PhenoImageShare: An image annotation and query infrastructure

Solomon Adebayo[1], Kenneth M[c]Leod[2], Ilinca Tudose[3], David Osumi-Sutherland[3], Richard Baldock[1], Albert Burger[2] and Helen Parkinson[3]*

[1]MRC Human Genetics Unit, IGMM, University of Edinburgh, Crewe Road, Edinburgh, UK, [2]Department of Computer Science, Heriot-Watt University, Edinburgh, UK, [3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

ABSTRACT

Motivation: High throughput imaging is now available to many groups and it is possible to generate a large quantity of images quickly. PhenoImageShare provides a lightweight image data query and annotation tool, and a single point of access backed by a Solr server for programmatic access to a federated image collection and related image PhIS enabling improved community access.

1 INTRODUCTION

As reference genomes and large-scale programs to generate model organism mutants and knock-outs are completed, there has been matching effort from projects such as the International Mouse Phenotyping Consortium (IMPC) to establish and codify phenotype with genomic coverage (Koscielny *et al.*, 2013). Current phenotyping effort will deliver annotations held in independent databases associated with the primary data, which may be searched individually, but there is no mechanism for integration, cross-query and analysis, especially with respect to human abnormality and disease phenotypes for image derived data. Therefore we have developed PhenoImageShare (PhIS). PhIS is a cross browser, cross repository platform enabling semantic discovery, phenotypic browsing and annotation of federate phenotypic images. Resources such as OMERO (Allan *et al*., 2012) allow users to store their images, but do not provide ontology enabled user facing tools to share and query image PhIS across repositories using standard query engines such as Solr. Further, images are often "siloed" by imaging methodology or domain and access to multiple image repositories requires bespoke development against multiple modes of programmatic access each of which is evolving as each resource changes. PhenoImageShare is species and imaging technology neutral and currently provides access to 94,751 images federated from different data resources with 10,000 regions of interest (ROI) associated to anatomy or phenotype annotations via ontologies. These can be accessed via the GUI beta.phenoimageshare.org or via web services and URL. To date PhIS is populated with images from Drosophila and three different mouse projects. Code is available from https://github.com/PhenoImageShare and https://github.com/ma-tech/PhenoImageShare.

2 METHODS

The PhIS platform consists of three major components. The Image Discovery Infrastructure, consisting of a schema for image related PhIS representation, a Solr index offering queries on the data, and a web service hosted at the EBI which is accessible to the project and also as a service to external users wishing to access image PhIS. The IQS component, run at Heriot-Watt University, provides an integrated query service (IQS) supporting access to spatial reasoning queries integrated with the web services for the Solr index. Finally the user interface components hosted at Edinburgh University provide an intuitive GUI to query cross platform image meta-data and will in future host the annotation service to allow image owners, and third parties to annotate images using ontologies.

2.1 Image Discovery Infrastructure

The query API is available from http://beta.phenoimageshare.org/data/rest/ . It offers a simple REST interface with a single response type (JSON). In future the annotation submission API will be exposed to programmatic access and will require authentication, however, at present we do not expose the image submission functionality to outside access. We currently perform batch submissions from XML export files. Data sources provide us with up-to-date XML exports of their data, following the PhIS schema (XSD). We then perform XML and semantic validation (checking URIs and that term labels match), followed by data enrichment before indexing. This includes addition of synonyms and ancestor relationships to the Solr index allowing an improved semantic search. The implementation is in Java and the query functionality is primarily via Solr. The database, XSD and Solr representations share a common high-level schema. The meta data is split into three types: image annotation, channel annotation and ROI annotation (described in detail below).

- The `image` entity is core and stores generic, immutable metadata. These are roughly encompassed by imaging procedure and details, data source,

* To whom correspondence should be addressed.

credits and sample information, including genotype or sample preparation techniques.
- The `channel` entity encapsulates visualization information, such as tags or labels and genomic information when needed, such as for expression images.
- The ROI entity holds both the coordinates, within the given image, of the ROI(s) as well as the annotation values.

The metadata stored does not overlap so all associated entity types should be queried. For example an ROI entity associating an expression value in liver to a spatial region on the image should be analyzed with the associated image document containing genetic information and with channel entity providing information about the reporter used. The user interface makes the browsing seamless for users across these three categories.

