Approaches for Involving Volunteers into the Process of Metadata Capture from Specimens

Report for the SYNTHESYS II project

Network Activity 3, Deliverable 3.1

Jörg Holetschek
Botanic Museum & Botanical Garden Berlin-Dahlem
Königin-Luise-Str. 6-8
14195 Berlin-Dahlem
j.holetschek@bgbm.org

Table of Contents

1. Motivation

Specimen collections house huge amounts of preserved organisms gathered by collectors throughout the world over the past 300 years. This stock is the result of innumerable working years of botanists and zoologists past and present. The specimens provide rich and verifiable documentation of the planet's flora and fauna throughout the centuries covered by these specimens. They are a valuable source of information and material for today's biological research, for example for getting a better understanding of certain organisms by new analytical methods and for finding evidence for past and ongoing changes (and losses) of our biodiversity.

Typical examples for such collections are herbaria. In the traditional way, to make use of the material in the herbaria, researchers had to travel in order to get physical access to the specimens, or the specimens had to be sent on loan. The disadvantages of this are obvious: apart from being time- and cost-intensive, the specimens could be damaged during transport, even when handled with care. To overcome this, the efforts for digitising the collections have grown in the past. Instead of being sent via mail, loan requests are acted on by creating a high-resolution image of the specimen, which is then made available on a web server. This does not completely substitute personal travel and manual inspection of herbarium sheets, but it greatly reduces the number of specimens that have to be sent and the time required for getting access to the desired specimen. In addition, some of the collections are digitised systematically. Some of these herbaria house millions of sheets, so this process might take years, even decades. The ultimate aim is to fully digitise the herbaria, making them easily accessible over the web.

As the number of digitised specimens grows, the need for efficient search mechanisms becomes more important. Cataloguing digitised specimens is done by associating them with metadata, information usually taken from the labels attached to the specimens. Typical pieces of metadata are the catalogue number and a potentially existing barcode or field number; the taxon name of the specimen (species in most cases, or a higher taxon if identified only to a higher level); gathering agent, gathering time and gathering location (country, state, province, town, detailed description of locality, geographic coordinates); type status information, if appropriate; and potential annotations added by past researchers. This information will help users to find the specimens they are interested in, for example specimens of a certain species or taxonomic group, specimens gathered by a certain collector or during a specific expedition, or specimens gathered in a certain geographic region during a specific period of time.

The need of capturing as much of these metadata items as possible is intensified by another trend: initiatives such as the SYNTHESYS and the Global Biodiversity Information Facility (GBIF) are setting up networks for connecting biodiversity data from distributed sources. This enables access to a multitude of collections, potentiating the amount of possibly useful data in which users have to find their pieces of interest. Currently (July 2010), GBIF offers access to 1.21 million images of specimens from 84 collections. Browsing these images is not feasible, even for a single one of these collections. Without metadata, finding the right specimens is like finding a needle in a haystack.

The collation of data from various sources can potentially enable findings that cannot be gained by examining individual data sources. Given an ample amount of underlying data, a temporal analysis of occurrence data can provide insights into biodiversity changes for certain organisms in a given geographic region. Combined with climate models, potential changes of biodiversity in the future can be predicted, or at least probabilities of certain scenarios (ecological niche modelling). This is important for the evaluation of the threats by invasive species or the potential spread of disease vectors. Capturing geographic and temporal metadata as accurate as possible enables the use of specimen information for this kind of analyses.

Traditionally, metadata are captured manually in the course of the digitisation process. Before or after the image is taken, the agent will type in label or ledger information into a database that is used to organise the specimen images. When connected to biodiversity networks such as BioCASE (Biological Collection Access Service) or GBIF, these data will be published along with the digital image.

As imaging technology advances and the time required for taking the picture diminishes, metadata capture becomes a bottleneck. Recently, specimen digitisation has moved to a new stage by setting up streamlined, conveyor-based digitisation workflows, even with several lines in parallel, allowing for digitisation on an industrial level. With these setups digitising up to 10,000 specimens a day, full metadata capture becomes the crucial bottleneck that cannot be resolved with traditional approaches.

This report tries to investigate the potential for metadata capture from label or ledger images by involving volunteers. “Crowdsourcing”, i.e. harnessing the labour of volunteers by using current web technologies, might enable or speed up large-scale metadata capture.

As it turned out during the investigations, most of the current developments focus on herbarium collections. The reason for this might be the fact that the digitisation of herbarium specimens is easier than for other types of collections, due to the two-dimensional nature of most herbarium specimens. Up to now, specimen digitisation on a large scale was done only for herbaria. Therefore, projects for metadata capture from specimen labels focus on herbarium sheets. However, the methods and algorithms developed for this area can mostly be applied to other types of collections once they are digitised.

