Initial plan - Image Analysis for Museum Insect Drawers

- **Author** Louise Evans (1332639)
- **Supervisor** Paul L Rosin
- **Moderator** Xianfang Sun

- **Module Number** CM3203
- **Module Title** One Semester Individual Project
- **Credits Due** 40 credits

Project Description

Museums have vast numbers of specimens including animal and mineral varieties. Often large quantities of specimens are stored for a range of purposes together with corresponding data used to identify them. The specimens found at museums are stored in standardised trays which are organised and labelled with respect to their genus and species. The entomology department comprises the collection and identification of insects.

Cardiff Museum has a particularly large entomology collection of which they have been exploring the idea of digitising. They would like to create a data archive of the insect specimens within Cardiff Museum to make available for the public to access online. With their vast quantity of insect trays and specimens this would be extremely difficult to do manually; therefore a digitised process of extracting metadata from the insect trays would be an ideal solution to this problem.

For my project task I would like to explore the feasibility of data extraction of museum insect trays using image analysis. Cardiff Museum has provided resources including high definition insect tray images and entomology data in order to help with this project. Additionally there are examples of previous projects which may contain useful ideas of how to tackle this task (Schmidt, Balke, and Lafogler, 2017).

This process of digitisation would begin by investigating optical character recognition of insect labels in order to extract label text from the insect drawer image. This text could be transformed into XML metadata based on the requirements given by Cardiff Museum and outputted to an external file.

I could expand this analysis by determining the layout of different elements within the trays, such as dividing boxes, bugs and labels. I could explore the segmentation of trays using the approximate shape and dimensions of the boxes used to organise the specimens. I would also need to develop a method of detecting labels and insects within these containers. This insect detection could be used to count the number of each species contained within the tray as further information to include within the XML data.

If time permits, I could carry out further analysis such as categorising a label as a genus or a species based on the label's physical features. This identification could be used to interpret which specimens are identified by each label. A collection of insects will have a species label and a collection of species will have a genus label. This task would require an understanding of the image layout and how each of these relationships is defined within the tray's structure.

I could also look deeper into the appearance of insects within the tray, analysing an insect's physical features and comparing specimens within a species. I could additionally develop a verification method for the data extracted from the images using the museum's entomology data or the GBIF Species API (Gbif.org, 2017).

Project Aims and Objectives

During my project I would like to complete the following core objectives:

- Research and implement suitable character recognition, image segmentation and/or edge detection algorithms.
- Investigate text extraction including:
    - Finding a method of detecting text regions within the test images.
    - Applying optical character recognition in order to extract text from identified label regions.
    - Transform extracted text into XML format using the example XML provided by Cardiff Museum.
- Investigate insect tray segmentation including:
    - Segmenting an insect tray into organisation boxes used to arrange the trays uniformly.
    - Detecting regions of interest within a tray which may contain specimens or labels.
    - Detecting empty boxes within trays.
    - Extraction of useful image regions for deeper analysis.
- Investigating more refined analysis including:
    - Detecting each collection of insects within a tray container.
    - Detecting label regions within a tray container.

If time permits I could additionally:

- Continue further insect analysis including:
    - Detecting an individual insect within a collection of insects.
    - Producing a count of the quantity of insects in each species.
    - Analysing insect features to detect anomalies within a collection of insects.
- Continue label structure analysis including:
    - Categorising labels as a genus or a species by analysing label properties such as font size and relative location.
    - Interpreting which labels apply to each set of specimens based on tray structure.
- Improve metadata extraction including:
    - Editing the XML formatting based on further understanding of the image layout.
    - Analysing or improving text accuracy using external data integration using the museum's entomology spreadsheet data or the GBIF Species API.

Work Plan

Throughout the development process I will be meeting with my supervisor each week at 11:00am on Fridays to discuss progress towards deliverables and the overall project development.

The following plan outlines what I aim to complete each week before the final project submission date 5/5/2017.

1) **Research**: I will begin by researching image processing methods and their uses in image analysis. This includes considering existing solutions to similar problems and how they may affect my solution. Selecting suitable image analysis methods is necessary to ensure they are relevant to the requirements for my project. (Week 1)
2) **Experimentation**: I will use my research to begin implementation of algorithms identified during the previous week. These algorithms should be thoroughly tested to verify their practicality for this project. (Weeks 2-3)
3) **Implementation**: My program implementation will begin with detection of text regions within insect drawer test images. This will require testing and improvement of the text extraction method developed in the previous weeks. The extracted text should ideally have minimal errors when compared with expected output from manual interpretation. The text can then be used to output XML data using the requirements specified by Cardiff Museum. (Weeks 4-5)
4) **Implementation**: The next stage requires high level segmentation of insect trays into smaller organisation box regions using previously investigated techniques. This task may be simplified with known data such as background colour or dimensions. I should also be able to identify empty space or features of interest. (Weeks 6-7)
5) **Implementation**: Deeper analysis of non-background regions can be developed within detected regions of interest. I will investigate how to identify types of regions based on region features. I should be able to identify the areas of a test image which contain a label or collection of specimens with reasonable accuracy. (Weeks 8-9)
6) **Additional Development**: Further development of desirable tasks can be continued where time permits. The selection of the final project tasks may be affected by progress made so far with the project, therefore this time should be spent developing and expanding the functionality implemented in previous tasks. (Weeks 10-13)
7) **Finalisation**: Make any final edits and refinements to the report or final program in preparation for project submission. (Week 14)

