pub/2007/DefiningImageAccess/FinalReport/`
4. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF, W3C recommendation (January 2008), `http://www.w3.org/TR/rdf-sparql-query/`
5. schraefel, m.c., Wilson, M.L., Russell, A., Smith, D.A.: mSpace: Improving information access to multimedia domains with multimodal exploratory search. Communication of the ACM 49(4), 47–49 (2006)
6. Huynh, D.F., Karger, D.R., Miller, R.C.: Exhibit: Lightweight structured data publishing. In: Proc. of the 16th International Conference on World Wide Web, Banff, Canada, pp. 737–746 (2007)
7. jSpace: a tool for exploring complex information spaces, `http://clarkparsia.com/jspace/`
8. Fielding, R.T.: Architectural styles and the design of network-based software architectures. ch. 5: Representational state transfer (REST). Ph.D. Thesis, Department of Information and Computer Science, University of California, Irvine (2000)
9. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: a generic architecture for storing and querying RDF and RDF Schema. In: Proc. of the 1st International Semantic Web Conference, Sardinia, Italy, June 2002, pp. 54–68 (2002)
10. Kruk, S.R., Decker, S., Haslhofer, B., Knežević, P., Payette, S., Krafft, D.: Semantic digital libraries (May 2007), `http://wiki.corrib.org/index.php/SemDL/Tutorial/WWW2007`
11. Kruk, S.R., Decker, S., Zieborak, L.: JeromeDL - Adding Semantic Web technologies to digital libraries. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 716–725. Springer, Heidelberg (2005)

12. schraefel, m.c., Smith, D.A., Carr, L.A.: mSpace meets EPrints: a case study in creating dynamic digital collections. Technical report, University of Southampton (January 2006)
13. FlyWeb: Data web for linking laboratory image data with repository publications, http://imageweb.zoo.ox.ac.uk/wiki/index.php/FlyWeb_project
14. Shotton, D.: Data webs for image repositories. In: World Wide Science: Promises, Threats and Realities, Oxford University Press, Oxford (in press, 2008)