SYNTHESYS

Synthesis of systematic resources

Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS3
Grant Agreement number:	312253
Workpackage:	4: Moving from physical to digital collections
Deliverable number:	4.1
Deliverable title:	Edge detection technology
Deliverable author(s):	Lawrence Hudson, Laurence Livermore and Vladimir Blagoderov, Natural History Museum London (NHM).
Deliverable contributor(s):	Vladimir Blagoderov (NHM) Alice Heaton (NHM) Pieter Holtzhausen (Stellenbosch University, South Africa) Lawrence Hudson (NHM) Laurence Livermore (NHM) Ben Price (NHM) Vince Smith (NHM) Stefan van der Walt (Stellenbosch University, South Africa)
Date:	16th December 2014

SYNTHESYS
Synthesis of systematic resources

Developing Edge Detection Technology for Natural History Images

Background

SYNTHESYS3 is a European Union-funded Integrated Activities grant which aims to create an accessible, integrated European resource for researchers in the natural sciences. The Joint Research Activity (JRA) is one of its three main activities and aims to improve the quality of and increase access to digital collections and data within natural history institutions' virtual collections.

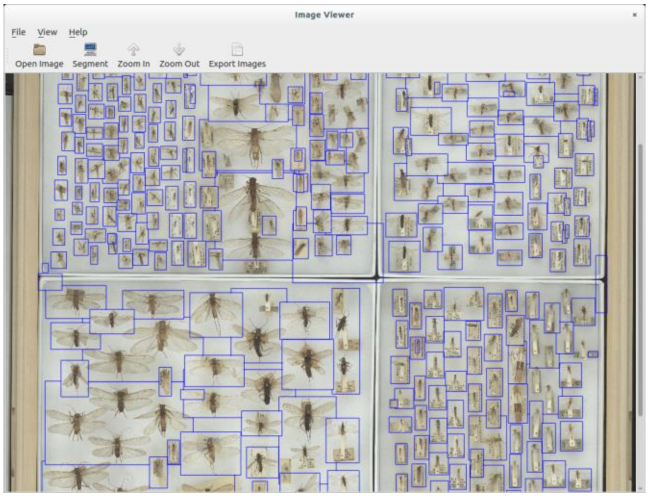
One of the JRA objectives (4.1) is to support and develop technology that automates data collection from digital images. As part of the NHM's (Natural History Museum London) contribution to this objective we have developed open source software that can recognise, process and annotate images that contain multiple specimens (e.g. whole drawer scans of pinned insects or slide arrays). A workshop was held in September 2014 to develop a specification and produce a functional software prototype (**Inselect**). Following the workshop this prototype was presented and demonstrated at the Taxonomic Database Working Group (TDWG) meeting in October 2014. Subsequent development has focused on creating a release candidate for wider testing and feedback from other SYNTHESYS3 Taxonomic Access Facilities (TAFs).

Initial Software Specification

A number of whole drawer scanning technologies currently exist but annotation and processing of images is the limiting step (Blagoderov et al. 2012). Typically collections want to record metadata about individual specimens which involves cropping specimens and entering metadata such as taxonomic name, location in collection etc.

Most software currently in use has limitations:

- No automation – all specimens have to be manually selected by a human operator;
- Poor user interface and user experience – most software has a very basic user interface and provides a poor user experience;
- Closed source and proprietary – making it costly or impossible for museums to collaborate, build upon and share software;
- Not cross-platform – limiting annotation and post-processing to Windows users only.