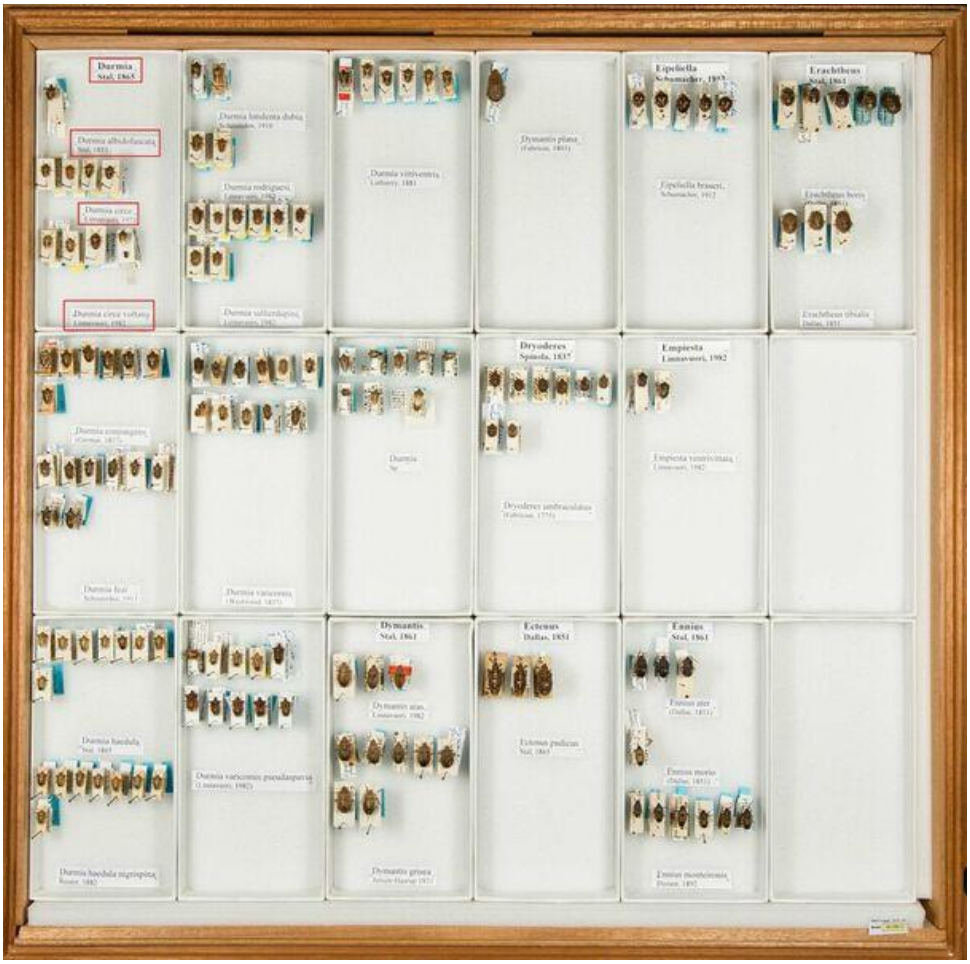
Task	**Week Commencing**													
	Jan 30th	Feb 6th	Feb 13th	Feb 20th	Feb 27th	Mar 6th	Mar 13th	Mar 20th	Mar 27th	Apr 3rd	Apr 10th	Apr 17th	Apr 24th	May 1st
1)														
2)														
3)														
4)														
5)														
6)														
7)														

Additional Resources

- Links to related systems/competitions:
  - http://zookeys.pensoft.net/articles.php?id=2916
  - http://zookeys.pensoft.net/articles.php?id=2913
  - https://beyondthebox.aibs.org/overview.html
  - https://naturalhistorymuseum.github.io/inselect/
- Global Biodiversity Information Facility (GBIF) - Species API:
  - http://www.gbif.org/developer/species
- Example output XML formatting:

```
<?xml version="1.0" encoding="UTF-8"?>
<DataSet>
<Unit> <!-- this is created per drawer image -->
<UnitID>0018610</UnitID> <!-- from the image filename eg. 0018610.jpg -->
<CaptureDate> dd/mm/yyyy </CaptureDate> <!-- from image metadata -->
<!-- creates a keyword from each label-->
<keyword>Durmia Stahl, 1865</keyword>
<keyword>Durmia albidofuscata Stahl, 1853</keyword>
<keyword>Durmia circe Linnivuori, 1973</keyword>
<keyword>Durmia circe voltana Linnivuori, 1982</keyword>
</Unit>
</DataSet>
```

- Example insect tray image:



References

Beyondthebox.aibs.org. (2017). *Beyond The Box*. [online] Available at: https://beyondthebox.aibs.org/overview.html [Accessed 27 Jan. 2017].

Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T. and Smith, V. (2017). *No specimen left behind: industrial scale digitization of natural history collections*.

Gbif.org. (2017). *Species API*. [online] Available at: http://www.gbif.org/developer/species [Accessed 27 Jan. 2017].

Schmidt, S., Balke, M. and Lafogler, S. (2017). *DScan – a high-performance digital scanning system for entomological collections*.

Trustees of the Natural History Museum, L. (2017). *Inselect*. [online] Inselect. Available at: https://naturalhistorymuseum.github.io/inselect/ [Accessed 27 Jan. 2017].