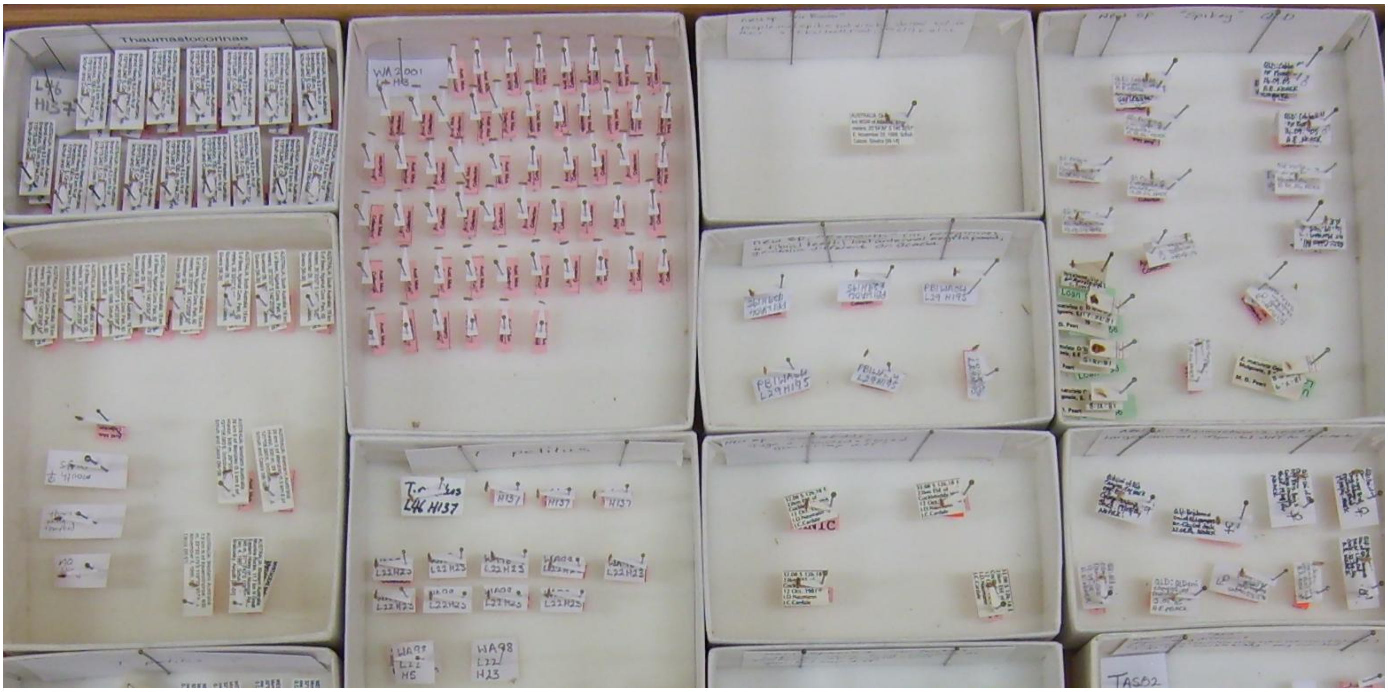


Australian Museum
Data capture of specimen labels using volunteers

John Tann & Paul Flemons

December 2008

Australian museum
nature culture discover

Australian Museum

Data capture of specimen labels using volunteers

Summary

This is a report of an attempt to speed up the capture of information on the labels of specimens held by the Australian Museum.

A trial was conducted using volunteers with a camera to photograph specimen labels and transcribe that data into a spreadsheet. Location information was georeferenced. The data in the resulting spreadsheet was then entered into EMu by a museum technician. Times and costs were compared to direct data entry, as well as with a previous trial using an off-shore data transcription service.

The outcome of the trial was successful in clarifying the following:

Importing data into EMu is not straightforward and is a specialist task. Having the data transcribed into a spreadsheet before import into EMu does not help. Errors, misspellings, and uncertainties on many of the labels meant that a spreadsheet of data became a clumsy and inefficient method of data entry.

Photographing a label has advantages – a photograph becomes a verbatim record in the database of the label for later referral, and makes the data entry process quicker by about 20%, as well as easier and more convenient.

Recommendations
The Australian museum could train and use a small team of volunteers to photograph specimen labels. These photographs would be saved on EMu as a record of the label, and subsequently used for data entry by AM technical staff.