2.2 Spatial Reasoning

Within PhenoImageShare, spatial reasoning is used to find phenotypes via a spatial description e.g. associated with a location or with respect to some other entity from which location can be inferred. For example, given the phenotype "cataract" the anatomical structure "eye" can be inferred. Spatial reasoning can also be used to search for particular combinations or patterns in the data that are spatially *similar.* A range of possible definitions exists for "similar spatial regions" including overlapping and adjacent areas. In order to facilitate this reasoning the images must include direct or inferred from phenotype spatial annotation, e.g. from pulmonary fibrosis the anatomical term lung is inferred; bridge ontologies provided by the Monarch initiative (Köhler et al., 2013) support this inference. When images have appropriate spatial annotations spatial reasoning can identify spatially similar ROIs. Two anatomical terms that have an is-a or part-of relationship are considered to be close. Reasoning is pre-computed with the results stored in Solr for query.

More sophisticated spatial reasoning can be undertaken by mapping the spatial information within the images onto an appropriate atlas that contains a representation of space including direction. Through the spatial model of the atlas a spatial relationship between two ROIs is obtained. For example, ROI *A* is distance *Y* further along the cranial axis than ROI *B.* Measurements of distance and direction present the user with the opportunity to select their own definition of *similar* by simply changing a threshold. The spatial reasoning API will be integrated into future releases via the Integrated Query Service (IQS). The IQS will turn the GUI's query into distinct queries for the spatial reasoning service and Solr. The two responses will be combined in a single result set and sent to the GUI.

2.3 User interface

PhIS delivers its functionality through a highly responsive and configurable web-based interface that allows users to register, search, query and provide phenotypic annotations for image data using standard ontologies. It provides discover, query and subsequently annotate image data held in independent databases. These annotations and associated metadata can be imported at registration and also exported in standard formats such as CSV, RDF, etc. The first release of the PhIS GUI provides annotation discovery and query, multi-word free-text and ontology-based search with autosuggest, facet-based browsing and visual analytics providing a unified view of the distributed image databases. Feature implementation is requirements-driven and is prioritized based on user test sessions. The Model-View-Controller (MVC) design technique has been adopted due to the nature of PhIS data. Portability, re-usability and responsiveness are at the core of Graphical User Interface (GUI) design considerations. The core logic of the web interfaces is developed using Python, Django, JavaScript and AJAX. Bootstrap and Flat-UI technologies are used for the interface theme and styling (Views), and are integrated with the core to support changing of theme dynamically, while allowing application components to be developed as widgets.

Figure 1. Users are presented with options for taking a tour of the website (**a**) or performing a quick free-text search (**b**) or navigating to the search interface (**c**). Clicking through points on the interactive charts launches the search page with imaging method filter (**d**) applied. Data labels can be used to toggle (**e**) the contributions of the data points and charts can be downloaded in various formats (**f**).

Users can explore PhIS from the index page through options ranging from taking a tour of the web application, through performing a quick search or navigating search interface via the application menu, to performing quick filtering/browsing by clicking through points on the interactive charts, which present visual summaries of the data in the repository. Search text can be an anatomy or phenotype term, gene or allele symbol from standard ontologies, or user entered free-text annotation term. Points – or slices – on the pie charts represent canned queries, which map to data filters on the search interface. Clicking through points on the charts launched the search page with the corresponding filters selected, and the matching image data displayed. Users can also interact with the charts by toggling the data

labels in order to visualize the contributions of each data point. This interactivity affords users the facility to quickly drill down to images of interest by filtering out dominating data within a category of interest. The charts can be exported in various formats such as JPG, PNG, and SVG for local use. These features of the index page are shown in Figure 1.
The search interface is the main entry point for users to browse through PhIS repository, through the use of facets or quick search feature. In addition to support for autosuggest, the search interface is responsiveness and search results are displayed in real-time.

Text-based search and faceting
Users can perform quick search for a gene or allele (using their symbols), anatomical or phenotypic terms from ontologies or free-text using the generic search box and autosuggest provides users with a drop-down list of terms predicted from their text entries. Search results are displayed in real-time as users enter their search text. Users can further narrow down their search by applying filters for taxon, anatomy, imaging method etc.