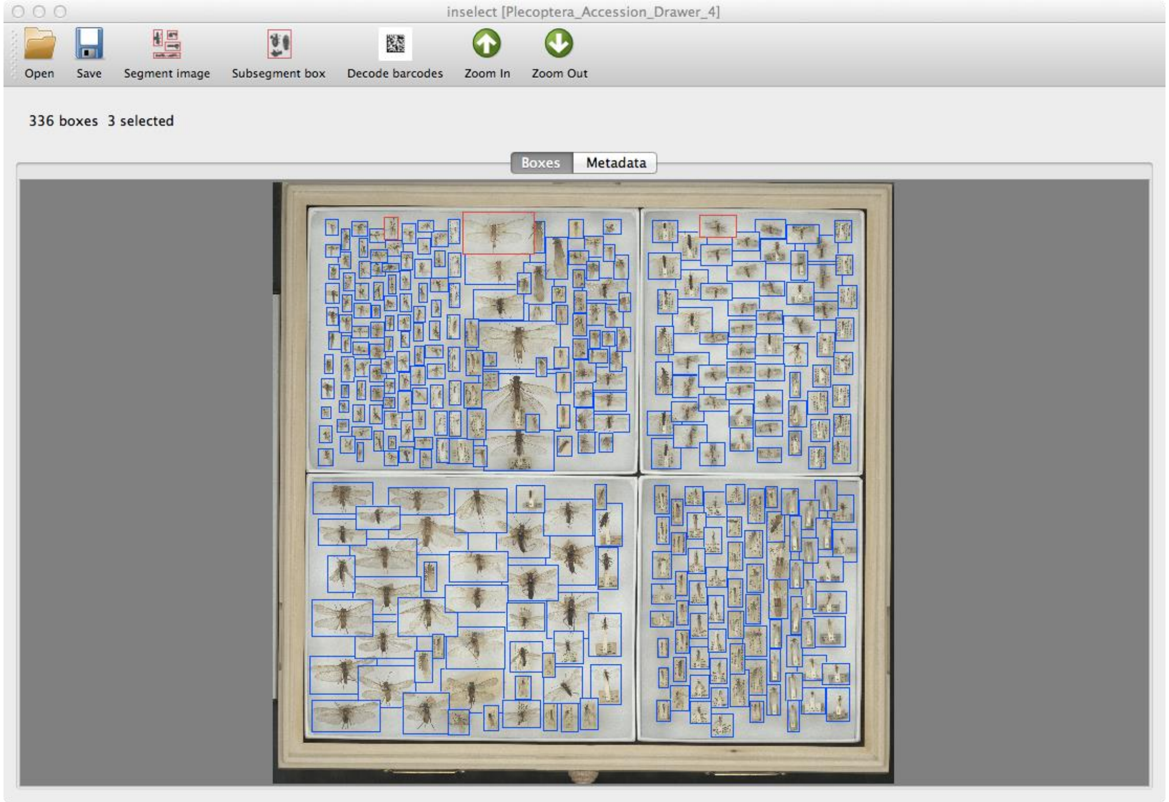
Screenshot of Inselect prototype showing automated recognition of individual specimens with blue bounding boxes.

The prototype was developed to solve limitations in existing software based on the following specifications:

- Open source, freely available and cross platform (works on Windows, OS X and Linux);
- Allows manual creation and editing of bounding boxes around specimens for cropping, tagging and saving individual specimen images;
- Segmentation of images into regions of interest (insects/slides) should be automated, with the ability to refine manually and edit the outputs of the automated process;
- Segmentation algorithms should be modular (easily extend to different regions of interest in the future);
- Modularity, in order to add extra functionality easily, for example integrating barcode reading technology in order to set metadata with the value(s) of barcode(s) within the cropped specimen image;
- Cropped specimen images should be saved at full available resolution;
- Possibility to batch process of images (by directory);
- Ability to annotate specimens individually/in bulk and export metadata;
- A design that considers both a single user working with a small number of scanned images and an organisation working with high-throughput digitisation of large collections;
- Extensively tested in realistic workflows by collections/digitisation staff.