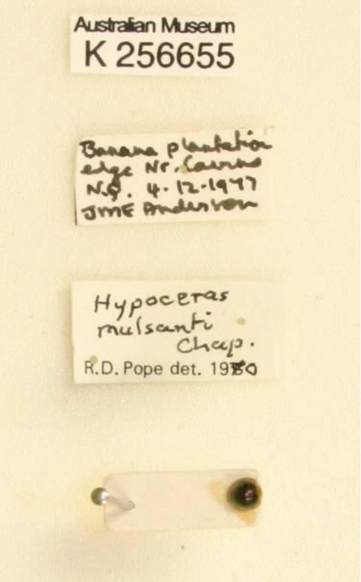
Investigate the EMu inline toolset as a possible route for engaging volunteers for accurate and reliable data entry.

John Tann & Paul Flemons
Australian Museum
December 2008

Contents

Data Capture of specimen labels using volunteers

This is an analysis of a trial held at the Australian Museum using volunteers to capture with a camera, the information on specimen labels in the insect collection. In this trial, volunteers took photographs of the labels attached to a selection of insect specimens. For a comparison, they transcribed label data into a spreadsheet both directly from the label and also from a photograph of the label. Location information was converted to latitude and longitude, and the data was then prepared and imported into EMu.



A breakdown of the strengths, weaknesses, costs and time for each step in the process is given. Comparisons are made with direct data entry using museum technical staff, as well as with a previous trial outsourcing data transcription.

Aim

To assess the potential of using volunteers to assist with capturing data on museum specimen labels and adding that data to EMu, the Australian Museum collection management database.

Background

Traditionally, collections within museums have been an ordered array of specimens. Those specimens would have a label attached and be physically arranged according to a well understood system. With the advent of computers the management of collections changed, and information associated with each specimen was made accessible through a database. Much of this specimen information such as species name, date and place of collection is also very useful to others – planners, natural resource managers, ecologists, field naturalists, quarantine services, and others working with biodiversity.

For more than twenty years the Australian Museum has been taking the information written in registers and on labels attached to specimens and entering it into their databases. This work is ongoing. There are now about 1.5 million records entered into EMu, the AM database. There are still many to go. For example AM holds approximately 800,000 insect specimens, and the information of about 150,000 of these has been added to EMu.

Capturing the information on specimen labels is largely manual work. Labels are often hand-written, small (insect labels are generally smaller than a postage stamp), and commonly quite old. There are difficulties with non-standard abbreviations, legibility, misspellings, ambiguous date formats, obscure locations, changing species names and mistakes. Labels are not written in a standard format, interpretation is often required, and an understanding of the collection process is important.

The capture process is slow, expensive and boring. Mostly it is the task of trained technicians working in a dedicated space. For example, with insect labels, each specimen needs to be carefully removed from its drawer, its label stack disassembled, and the information on these labels transcribed into the relevant fields in EMu. The label stack is then re-assembled and returned to the drawer. Any complication – such as information interpretation, database difficulties or minor maintenance – is dealt with at the time.

Some methods have been investigated in the past to speed up the data capture process. In 2002, KPMG produced a business case for digitisation of collections records for the Australian Museum[1]. In 2007, as a trial, a set of photographs of specimen labels were sent off-shore for transcription; data from the resulting spreadsheet was then entered locally into EMu.

The Australian Museum has strong volunteer support. Currently there are about 150 volunteers working both in front-of-house and behind-the-scenes. In the collections area, volunteers are available and willing to work on specialist tasks and, with training, can handle delicate specimens with appropriate care.

In the past, AM has made little use of volunteers working at home. There are many internet-based activities where people freely give their time for projects such as Wikipedia, open-source software, and proofreading public domain e-books. As an example of a cultural institution making good use of at-home community support, the National Library of Australia recently opened their trial newspaper digitising project for the community to comment and correct the machine-generated text. After four months they have 1,400 text correctors in their user community, having corrected over a million lines of text in 60,000 articles[2].

Method

In order to make best use of volunteers, the process of data capture was broken into discrete components. The tasks were:

- Photography of labels of insect specimens
- Transcription of information from the label photographs to a spreadsheet
- Transcription of information directly from the label to a spreadsheet for comparison
- Georeferencing of locations
- Import of data into EMu using AM technical staff

A comparison of the above techniques was done with:

- Using a transcription service off-shore
- Direct data entry using AM technical staff

Each step of the process was evaluated in terms of time, costs, use of resources, and potential for damage to the collection.

