

Towards Metadata Patterns in Image Databases

Joris Slob, Amalia Kallergi, Julia Dmitrieva, Fons J. Verbeek

LIACS, Leiden University, the Netherlands

Abstract. Pattern recognition in image databases has focused on homogeneous images from single or limited number of modalities. Biological problems require a more integrated approach that crosses scales and modalities. We present a database architecture that focuses on image metadata enriched with ontology term annotations. A use-case within the research field of Alzheimer’s disease shows how such a database architecture can be a valuable source of investigation and exploration.

Keywords: Biological databases, integration, visualisation, ontology, annotation

1 Introduction

Biological research centers are struggling to organize their image repositories. In practice researchers keep their microscopy images on local storage devices and only a handful are made available for the community. There are benefits to sharing these images after paper publications have been made about the research: the ability to look up the cited experimental data for reexamination, for educational purposes or for inspiration.

When inspecting biological images there are certain factors we have to take into account. If we examine biological samples *in vivo*, we cannot disregard the context of the sample. To understand the function of a particular biological system, it is important to have a clear view of both the details of its parts and the way they combine on a larger scale. For example, to appreciate the importance of a particular pathway it is necessary to examine its implications on different scales. It might be impossible to view these different scales in the same modality.

We present a way to organise and share microscopy images in such a way that researchers can look at available resources in the context of their experiments and that allows different fields of biology to benefit from their data. Concretely, we are developing a system that deals with images from different modalities, with different number of dimensions and on different scales. We take scale as a way to relate data, but we do not limit ourselves to scale as the only connection between different images.

2 Materials and Methods

To create a system that connects images from different sources, we have constructed a database which can model the wide variety of images. In our approach

we focused more on the metadata of images than the actual pixel/voxel content. To understand the content of the images we apply semantic annotation from multiple ontologies. We present a way to unify the concepts by defining an ontological context. We used data visualization techniques to allow the users to browse through the database in a more natural explorative way.

2.1 System Architecture

Our system can be described as a set of subsystems which work together to create a repository for microscopy images. We have described our system in previous papers[1][2][3], but give a small overview here.

At the heart of our system, we have an SQL database that keeps track of all the metadata of the images. A second SQL database keeps track of all the ontology terms that are available for annotation. These terms are extracted from the OBO Foundry[4]. Communication with these two SQL databases go through a SOAP interface. In the current implementation these interfaces are only available internally.

We use Apache Tomcat as our servlet container to provide services for our web user interface. This interface provides easy access for researchers to upload, annotate and explore images. We also expose a SOAP endpoint with restricted access to the data for third party developers. The Cyttron Visualization Platform¹ uses this endpoint.

2.2 Data versus metadata

While traditional pattern recognition techniques mainly focus on the actual pixel data and the derived features, we focus mainly on the metadata of the images. Without the image metadata it would be hard and in some cases impossible to process the images, for example images where specific biomarkers were used. Without metadata that tells us to what these biomarks bind it is impossible to fully interpret the images. Our main aim is to support researchers in browsing the available data and discovering unforeseen relations. The metadata includes:

- which user created the image
- which research group the user is from
- the time and date the image was submitted
- the microscope settings used to acquire the image
- the ontology terms used to describe the content of the image
- if the image is included in a dataset (created by a user)

We have chosen for semantic annotation to ensure that image owners share a common vocabulary of terms that have a richer structure than normal controlled vocabularies. There are over 90 available ontologies In the OBO Foundry, out of which we have incorporated 50 in our database. Choosing terms from this list confused users.

¹ <http://graphics/tudelft.nl/cvp>

2.3 Ontological context

Although ontologies describe different domains, it could be possible that domains overlap. A concept of interest can occur in different ontologies and can be described from different points of view. This gives us the possibility to reuse this overlapped information to create a new ontology about a particular concept of interest. In our application we have used the approach [5] where an *ontological context* is created around the concept of interest. The concept of interest in our use-case was 'Alzheimer'. First, we have extracted similar terms from NCI THESAURUS ontology, because this ontology contains broad information about diseases and related genes. The extracted concepts were used as *seed terms* in order to trigger the search process in other ontologies, e.g. PATHWAY, GO, MESH. This gives us the set of terms which is referred to as the *global ontological context*. For each concept from the *global ontological context* we extract a module/graph from their original ontologies based on the hierarchy and other relationships between concepts, e.g. *part_of*, *develops_from*, *Chemical_Or_Drug_Affects_Gene_Product*, et cetera. The graph is generated till some level of depth, in our case we have used 3 as a limit. The subgraphs extracted from original ontologies are referred to as *local ontological context*. The subgraphs combined in one structure can be considered as an integrated ontology dedicated to the term of interest.