2. Specimen Metadata

A typical example of a collection specimen is a herbarium sheet as the one shown in Fig. 1. Besides showing the actual specimen (the organism, 1) it may hold additional plant material (seeds, fruits: 2) in a paper pocket. Attached to the specimen sheet, written or stamped directly on the sheet, you may find the following types of information:

- The herbarium label (3) identifies the specimen either on species or a higher level. Apart from the identified taxon it usually records the identifier person and the identification date, the collector and the gathering date, and information on the gathering site – country, state or province, town or named area and a description of the locality. So the herbarium label holds very important pieces of metadata. It is mandatory for all sheets, but the amount of information can differ. It may be handwritten or typewritten.
- Additional identification labels (4; can be one ore more) record subsequent identification events. They may confirm the first identification result or contradict; together with the identification result it lists the identifier and the identification date. Sometimes, they can cite a publication that refers to this specimen.

- If digitized, the sheet usually contains a scale (5), a colour scale (6) and a digitization stamp (7). Some specimens are digitised twice, so more than one digitisation stamp can be found. This happens when an organism gets re-identified with another result or a specimen is of such high importance that it needs to be re-digitised with a higher resolution (typical for type specimens).
- A barcode label (8), if the specimen has been assigned a barcode.
- A type designation (9) marks the specimen as a type. An additional label (10) may specify the type in more detail.
- A stamp identifies the herbarium (11) and potential previous herbaria owning the specimen (12).
- Loan numbers (13) can be used to tell when and how long the specimen was given on loan.

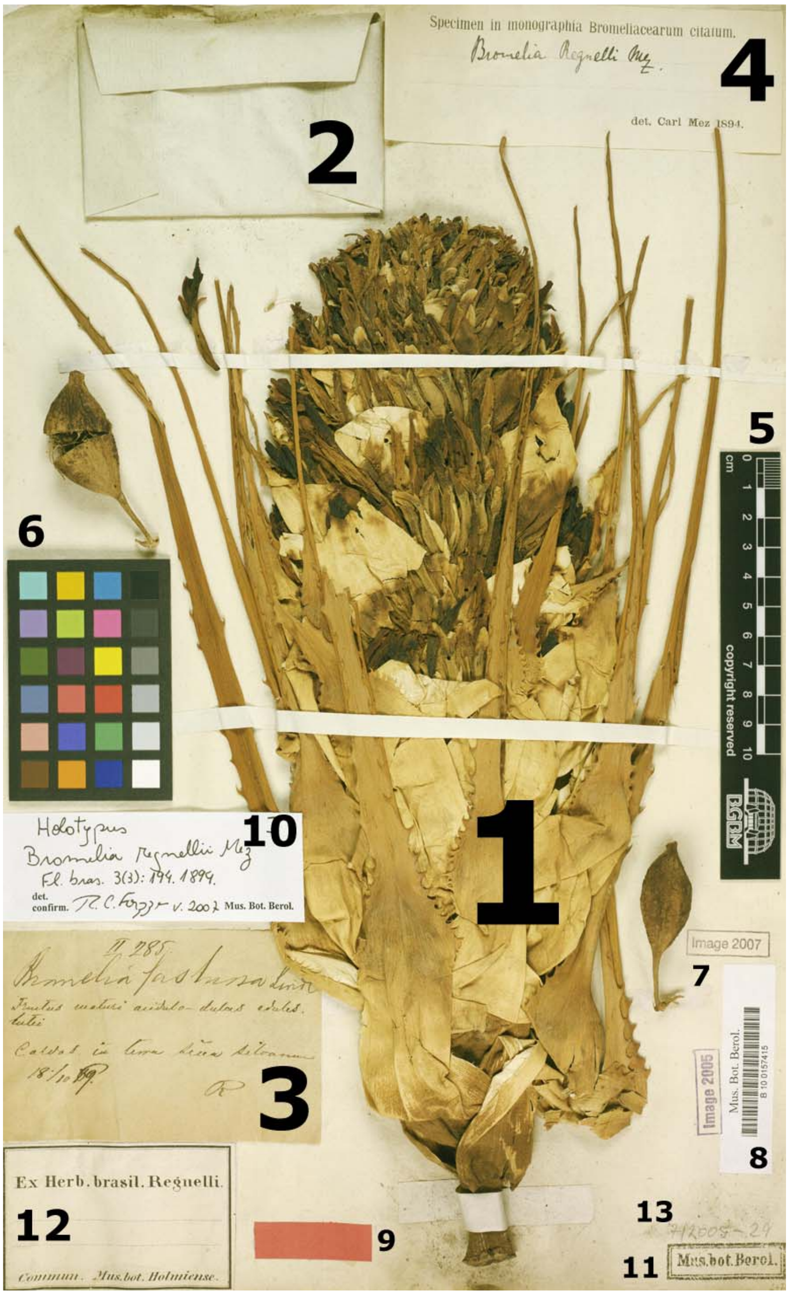


Figure 1: Typical herbarium specimen sheet with different objects attached

Accession numbers (not shown Fig. 1) are an additional metadata item often assigned to specimens and stamped or written on the sheet.