Browsing and data filtering with facets
Facets and sub-facets allow users to browse through PhIS dataset without having to type in text in the generic search box. This is useful for users with limited prior knowledge of what to search for. They can expand facets and check filters (checkboxes) - e.g. Mutant-Expression applicable to their search. Applied checkbox filters are displayed as tags above the facets. Users can click on the tag(s) to unselect the corresponding facet(s), and the updated records. Users can also perform ontology-based facetted search for gene or allele symbol, anatomy or phenotype term or a combination of these. Figure 2 shows the search functionality of PhIS web application.

2.3.1. Detail & Annotation Interface

When a user finds an image of interest, clicking on it takes them to the detail. Genotype information and image metadata are presented on this page along with ROI and annotations associated with the selected image. Annotations associated with the image or ROI within the image are displayed. These can be downloaded from the interface. At present genotype information is captured at image registration, whereas phenotypic annotations are captured either at registration or through the user facing annotation submission interface to be released in the next version of the application.

3 DATA AND ONTOLOGIES

The current PhIS release contains data from two sources, which cover a variety of image types (X-ray, macro, histopathology slides, and lacZ expression). 1. The TRACER database (Chen et al., 2013) contains mouse data embryos carrying a transgenic insertion generated by the Sleeping Beauty transposon-based system. Embryo expression images are recorded and annotated with a controlled vocabulary for anatomy parts and are therefore suitable for inclusion in PhenoImageShare. 2. Images from the Wellcome Trust Sanger Institute represent single gene knock-outs and are of types: X-ray, histopathology or gene expression images using lacZ reporter. Genotype, anatomy and phenotype annotations are provided for these data. Further data sources will be included in the next release, and will include Drosophila data from Virtual Fly Brain and mouse data from EMAGE.

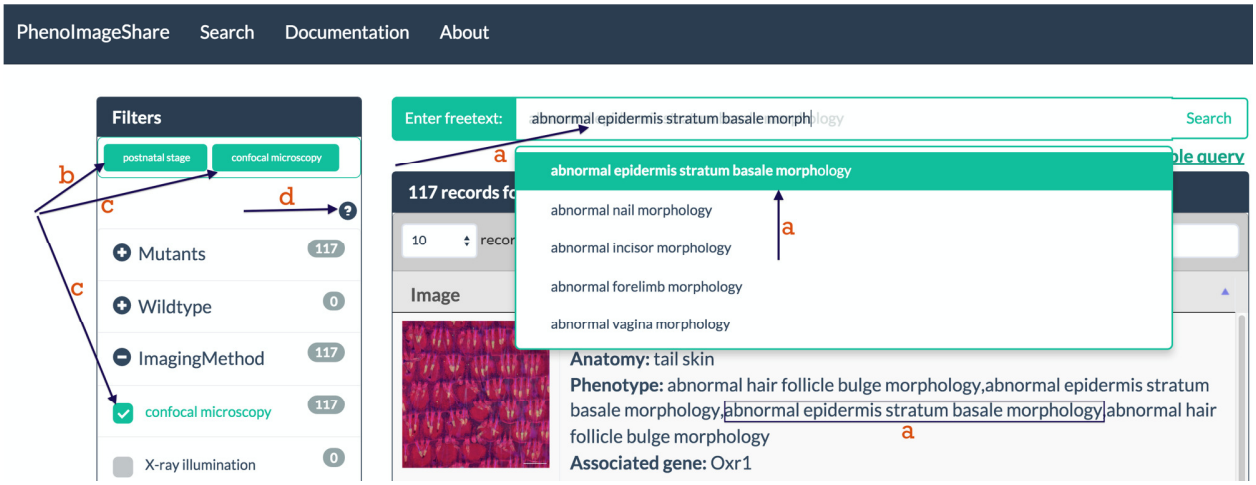

Figure 2. Query interface with ontology-based search and faceting. Results of auto-suggest for abnormal epidermis stratum basal morphology (**a**) filtered by development stage (**postnatal**) (**b**) and confocal imaging method (**c**). Example query and help on using facets (**d**) are also provided on the interface.