2.4 Interface

In our web interface, we make a conscious effort to accommodate and motivate our users in their interaction with the data. We believe that data accumulated in the repository will require not only powerful algorithms for mining but also effective displays and interactive tools to examine and explore this data. To this end, ideas from information visualization, a well-established field exploiting the visual capacities of humans to reinforce the understanding of data, are worth investigating. Note that in the context of interacting with life science repositories, we are interested in the visualization of the metadata of the images, and not of the raw data.

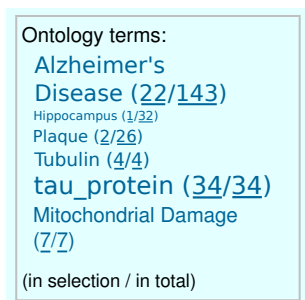


Fig. 1. Ontology term cloud

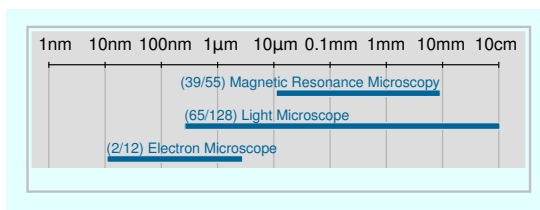


Fig. 2. Modality Graph

Simple information visualization elements are introduced to facilitate user interaction with the database. Consider e.g. the ontology viewer[6] that exemplifies ontologies using a graph visualization. More recently, attention has been paid on the visualization of search results: The standard result list (in thumbnail or table view) has been updated with two panels, the ontology term cloud and the modality graph. The intention is to provide a better overview of the search results, in a faceted-like display, and a means for the user to browse the data. First, the ontology term cloud (see Fig. 1) is a list of all ontology annotations appearing in the given result set, presented in the popular tag cloud metaphor. This term cloud functions as a combination of a tag cloud and a drill cloud², allowing to either refine the current search (drill cloud-like interaction) or start a new search (tag cloud-like). Number of hits per option is indicated and may provide a perspective on the usage of terms across the database. Second, the modality graph (see Fig. 2) is a dynamic graphic of the resolution scale supported in the database. The graph displays all modalities in the given result set drawn at their approximate resolving ranges and allows queries on modality in a fashion similar to the term cloud. By means of simple interactive and visual elements, these panels provide a richer context for a given search. Moreover, they invite the user to examine the results (and the collection) as a whole rather than as an enumeration of individual entries. Currently, we are considering more elaborate information displays to invite our users to connect and relate data by means of visual inspection and interactivity.

3 Results

As a test-case we have asked researchers in the field of Alzheimer’s Disease to enter and annotate their images in our database. We will describe how three users (Researcher A,B,C) from our database entered annotated data and how a fourth user can browse through the results (see Fig. 3).

Researcher A, studied the effects of tau proteins on mitochondria in Chinese hamster ovary cells. He submitted images he acquired with fluorescence microscopy and annotated them with terms like *tau protein*, *mitochondrial damage* and *tubulin*.

Researcher B, studied plaque formation with a bright field and confocal microscope. She annotated her images with terms like *Alzheimer’s disease*, *plaque*, *tau protein* and *hippocampus*.

Researcher C, studied the progression of Alzheimer’s Disease patients over multiple years with magnetic resonance microscopy. He annotated his images with terms like *Alzheimer’s disease* and *hippocampus*.

Researcher A did not label his images with *Alzheimer’s disease*, because in this research they didn’t look at brain cells. Researcher B did annotate the images with *Alzheimer’s disease* and *tau protein* thus forming a link between the two keywords, that did not have to exist before. Researcher C did not label his images with *tau protein* because at this scale it would not make much sense.