Apart from the herbarium label, all these pieces of information are optional. They may be missing, or they may exist several times. They may be handwritten or typed. There is no defined position on the sheet for them; moreover they can be sideways or even upside down.

Keeping these things in mind, the task of acquiring metadata from specimen labels can be subdivided into three subtasks, namely

(1) Identification of labels or parts of the specimen sheet that hold metadata: the herbarium label, identification and annotation label(s), type markers and type specifications, notes on previous herbaria, accession and loan number(s), barcodes.

(2) Classification of features holding metadata. This helps with the actual data capture in the next step, when OCR (optical character recognition) is used or when labels are assigned to specialists. In both cases knowing the type of information the feature is depicting is helpful; OCR can be supported by specialised dictionaries, and different people can be assigned to task like deciphering geographical information or taxonomic names. Moreover, different strategies can be applied for the different types of features, or it can be decided that only some of them get captured.

(3) Capture of the metadata.

The following chapter will introduce several initiatives and approaches trying to cope with these tasks:

- HerbarDigital aims at digitising herbarium specimen sheets; some of the approaches developed for this project refer to subtask (1) and will be outlined in the first section.

- The second section describes ideas for organising the workflow between the three subtasks and for optimizing the capturing process.

- ReCAPTCHA is a service using the deciphering capacities of humans to defend against web access by automated programs. It can be put to use in the context of subtask (3).

- Herbaria@home is a non-commercial project in the United Kingdom for capturing metadata from digitised herbarium sheets by volunteers; it concerns all three subtasks.

- Helping Science is a system developed by Silver Biology, a commercial software company in the United States. It uses a different approach than Herbaria@home but also aims at metadata capture from digitised herbarium sheets referring to the three subtasks.

3. Current Initiatives and Approaches

3.1 Feature Detection in the Herbar Digital Project

This section outlines approaches that have been developed for the task of identifying parts of the voucher that hold metadata. Even though this is not directly related to the involvement of volunteers, it is included in this report because it is a necessary preparatory step required for finding the pieces of metadata that can be captured later by volunteers.

The project “Herbar Digital – Rationalizing the virtualisation of botanical document material and their usage by process optimization and automation” by the University of Applied Sciences in Hannover has the aim of developing an efficient workflow and powerful hardware and software solutions for digitising the 3.5 million specimens of the Herbarium Berolinense. Portions of the project deal with analysing the high resolution images created in the digitisation process, and several approaches are used for identifying the elements of a specimen sheet in an image.

3.1.1 Optical Character Recognition

Commercial OCR software is capable of reading high quality printed texts with a low error rate. However, they run into great difficulties when attempting to process text located next to non textual objects. Specimen sheets as the one shown in Fig. 1 are a very complex setup: handwritten and typed text is mounted side by side with plant material, pockets, stamps and scales, sometimes overlapping each other or upside down. Even though OCR solutions exists for recognising text in images, such as book covers or scanned filled-in forms, they are mostly dealing with printed or typed text. The medley of old-fashioned handwriting of different persons, printed text in different sizes and fonts and stamps is too much for the existing software, so a different approach is needed.

3.1.2 Localising text in an Image

Since OCR often fails in identifying the location of text in complex surroundings, the text needs to be identified and singled out for processing. This can be done using the special characteristics of text:

- Text is distinguished by high contrast.
- It usually appears in rows.
- It is letter-sized.
- It has neighbours in the same row.
- It can be found mostly in text-like surroundings.

The approach developed by Herbar Digital consists of a series of steps. In essence, it consists of the following stages:

1. Gray scaling and reducing the image resolution from 600 to 150dpi in order to keep the required computation time in reasonable boundaries.
2. Contrast amplification to make the elements and background easier distinguishable.

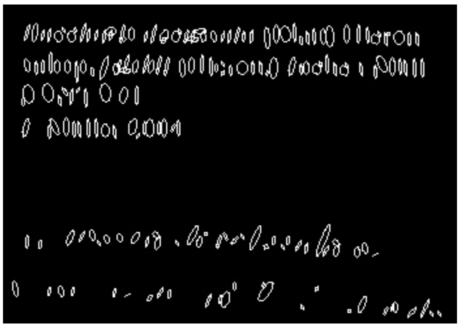


Figure 2: Approximation of text by ellipses

3. The image is then approximated by ellipses, each defined by centre, height, width and tilt (example result in Fig. 2). This reduces all components to ellipses, the organism as well as stamps and text; heuristics will then be used to filter ellipses with text-like features.
4. In a last step, the filtered ellipses are merged into lines.