[1] KPMG report *Digitisation of Collections Records, Business Case*, Australian Museum 2002.
[2] ANDP-announce newsletter 27 November 2008. See also http://www.nla.gov.au/ndp/

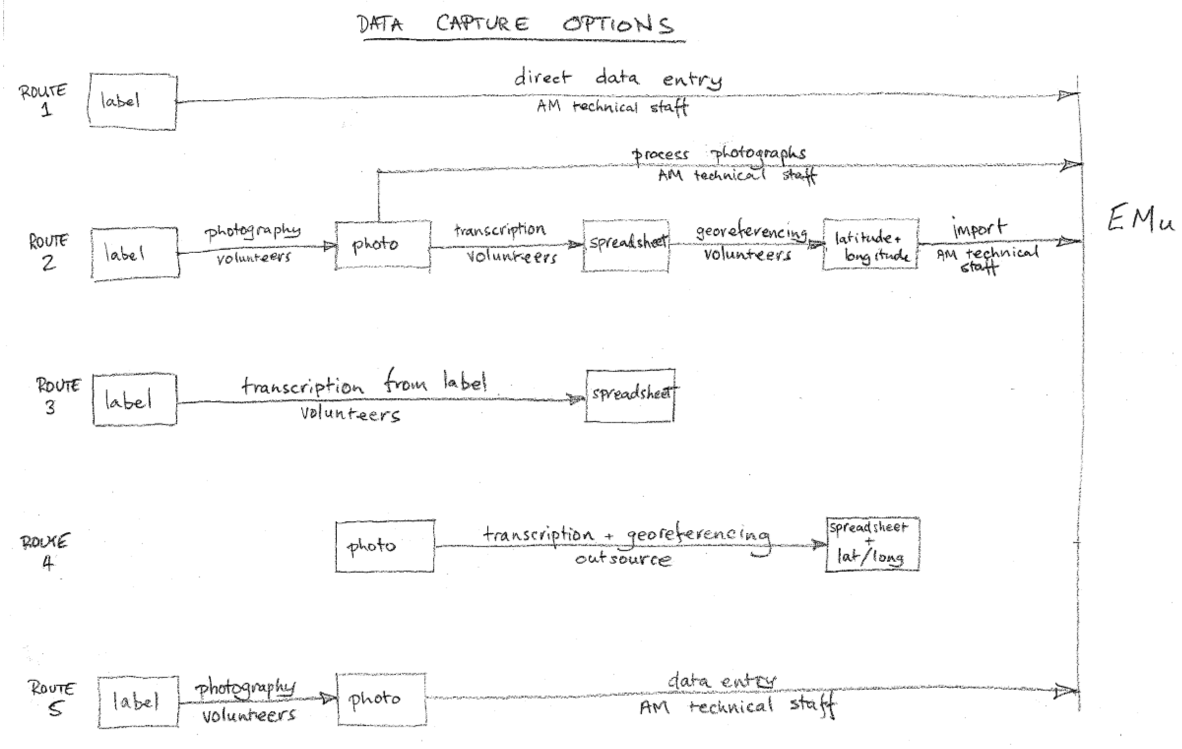
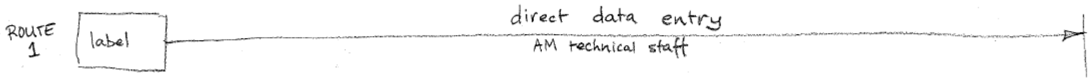


Figure 1. Possible options to capture data from specimen labels. Route 1 is the current way of entering data from the specimen label directly into EMu. Route 2 shows the data capture process broken into tasks to maximise the effectiveness of volunteers. Route 3 shows the process of volunteers entering data from the specimen label directly into a spreadsheet, without a photograph. Route 4 can be done off-shore by a professional transcription and georeferencing service. Route 5 shows a potential hybrid technique using volunteers to take a photograph, and technical staff to enter data directly into EMu from that photograph.

Route 1. Direct data entry



Data is extracted from specimen labels and entered directly into EMu.

Strengths

- Work is carried out in-house.
- Experienced and dedicated personnel. Minimal supervision required. Minimal chance of damage to specimens.
- Regular and contained work – eg, today work only on this drawer of specimens, and when that is finished, move onto the next one.
- Minimal set-up costs, no specialised software maintenance costs, no volunteer management costs

Weaknesses

- Expensive use of technician. This work needs to be done by an EMu-qualified person
- Boring
- Slow – at current estimates there is 50 man-years work to enter the data for the insect collection into EMu[3]
- No record of label for later reference. To transcribe the label again, for example as a check, is the same process as doing it once, with its similar costs and dangers
- The small size of the label leads to legibility problems
- The location for doing this work is constrained by where the specimens are located

Resources and costs

- Bench space
- Computer
- $4.55 per record.[4] (Other institutions range from $2.43 to $13.50 per record)[5]

Time

	hourly rate[6]	**quantity per day**[7]	**time to enter 1000 records**
Direct data entry using AM technical staff[8]	11	55	91 hours