² <http://lab.cisti-icist.nrc-cnrc.gc.ca/cistilabswiki/index.php/Drill.Clouds>

A user can start looking for images annotated with the term *Alzheimer's disease*. He will find the images of researcher B and C. Now, the user wants to explore the concepts relating to *Alzheimer's disease*. In the ontology term cloud he sees that some of the images are co-annotated with the *tau protein* term. He can now choose to narrow his search for images that have both tags (like the images from researcher B, or search for images that are annotated with the term *tau protein*, but not necessarily with *Alzheimer's disease*, like the images from researcher A.

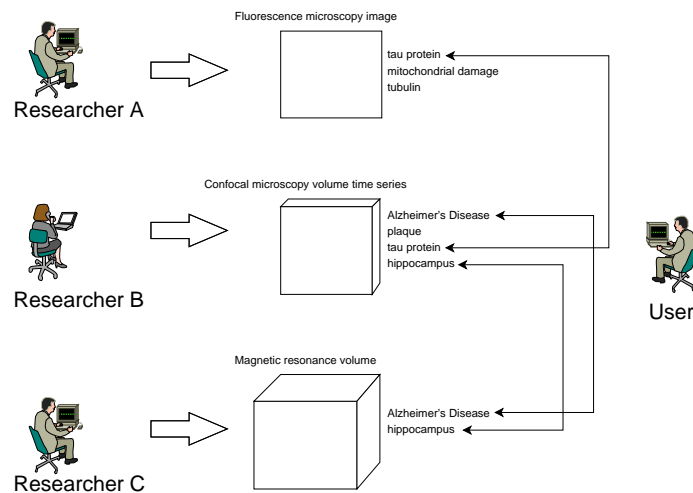


Fig. 3. test-case: Researchers A,B and C contribute their images to our database and annotate them using ontology terms. Another user can now browse from image to image because the annotations co-occur.

4 Discussion and Conclusions

We have seen that images and their annotations form a virtual path in our database that can be followed by any user. This is an example where human researchers have annotated their images in such a way to create a new way to browse the available data. Even with limited images in our database at this time, we observe how relations can emerge out of multiple users interacting with the system.

Modality plays an important part in our database, because it is also a measure for the scale of the images. The ontology terms transcend modality and scale and form relations across multiple fields of biology. Nature doesn't conform to the scales of our equipment, so information systems should find a way to remove the artificial boundaries that modalities impose on our perception of our samples. Our system bridges these boundaries and connects images from different sources.

We plan to replace our SQL database implementation of the available ontology terms with a triplestore and test if we can gain more information by applying different reasoning strategies.

Future work would include mixing metadata and pixel/feature data into a feature vector and using that to find new patterns and using the inherent structure of ontologies to help making new relations.

References

1. Bei, Y., Belmamoune, M., Verbeek, F.J.: Ontology and image semantics in multi-modal imaging: submission and retrieval. In: SPIE Internet Imaging VII. Volume 6061. (2006)
2. Bei, Y., Dmitrieva, J., Belmamoune, M., Verbeek, F.J.: Ontology driven image search engine. In: SPIE, MultiMedia Content Access: Algorithms & Systems. Volume 6506.
3. Kallergi, A., Bei, Y., Kok, P., Dijkstra, J., Abrahams, J.P., Verbeek, F.J.: Cyttron: A virtualized microscope supporting image integration and knowledge discovery. In Backendorf, Noteborn, T., ed.: Cell Death and Disease Series: Proteins Killing Tumour Cells. ResearchSignPost (2009) 291–315
4. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* **25**(11) (2007) 1251–1255
5. Dmitrieva, J., Bei, Y., Verbeek, F.J.: Ontological context visualization. In Golbreich, C., Kalyanpur, A., Parsia, B., eds.: OWLED. Volume 258 of CEUR Workshop Proceedings., CEUR-WS.org (2007)
6. Kallergi, A., Bei, Y., Verbeek, F.J.: The ontology viewer: Facilitating image annotation with ontology terms in the csidx imaging database. In: VISS-WS. (2009)