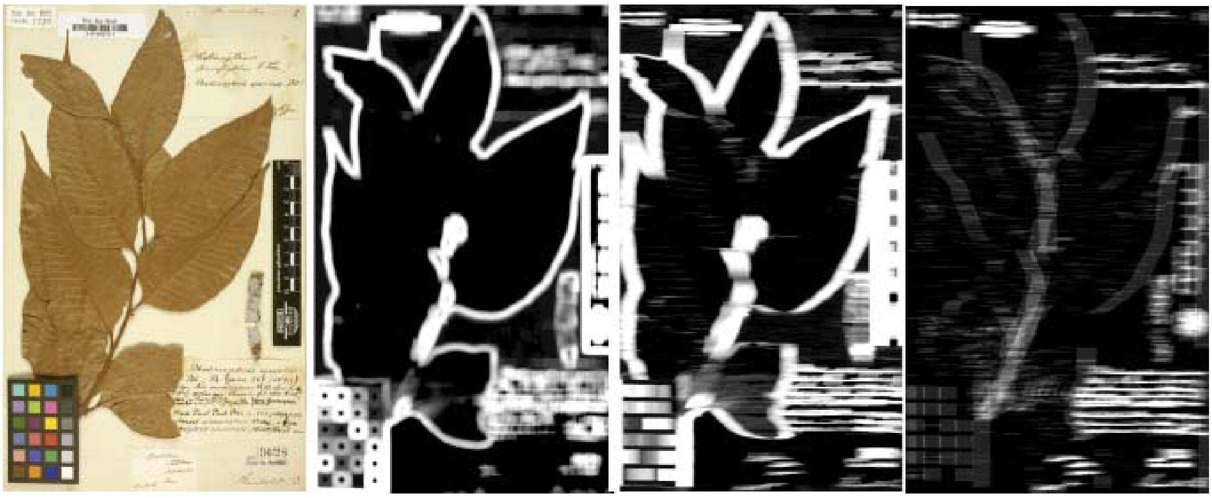


Figure 3: Original specimen image and different types of contrast images

One of the classifiers used in step 3 is the contrast. For each pixel of the original image (leftmost in Fig. 3), the contrast in the surrounding square is calculated and assigned a brightness value (second left). It clearly shows that not only text has a high contrast, but also leaf edges, the ruler and colour scale. Since the horizontal contrast particularly distinguishes text from other objects, calculating the contrast values of the horizontal surrounding for each pixel leads to an improved contrast image (second right). An even better (and faster) result can be produced by an algorithm that counts the number of black/white alternations in a horizontal surrounding of each pixel (rightmost). Plant and colour scale are widely eliminated, barcode and textual components are emphasized. Rows of text show up as bright lines. Bright areas in the image have a high probability of representing text, so ellipses that fall into bright areas of the contrast image get classified with a high probability score.

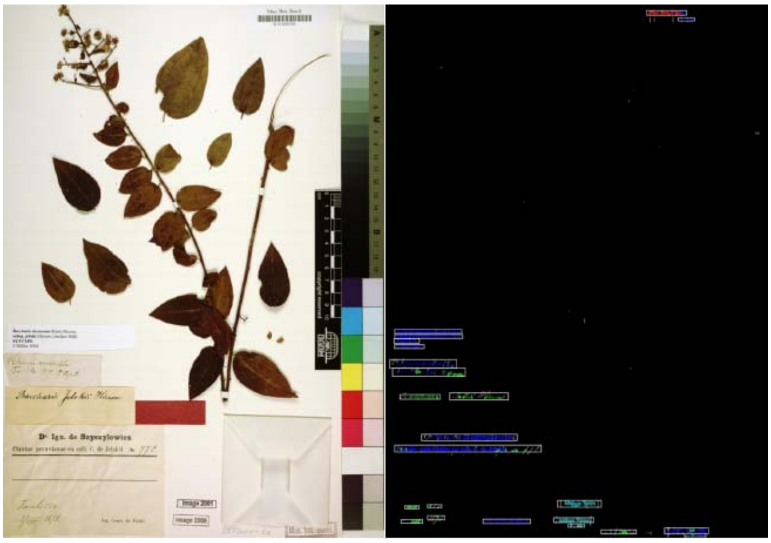


Figure 4: Original specimen and text objects detected

Figure 4 shows an example of an original specimen and the final result of this process. The plant, the ruler and the colour scale are completely removed, only handwriting and printed text as well as barcode and stamps are left.