[3] At 55 specimens per day

[4] AM figures, based on fulltime TO doing only databasing, calculated by total number of records entered, divided by cost of TO

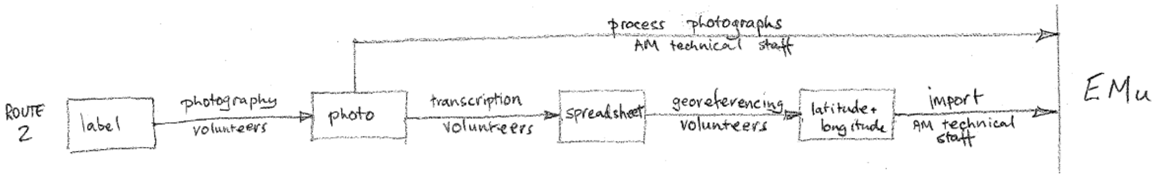
[5] Email Dave Britton, 13 Oct 2008. Databasing costs of QLD DPIF, MV, UQIC, NZAC, AM.

[6] KPMG Appendix A assumed 7 hours per day, whereas I assume five hours on-the-job. The hourly rate is shown to correlate with Table 3.2: entering data for 192,857 specimens in 15 man-years.

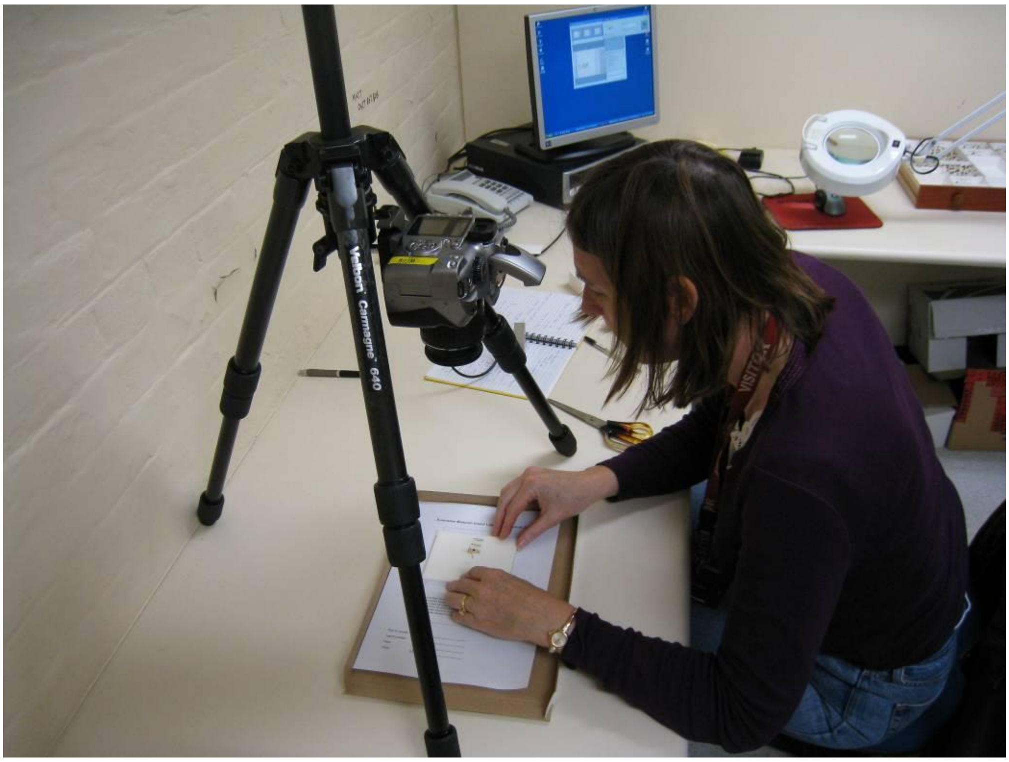
[7] KPMG Table 3.2 and Appendix A suggest 3 staff x 5 years x 240 days/year = 192,857 specimens, ie 55 specimens per day.

[8] Derived from KPMG report *Digitisation of Collections Records, Business Case*, Australian Museum 2002

Route 2. Volunteer tasks

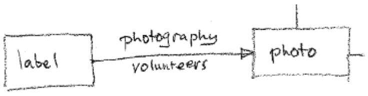


In order to make the best use of the efforts of volunteers, the data capture process is broken up into discrete tasks – photography of labels, transcription of photographs and georeferencing of locations. The semi-automatic task of processing photographs and the task of importing data into EMu is done by museum technical staff.



Volunteer at a data capture station. Each set of specimen labels is photographed separately, and the captured images are transferred directly to the computer.