Current Software Description

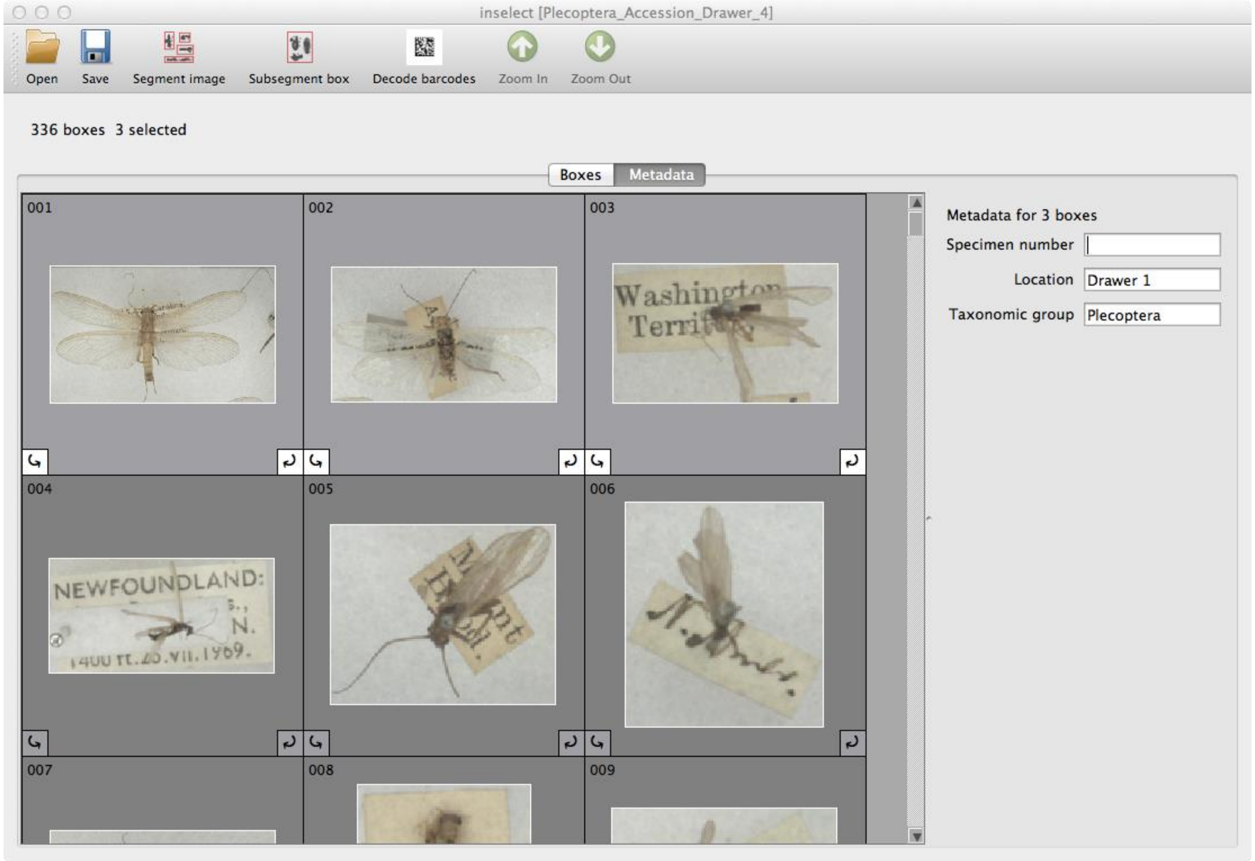
The proof-of-concept prototype developed in late 2014 has been re-engineered to provide the modular architecture required by the specification. In this architecture, Inselect maintains a document, which is made up of original full-resolution scanned image, a lower-resolution thumbnail suitable for display and a collection of bounding boxes together with their metadata. Separate to this, Inselect contains a number of views which determine how the user interacts with different elements of the document. The ‘Boxes’ view shows the locations of the bounding boxes around each specimen, and allows boxes to be created, deleted, moved and resized:



336 boxes 3 selected

The ‘Boxes’ view, showing bounding boxes around 336 specimens. Three boxes are selected (shown in red).