3.1.3 Template matching

Some of the steps involved in the approach described in the previous section cost considerable amounts of computing time. Reducing the complexity of the original specimen image by eliminating non-textual objects can save computing time that could be invested later into the algorithms for identifying text. Herbar Digital uses template matching for finding and eliminating features on the specimen images that are irrelevant for subsequent processing steps.

The objects on a specimen sheet can be classified according to their variability in respect to size and orientation:

- Stable: Colour scales, ruler
- Less stable: Stamps, Labels with printed titles
- Variable: Type designation mark, barcode
- More variable: envelopes
- Extremely variable: Hand-written annotations, plant.

For all but the last of these classes it is possible to create a collection of templates, i.e. digital examples representing the typical colours, shapes etc. in these features. A template matching algorithm can then be used to find these templates in a given specimen image and, if it is irrelevant for subsequent steps, replace the respective region with blank space. Figure 5 shows the result of this feature elimination: The ruler, grey and colour scale, the barcode and the "Image" stamps were removed.

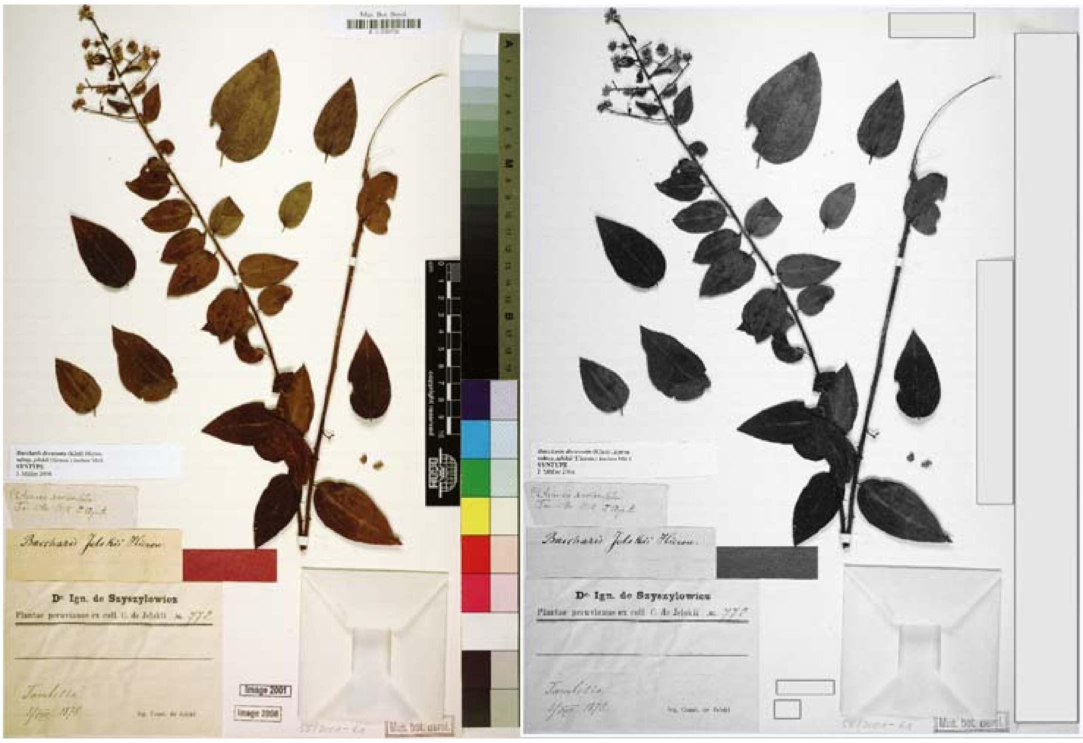


Figure 5: Original Specimen and objects removed with template matching algorithms

3.1.4 Joint component identification

The largest feature in the image is usually the plant specimen itself. Joint component identification is able to recognise the plant as a set of big joined components that can be filled with colour and deleted to facilitate feature recognition as well as the metadata capture itself.

3.2 Optimizing the Workflow for Metadata Capture

As outlined before, the features that can be found on a specimen are very diverse, with respect to their metadata content, its semantics, as well as with respect to form (in texts: handwritten, typewritten, and stamped). Therefore, the methods for metadata capture need to be flexible, too. Typewritten and stamped information are good candidates for OCR software; if linked with the growing number of web services that provide taxonomic terms or geographic names (including historic place names), they can deliver good results. Handwritten texts require more sophisticated algorithms for handwriting recognition; in many cases, they can deciphered reliably only by a human.