To add semantic value and make the integration possible we support the use different ontologies e.g. by species and application area and bridge between ontologies. For anatomy annotations we use the MA, EMAP, EMAPA and ontologies and for cross species integration UBERON (Mungall, Torniai, Gkoutos, Lewis, & Haendel, 2012) and the Monarch Initiative's bridge files. We also have phenotype currently from the MP (Smith & Eppig, 2012) and we expect to extend these to pathology, and other species in future releases. The species specific developmental stage ontologies and the Fbbi ontology for biological imaging methods, originally developed for the Cell Image Library (Orloff, Iwasa, Martone, Ellisman, & Kane, 2013) are critical for the development of PhenoImageShare. In order to accommodate data from different resources free text is also permitted, although the annotation interface will encourage the use of ontology terms through tailored auto-suggest and tools such as Zooma (http://www.ebi.ac.uk/fgpt/zooma/), which can suggest ontology terms based on the free text description entered.

4 CONCLUSION

The PhIS platform provides underlying infrastructure for informatics users and user facing tools for biologists enabling the query and annotation of federated images. It supports the use of ontologies and builds on these to deliver spatial reasoning. The releases of an ontology enabled annotation tool for image generators and third parties will allow projects such as IMPC (Koscielny et al., 2013) to federate image annotation tasks and will provide a collaboration platform for annotators. Features to be supported in future releases include query sharing and bookmarking, grid-style image gallery view, image discovery by similarities, complex query builder, drawings and annotations of ROIs on images, cross-species search, integration and mappings, spatial reasoning, links to reference atlas frameworks such as EMAP (Baldock et al., 2003), support for 3D image annotation, browsing and annotation of image collections, and query storage and analytics. The next data release includes EMAGE database and the Virtual Fly Brain (Milyaev et al., 2012) (VFB). The VFB dataset offers an interesting use case for cross species integration, but also adds value to our anatomy coverage with its focus on neuroanatomy and gene expression in the adult Drosophila melanogaster brain. Future infrastructural plans include offering the PhIS services as portable widgets such as those offered by BioJS (Gómez et al., 2013) for inclusion in third party tools. We encourage owners and generators of image datasets to expose their data via PhIS and view it as sustainable platform for the sharing of image data which requires minimal investment in disk, and supports a federated model of image data sharing.

ACKNOWLEDGEMENTS

PhenoImageShare is funded by BBSRC BBR grant BB/K020153/1 and EMBL Core funds to HP. Thanks to Jason Swedlow for useful discussions; Yiya Yang for export of Mouse Atlas data; Tony Burdett, Jeremy Mason and Gautier Koscielny for technical guidance; and Matt Pierce for assistance with Solr.

REFERENCES

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., … Swedlow, J. R. (2012). OMERO: flexible, model-driven data management for experimental biology. *Nature Methods*, *9*(3), 245–53. doi:10.1038/nmeth.1896

Baldock, R. A., Bard, J. B. L., Burger, A., Burton, N., Christiansen, J., Feng, G., … Davidson, D. R. (2003). EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics*, *1*(4), 309–25. doi:10.1385/NI:1:4:309

Chen, C.-K., Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Smedley, D., & Spitz, F. (2013). TRACER: a resource to study the regulatory architecture of the mouse genome. *BMC Genomics*, *14*(1), 215. doi:10.1186/1471-2164-14-215

Gómez, J., García, L. J., Salazar, G. A., Villaveces, J., Gore, S., García, A., … Jiménez, R. C. (2013). BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics (Oxford, England)*, *29*(8), 1103–4. doi:10.1093/bioinformatics/btt100

Köhler, S., Doelken, S. C., Ruef, B. J., Bauer, S., Washington, N., Westerfield, M., … Mungall, C. J. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, *2*, 30. doi:10.12688/f1000research.2-30.v2

Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T. F., Morgan, H., Atienza-Herrero, J., … Parkinson, H. (2013). The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research*. doi:10.1093/nar/gkt977

Milyaev, N., Osumi-Sutherland, D., Reeve, S., Burton, N., Baldock, R. A., & Armstrong, J. D. (2012). The Virtual Fly Brain browser and query interface. *Bioinformatics (Oxford, England)*, *28*(3), 411–5. doi:10.1093/bioinformatics/btr677

Mungall, C. J., Torniai, C., Gkoutos, G. V, Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, *13*(1), R5. doi:10.1186/gb-2012-13-1-r5

Orloff, D. N., Iwasa, J. H., Martone, M. E., Ellisman, M. H., & Kane, C. M. (2013). The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Research*, *41*(Database issue), D1241–50. doi:10.1093/nar/gks1257

Smith, C. L., & Eppig, J. T. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome*, *23*(9-10), 653–668. doi:10.1007/s00335-012-9421-3