Photography of specimen labels



This task involves using a camera connected to a computer to take a photograph of the registration number, the specimen labels and the specimen itself. The labels need to be carefully removed from their pins, arranged, photographed, re-assembled and returned to their tray without damage or mix-up.

Strengths

- The photographic process is reasonably fast. Although labels need to be removed from their specimen pins and then re-assembled, over a hundred photos can be readily captured in a half-day session.
- The photograph becomes a record of the label, available on EMu for later referral. Without a photo, checking original labels can be time consuming and awkward.
- A simple image of the specimen is also captured, allowing for preliminary species verification and if a scale is included then rudimentary measurements can be made.
- A photograph that includes the specimen may potentially act as a surrogate for when the original specimen is on loan.

Weaknesses

- Potential for damage due to handling by semi-experienced volunteers[9]
- Potential for label mix-up[10]
- Boredom burnout[11]

Resources and costs

Set-up and run an in-house volunteer photography program.

The equipment and resources required for the photographic process need not be elaborate. Essential elements are:

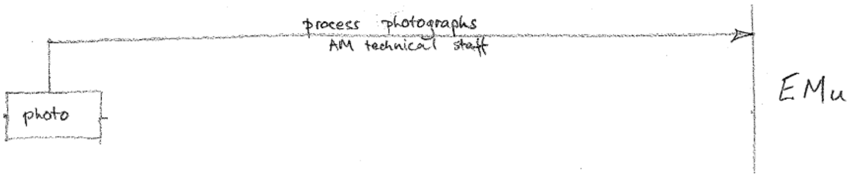
- Several metres of bench space
- Camera
- Computer

Direct set-up costs $2689

Time

	hourly rate[12]	**quantity per day[13]**	**time to photograph 1000 specimens**
Photography of labels using volunteers[14]	50	250	20 hours
Supervision		1 hour	4 hours

Processing of photographs



This task involves using a script to automatically extract the registration number from the image, and use that number to create the filename. A list of unique and identifiable filenames of images is created. These images are then imported into EMu.

[9] In our trial 3% of specimens needed maintenance (re-gluing) or were damaged – broken insect leg, elytra

[10] Good processes can keep this to a minimum. Only one instance of label mix-up detected in our trial of 300 labels.

[11] Boredom may not be such a problem with volunteers, who often work one day a week, compared with someone who may be employed full-time.

[12] Best rate after several days experience.

[13] Assume five hours on-the-job per day.

[14] These times are based on measurements made of volunteers who had little experience. The overarching emphasis was on care and accuracy, not speed.

Strengths

- Semi-automatic process to extract specimen number from photograph – minimal interaction is required
- EMu quickly has a record (the photograph of the label) associated with an issued specimen number

Weaknesses

- OCR methods are imperfect. Human assistance is required for computer processing

Resources and costs

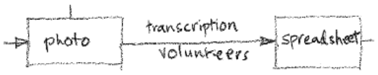
Photographs need to be processed and stored

- Once-off cost to convert proof-of-concept script to an operational program
- Space on database (3 GB per thousand specimens)
  Note that the quantity of space required on EMu is not considered to be a problem. However there are continuing concerns about the growing size of the EMu database and its impact on performance.

Time

			time to process 1000 images
Run batch process[15]			2 hours
Import photographs into EMu[16]			2 hours

Transcription from photographs



This task involves transcribing all the information on each specimen label into a spreadsheet. Fields are available in the spreadsheet for registration number, taxonomy, collector, location, date, identifications, and notes. Some interpretation is required

Strengths

A choice can be made where to transcribe the photograph:

1. **In-house**. Transcription can be done within the museum by volunteers. This has an advantage of having direct access to local knowledge. The workflow can be also be readily monitored.
   Transcription can take place at any computer.
2. **Off-site**. Given appropriate infrastructure and software, transcription can be carried out by volunteers from home. A volunteer needs only a computer, web browser and an internet connection, and can work in their own time at their own pace. This allows the museum to engage with a greater pool of volunteers, without having to accommodate them within the building.

Transcribing from a photograph has advantages

- The resolution of each photo enables the smallest print to appear large on a screen.

[15] Estimate based on 4 batches of 250 images, 30 minutes per batch.

[16] Estimate based on 4 batches of 250 images, 30 minutes per batch.

- With the size of a label being about half the size of a postage stamp, legibility problems and non-standard abbreviations are common. Queries and uncertainties can readily be discussed with, and passed on to, others.
- The text in the image can be transcribed multiple times for comparison, greatly improving accuracy.

When working with volunteers, accuracy and speed comes from experience.