If it comes to historic specimens, only highly skilled curatorial personnel are capable of dealing with all pieces of metadata; most people hired for data capture frequently run into difficulties resulting in information which is either erroneous or incomplete. This can be caused by insufficient taxonomic knowledge or by difficult historic place names; sometimes the handwriting of certain collectors is particularly challenging, so can be the habits of some collectors in using abbreviations etc. On the other side, some of the pieces can be captured without difficulties by less qualified personnel.

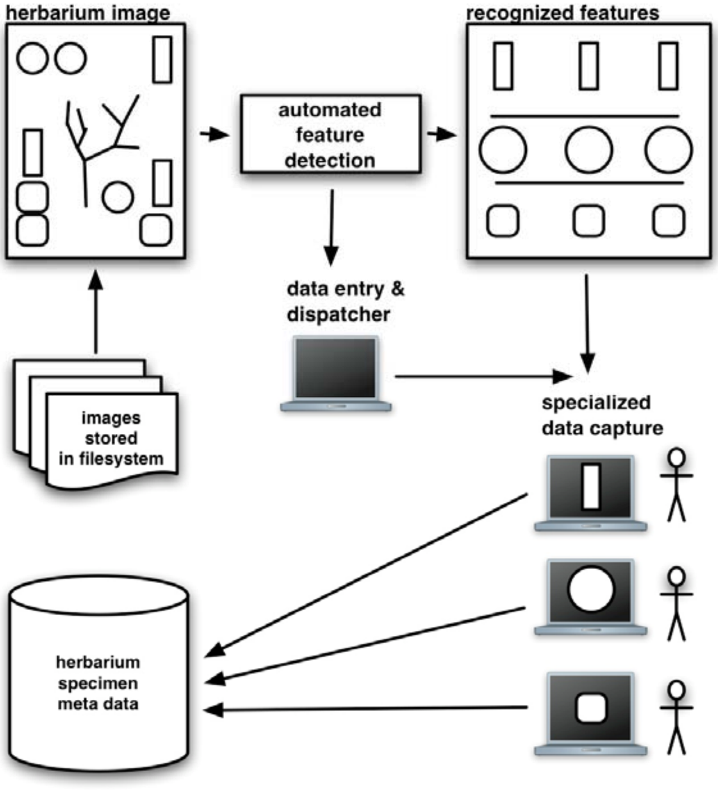

herbarium image
recognized features
automated feature detection
data entry & dispatcher
images stored in filesystem
specialized data capture
herbarium specimen meta data

Figure 6: Optimised workflow for metadata capture

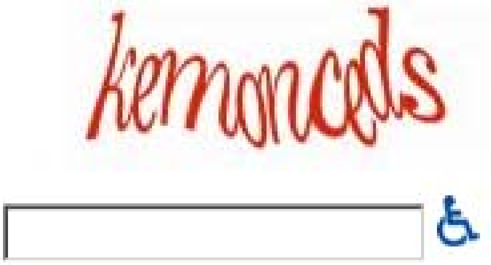
Hence, assigning the entire process to a single person may either result in erroneous data or in wasting the specialist's time on routine tasks. The delegation of specific sub-tasks to specific personnel or to appropriate software components such as OCR or handwriting recognition presents a great potential for improving the efficiency of metadata capture. Delegation to specialised personnel would hand on the labels written by a certain collector to someone who is familiar with his handwriting, and let someone who is good at geo-referencing free-text descriptions of places in historic France process label text that potentially includes such data.

Starting with the digitised herbarium specimen, the feature detection and localisation described in the previous section could be used to find features on a herbarium sheet (see Fig. 6). Recognized features would be stored as partial images. OCR techniques could be applied whenever typewritten or stamped information, such as label information of modern herbarium sheets, has been identified; the quality of OCR can be increased by connecting the components to taxonomic thesauri or geographic gazetteers. Handwritten text could be subjected to handwriting author recognition based on samples of handwritten text from known collectors, and, depending on the results, the features could be assigned to the person most skilled for the handwriting of the identified person.

Depending on the type of feature detected, a dispatcher component would delegate the task of specialized data capture to either a person with the most appropriate knowledge or skills or to an applicable software service. The dispatching process could be performed by a person or, as the rules for dispatching are improved over time, by software. Existing services and tools for quality checking of primary biodiversity data could be integrated into the data capture.

3.3 ReCAPTCHA

CAPTCHA is a challenge response test used on the World Wide Web to determine whether a user is a human or a computer program. The acronym stands for Completely Automated Public Turing test to tell Computers and Humans Apart. A typical CAPTCHA is an image containing several distorted characters that appears at the bottom of web registration forms (see figure at right). Users are asked to type the wavy characters to prove they are human. Current computer programs cannot read distorted text as well as humans can, so CAPTCHAs act as sentries against automated programs attempting to abuse online services.