The ‘Metadata’ view shows specimen images in a grid:



336 boxes 3 selected

Metadata for 3 boxes

Specimen number

Location: Drawer 1

Taxonomic group: Plecoptera

The 'Metadata' view.

The user can rotate specimen images and, using the controls on the right of the view, can edit their metadata.

This model-view design (Gamma et al 1994) separates the way that data is stored from the way that it is presented to the user, making it relatively straightforward to add new ways of viewing and editing Inselect's data without changing the underlying data structures.

Inselect also provides a system of plugins - code modules that are allowed to examine and possibly modify the document's bounding boxes and associated metadata. Plugins can be operated on any part of the document - the full-resolution scanned image, the lower-resolution display thumbnail and/or the bounding boxes. Inselect currently has plugins for automated segmentation, sub-segmentation and for reading barcodes. New functionality can be added to Inselect simply by adding a new plugin, for example alternative segmentation algorithms.

Dissemination

Research Presentation: SPNHC (Society for the Preservation of Natural History Collections) 2014 (June 2014)
Vladimir Blagoderov, Laurence Livermore, Ben Price, Stéfan van der Walt. Image segmentation in high throughput digitisation workflows.
http://prezi.com/doomtbgkmqig/?utm_campaign=share&utm_medium=copy&rc=ex0share

Research Presentation: SYNTHESYS3 Annual General Meeting (September 2014)
Livermore, Laurence; Blagoderov, Vladimir; Heaton, Alice; Holtzhausen, Pieter; Hudson, Lawrence; Price, Ben; Walt, Stéfan van der (2014): Applied Edge Detection in Zoological Collections: Inselect Prototype. figshare.
http://dx.doi.org/10.6084/m9.figshare.1165492

Research Presentation: TDWG (October 2014)
Holtzhausen, Pieter; Walt, Stéfan van der; Heaton, Alice; Livermore, Laurence; Blagoderov, Vladimir; Price, Ben; Hudson, Lawrence; **Smith, Vincent** (2014): Moving beyond the box: automating the digitisation of insect collections. Slideshare.
http://www.slideshare.net/vsmithuk/moving-beyond-the-box-automating-the-digitisation-of-insect-collections

Future Development

The following functionality is planned but not yet implemented:

- User-interface refinement
- User-interface for plugins, for example, tuning parameters of segmentation algorithms
- Server-side tools for batch ingestion, segmentation, barcode reading etc.
- Unit tests
- Order of segments
- User-defined metadata fields
- Localisation
- Allow user to define trays
- Undo
- Multiple monitor support(?)
- Make improvements based on user testing
- Find and fix bugs

A list of issues is maintained at https://github.com/NaturalHistoryMuseum/inselect/issues.

Release plan
First alpha release in Q1 2015.

Testing
Extensive testing on a wide range of specimen types, using NHM's collection of SatScan images.

- (Potentially) Formalise a measure of segmentation algorithm performance
- User testing of Inselect
- User testing of workflow tools

Gather statistics on several aspects of performance:
- Time to ingest (i.e. creation of low-resolution thumbnail)
- Time to run automated segmentation
- Time to refine results of automated segmentation

Collaboration
- Design home page with documentation
- Advertise first release

Links

Inselect code repository: https://github.com/NaturalHistoryMuseum/inselect

Inselect project page: http://naturalhistorymuseum.github.io/inselect

References

Gamma, Helm, Johnson, and Vlissides. 1994. Design Patterns - Elements of Reusable Object-Oriented Software, ISBN 0-201-63361-2.

Blagoderov, Vladimir A., Ian Kitching, Laurence Livermore, Thomas Simonsen, and Vincent Smith. "No Specimen Left behind: Industrial Scale Digitization of Natural History Collections." ZooKeys 209 (July 20, 2012): 133–46. doi:10.3897/zookeys.209.3178.