- Expertise in this field comes through experience. Using a smaller number of dedicated volunteers is preferable to a larger number of people with little or no experience.
- Locations, dates, collectors, and handwriting will become more familiar with exposure and, with time, will lead to faster and more accurate transcription with fewer uncertainties.
- If larger numbers of volunteers are available, they can be used for cross-checking other entries.

Weaknesses

- Interpretation is necessary. Guesses and uncertainties will produce errors
- Whichever method of transcription is chosen, some organisation will be required
- Workflow needs to be maintained. There must always be work available for volunteers.

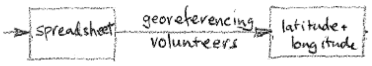
Resources and costs

1. In-house
    - computer use
    - desk space
    - supervision of volunteers
2. Off-site
    - once-off cost to develop software to enable transcription to be done over the internet
    - Web-based computer access
    - supervision of numbers of volunteers, work-flow management
    - disk space for non-EMu database: 10 MB per 1000 records

Time

	hourly rate[17]	**quantity per day**[18]	**time to transcribe 1000 photographs**
Supervision of transcription volunteers[19]		1 hour	7 hours
Transcription from photograph[20]	27	135	37 hours

Georeferencing from spreadsheet



This task involves finding the latitude and longitude for a given position. Gazetteers and mapping software like Google Earth are used, and the results are entered back into the same spreadsheet.

[17] Best rate after several days experience.

[18] Assume five hours on-the-job per day.

[19] Estimate of help needed with new volunteers, maintaining work flow and interpreting difficult labels

[20] These times are based on measurements made of volunteers who had little experience. The overarching emphasis was on care and accuracy, not speed.

Strengths

Batch processes are applicable – eg BioGeomancer

Similarly to the transcription process, georeferencing can be carried out by volunteers either in-house or off-site.

1. **In-house**. Georeferencing can be done within the museum by volunteers. This has an advantage of being directly accessible to local knowledge. The workflow can be also be readily monitored.
2. **Off-site**. Given appropriate infrastructure and software, georeferencing can be carried out by volunteers from home. A volunteer needs only a computer, web browser and an internet connection, and can work in their own time at their own pace. This allows the museum to engage with a greater pool of volunteers, without having to accommodate them within the building.

Accuracy and speed comes from experience.

- Expertise in this field comes through experience. Using a smaller number of dedicated volunteers is preferable to a larger number of people with little or no experience.
- Locations become more familiar with exposure and, with time, will lead to faster and more accurate transcription with fewer uncertainties.
- If larger numbers of volunteers are available, they can be used for cross-checking other entries.

Weaknesses

- Batch processes are only semi-automatic and significant human interpretation is required
- Ambiguities and obscurities seriously hamper georeferencing – often more details of the collection event are needed
- Different people may interpret the same location in different ways, leading to errors and unwanted duplication

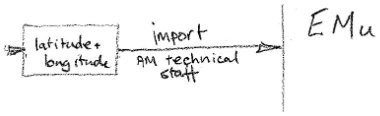
Resources and costs

- **In-house**. Desk space and computer
- **Off-site**. One-off cost to develop software to enable georeferencing by volunteers
- Supplementary support is required for interpretation difficulties

Time

	hourly rate[21]	**quantity per day**[22]	**time to georeference 1000 locations**
Georeferencing - manual	60	300	17 hours
Interpretation support[23]			5 hours

Importing data from spreadsheet into EMu



[21] Best rate after several days experience.

[22] Assume five hours on-the-job per day.

[23] Support level needed if 30% of locations cause difficulties. This time is a rough estimate, and because these are the difficult interpretations, they may take a disproportionate amount of time.

This step involves significant preparation of the data to ensure compatibility with EMu. The data in each field of the spreadsheet needs to be closely checked against the existing contents of EMu, to ensure duplications and bad data are kept to a minimum. Once corrections are done and new events are created, the data can be entered.

Strengths

Time spent here prevents duplication of data in EMu

Weaknesses

- This work needs to be done by an EMu-qualified person.
- Adding bulk data into EMu is not trivial – taxonomy, locations, collectors, dates, collection events – all have to be individually checked and entered

Resources and costs

- Desk space and computer

Time

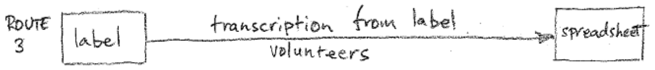
	hourly rate	**quantity per day**	**time to prepare 1000 records**
Preparing data for import into EMu[24]			60 hours

Comment

The complexity of data and level of checking required prevent this task from readily becoming an automated process. For example, with the test set of data, only 11% of the specimens had a determination with a currently accepted name. Multiple spellings, misspellings and synonymies need to be checked carefully to avoid errors. Many assumptions need to be made when deciding what data actually is entered into the EMu database, and it is these assumptions which could not easily be recreated by a machine.