According to the creators of reCAPTCHA, humans around the world type more than 100 million CAPTCHAs every day, in each case spending a few seconds typing the distorted characters. In aggregate, this amounts to hundreds of thousands of human hours per day. ReCAPTCHA tries to make positive use of the time spent by humans solving CAPTCHAs, because deciphering text requires them to perform a task that computers cannot.

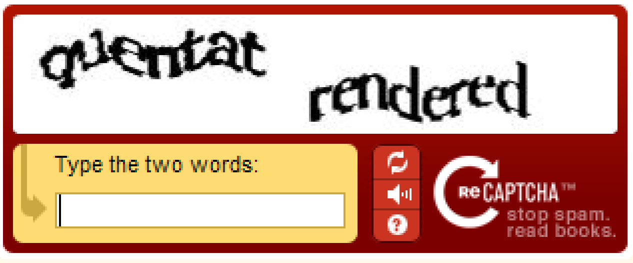
ReCAPTCHA[1] is a project originally founded at Carnegie Mellon University, but acquired meanwhile by Google. The following synopsis has been extracted from Ahn (2008)[2]:

[1] http://recaptcha.net/

The idea behind reCAPTCHA is to make use of the people's deciphering power for capturing typeset texts from digitised sources – mainly books, such as the Google Books Project, or other documents, such as the non-profit Internet Archive. The pages are photographically scanned on a large scale and the resulting images converted into text by optical character recognition software. But as mentioned earlier, OCR software is far from perfect. For older prints with faded ink and yellowed pages, OCR cannot recognize about 20% of the words. In contrast, humans are much more accurate at transcribing such print. For example, two humans using the "key and verify" technique (where each types the text independently and then any discrepancies are identified) can achieve more than 99% accuracy at the word level.

Whereas standard CAPTCHAs display images of random characters rendered by a computer, reCAPTCHA displays words taken from scanned texts. The solutions entered by humans are used to improve the digitisation process. To increase efficiency and security, only the words that automated OCR programs cannot recognize are used. The guesses of two different OCR applications (the sequence of letters they detected) are stored for later comparisons with the user's guesses.

To meet the goal of a CAPTCHA, the system needs to be able to verify the user's answer. To do this, reCAPTCHA presents the user with two words, the one for which the answer is not known, and a second "control word" for which the answer is known (see figure at right). In this example, "quentat" was unrecognizable by OCR software. ReCAPTCHA isolated the word, distorted it using random transformations and presented it as a "challenge word" to users. If users correctly type the control word ("rendered" in this case), the system assumes they are human and gains confidence that they also typed the unrecognized word correctly.



To account for human error in the typing process, reCAPTCHA sends every unrecognised word to multiple users, each time with a different random distortion. At first, it is displayed as an unknown word. If a user enters the correct answer to the associated control word, the user's answer is recorded as a plausible guess for the unknown word. If the first three human guesses match each other, but differ from both of the OCR's guesses, the word becomes a control word in other challenges. In case of discrepancies among human answers, reCAPTCHA sends the word to more humans as an unknown word and picks the answer with the highest number of "votes", where each human answer counts as one vote and each OCR guess counts as one half of a vote. A guess must obtain at least 2.5 votes before it is chosen as the correct spelling of the word. Hence, if the first two human guesses match each other and one of the OCR's, they are considered a correct answer; if the first three guesses match each other but do not match either of the OCRs, they are considered a correct answer, and the word becomes a control word. 67.87% of the words required only two human responses to be considered correct, 17.86% required three, 7.10% required four, 3.11% required five, and only 4.06% required six or more (this includes words discarded as unreadable).

[2] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, Manuel Blum: „reCAPTCHA: Human-Based Character Recognition via Web Security Measures". Science Magazine, Issue 321, 12 September 2008, pp. 1465-1468. Published online 14 August 2008 [DOI: 10.1126/science.1160379]

To account for unreadable words, reCAPTCHA has a button that allows users to request a new pair of words. When six users reject a word before any correct spelling is chosen, the word is discarded as unreadable. Visually impaired users can click the audio button to hear a set of words that can be entered instead of the visual challenge.

ReCAPTCHA can be used like a traditional CAPTCHA system to prevent web forms from being abused by scripts and programs. With mailhide, authors of web pages can protect email addresses from being harvested and used for spam: Visitors are asked to solve a reCAPTCHA before they can view the email address. ReCAPTCHA is implemented as a web service – potential users just have to include some lines of code into their sites to make use of it.