[24] Estimate from sample size of 300 taking about 20 hours.

Route 3. Transcription directly from specimen label to spreadsheet



Volunteers transcribe each specimen label directly – without taking a photograph. This task is compared to the two-step process above where volunteers transcribe the information on each specimen label from the photograph of that label.

Strengths

- Camera equipment not required
- Traditional method of operation. Photographs have not been used in this way in the past.

Weaknesses

- Potential for damage due to handling by semi-experienced volunteers
- Potential for label mix-up
- Boredom burnout
- No record of label for later reference. To transcribe the label again, for example as a check, is the same process as doing it once with similar costs and dangers
- The small size of the label leads to legibility problems, and non-standard abbreviations are common
- The location for doing this work is constrained by where the specimens will be located.

Resources and costs

The equipment and resources required for direct transcription from label to spreadsheet is minimal.

Essential elements are:

- Several metres of bench space
- Computer

Time

Single-step process	**hourly rate**[25]	**quantity per day**[26]	**time to transcribe 1000 labels**
Transcription directly from label	18	90	55 hours
Error-checking[27]	2		

Comparing this direct transcription process with the two-step process – photography of specimen labels followed by transcription from the photographs:

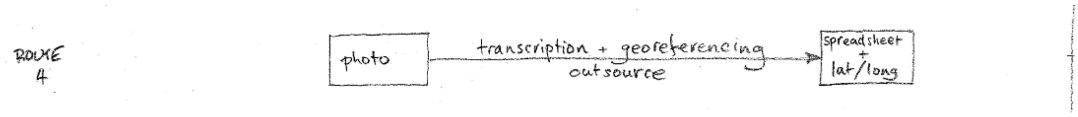
Two-step process	**time to transcribe 1000 labels**
Step 1 – Photography of labels using volunteers[28]	20 hours
Step 2 – Transcription from photograph[29]	37 hours
Total time of Step 1 plus Step 2	57 hours

[25] Best rate after several days experience.
[26] Assume five hours on-the-job per day.
[27] Estimate of time to find specimen, check label, and replace
[28] From **Photography of specimen labels** above
[29] From **Transcription from photographs** above

If photography of specimen labels is done as a process, then the time taken to photograph a specimen label, *plus* transcribe the label from the photograph into a spreadsheet, is very similar to the time taken to transcribe directly from a specimen label into a spreadsheet.

For both processes, errors with transcription were common – scientific names were easily misspelt, and locations were often obscure and unknown. However, when transcribing directly from the label, the consequences of errors is more serious. Many errors could be fixable by instruction – eg punctuation such as commas, abbreviations such as Is., Hd, and F (female). Misspellings may be detectable with dictionaries. However, if there is only one pass at transcription, errors such as missing words, incorrect translations eg male symbol transcribed as Female, and wrong dates may go unnoticed. Poor handwriting would exacerbate these errors.

Route 4. Outsourcing – transcription and georeferencing



The Australian Museum carried out a trial in 2007, where photographs of labels were sent off-shore and transcribed cheaply and accurately.

Strengths

- a known quality and price
- few overheads in terms of managing staff or volunteers

Weaknesses

- some organisation will be required
- quality concerns
- follow-up of errors

Resources and costs

- \$0.40 per record, \$400 per 1000 records[30]
- liaison with outsourcing agents, quality control

Time

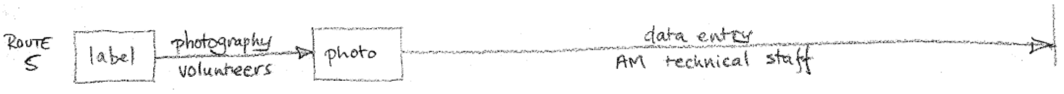
	hourly rate	**quantity per day**	**time to manage 1000 records**
Liaison with outsourcing agents[31]			4 hours
Quality control and correction follow-up[32]	100		10 hours

[30] Doug Rogan, International Conservation Services, email 15 September 2008. March 2007 price was US\$0.30

[31] Estimate

[32] Estimate assume 5% follow-up rate

Hybrid Route – photography of specimen label using volunteers, data entry using museum staff



This method combines the strengths of two tasks - using volunteers to take photographs of specimen labels, and using museum staff to enter the label information from those photographs, directly into EMu.

Strengths

The photographic process using volunteers has advantages

- The photographic process is reasonably fast. Although labels need to be removed from their specimen pins and then re-assembled, over a hundred photos can be readily captured in a half-day session.
- The photograph becomes a record of the label, available on EMu for later referral. Without a photo, checking original labels can be time consuming and awkward.
- A simple image of the specimen is also captured, allowing for preliminary species verification and if a scale is included then rudimentary measurements can be made.
- A photograph that includes the specimen may potentially act as a surrogate for when the original specimen is on loan.

Entering data into EMu from a photograph with museum technical staff has advantages

- Work is done in-house. This data entry into EMu can be done at any desk with a computer.
- Experienced and dedicated personnel. Minimal supervision required.
- The resolution of each photo enables the smallest print to appear large on a screen.
- With the size of a label being about half the size of a postage stamp, legibility problems and non-standard abbreviations are common. Queries and uncertainties can readily be discussed with, and passed on to, others.

Entering data from a photograph is faster than entering it directly from a label. The time spent taking photographs is approximately the same as the time saved entering the data from the photograph.[33]

Weaknesses

- Potential for damage due to handling by semi-experienced volunteers
- Potential for label mix-up in the photographic process
- Slow
- Boredom burnout
- Data capture now becomes a two-part process

Resources and costs

Set-up and run an in-house volunteer photography program.

The equipment and resources required for the photographic process need not be elaborate.

Essential elements are:

- Several metres of bench space
- Camera

[33] See comparison in **Route 3. Transcription directly from specimen label** to spreadsheet above

- Computer

Direct set-up costs $2689

- Staff labour costs approx $3.60 per record.[34]

Photographs need to be processed and stored

- Once-off cost to convert proof-of-concept script to an operational program
- Space on database (3 GB per thousand specimens)

Time

	hourly rate	**quantity per day[35]**	**time to capture 1000 specimens**
Photograph labels using volunteers[36]	50	250	20 hours
Supervision of volunteers		1 hour	4 hours
Data entry from photographs[37]	14	70	71 hours

[34] The cost per record for data entry without a photograph of $4.55 has been reduced by 20% for data entry with a photograph.

[35] Assume five hours on-the-job per day.

[36] These times are based on measurements made of volunteers who had little experience. The overarching emphasis was on care and accuracy, not speed.

[37] Estimate. Based on transcription times with volunteers – with and without photography. We expect a similar effect here – an hour spent photographing, is an hour not needed for data entry.

Discussion

Volunteers are able to photograph specimen labels and transcribe that information on those labels into a spreadsheet. However, this study has pointed out strongly that data entry into EMu is the sticking point for capturing data on specimen labels and currently this needs to be done by experienced technical staff. This is due to the vagaries of the database, the knowledge required to determine what data is fit for import, and the interpretation of incomplete information.

There is an EMu online toolset that has potential for remote data entry. If this proves to be suitable, this toolset could be used to check data against acceptable terminology at the time of data transcription, and prepare it in a form for direct data entry into EMu. These steps could be done by volunteers. The final step of importing into EMu would have the data pre-validated and in the correct form, leading to a reduced amount of technical checking and supervision.

Conclusion

In an effort to speed up the capturing of data from museum specimen labels, a method was trialled that broke the data capture process into a set of tasks that could be done by volunteers. A photograph was taken of each specimen label and the information written on that label was transcribed into a spreadsheet. Locations were converted to latitude and longitude. This spreadsheet of data was then prepared and entered into EMu by an Australian Museum technician.

The trial determined that:

The task of taking data from a spreadsheet and entering it into EMu needs to be done by a person familiar with EMu. This task is not straightforward and took an inordinate amount of time. Errors, misspellings, and uncertainties on many of the labels meant that the spreadsheet of data became a clumsy and inefficient method of data entry.

There are some advantages to taking a photograph of a specimen label before capturing the data from that label. The photograph becomes a verbatim record in the database of the label for later referral, and the following transcription process is quicker by about 20%, easier and more convenient. The amount of time spent photographing labels, is about the same as the time saved by entering data into the database from the photographs. The data can be transcribed several times to ensure accuracy.

Volunteers could readily and safely photograph specimen labels within the museum. However, unless we build appropriate safeguards, entering data from those photographs into EMu is a task best done by AM technical staff. There is an EMu online toolset that may at least partially meet these needs.

Recommendations

To speed up the process of capturing data on specimen labels, the Australian Museum can make use of volunteers.

Volunteers can be put to good use by having them take photographs of specimen labels. These photographs would be saved on EMu as a record of the label, and subsequently used for data entry by AM technical staff at a convenient time.

A team of volunteers could be trained to take photographs of specimen labels. Using a single camera and computer set-up, volunteers could work in pairs for limited sessions to avoid boredom.

Investigate the EMu online toolset as a means to use volunteers for data entry into EMu.