Using reCAPTCHA instead of traditional CAPTCHAs has some advantages for the owner of the website. One apparent reason is that the human processing power otherwise wasted will be used for solving problems that computers cannot solve yet. If the result is beneficial for a certain community – as in the case when books get digitised and are made freely accessible to the society –, this contribution is a clear advantage over a traditional CAPTCHA.

Another reason is of a much more pragmatic nature: Since reCAPTCHA uses only words upon which OCR algorithms failed, it is currently more secure than conventional CAPTCHAs that generate their own randomly distorted characters. The words displayed by reCAPTCHA have three different types of distortion: First (and most importantly), there are natural distortions that result from the underlying texts having faded through time. Second, the scanning introduces noise. Third, reCAPTCHA applies artificial transformations as the ones used by standard CAPTCHAs, such as waving the baseline and striking through. In combination, this produces a hard challenge; solving reCAPTCHA programmatically would require an OCR algorithm that is a real advance in the state of art.

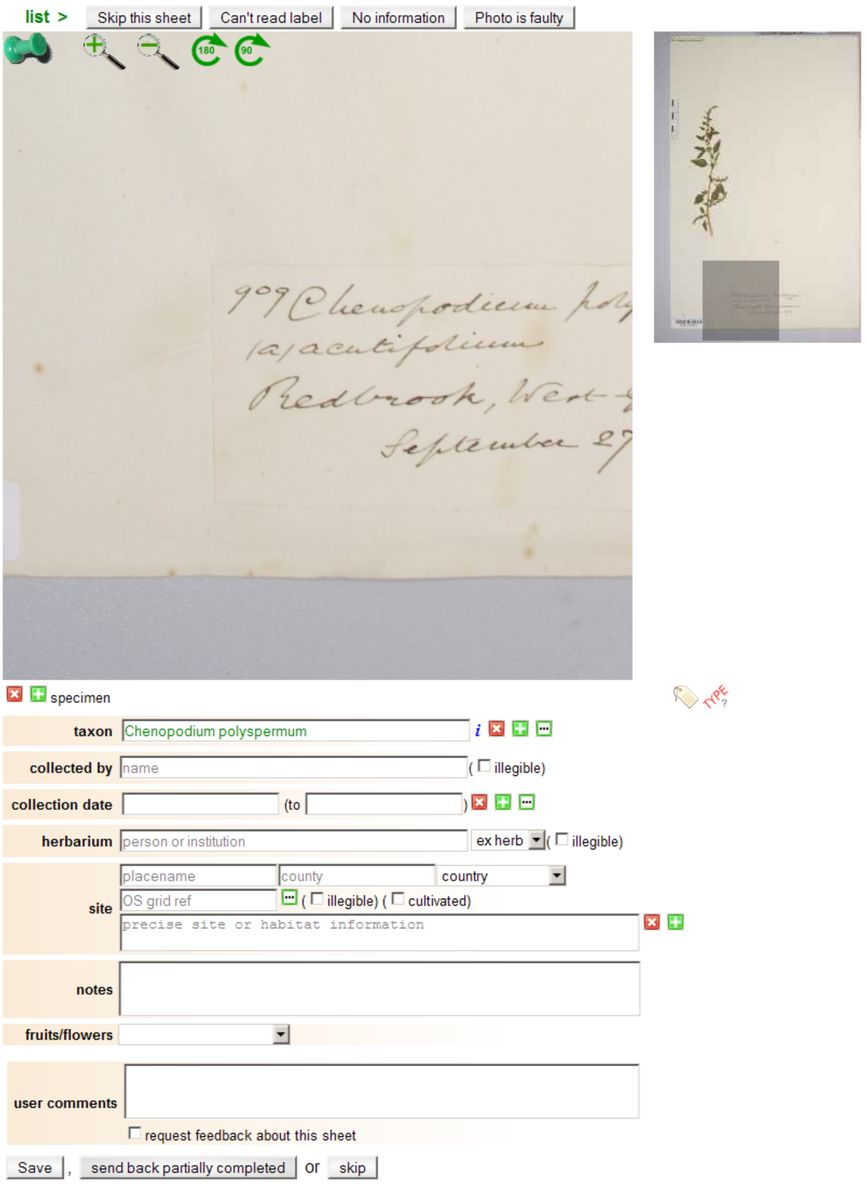
Surprisingly only at first sight, it typically takes no more time for users so solve a reCAPTCHA, even though it presents two words instead of one. Standard CAPTCHAs present six to eight randomly chosen characters without any meaning, whereas the words used by reCAPTCHA are natural language words with patterns a human is accustomed to. Deciphering two meaningful words does not take more time than deciphering a sequence of arbitrary letters, so reCAPTCHAs and CAPTCHAs take about the same time to be solved.

ReCAPTCHA has proved to harness large amounts of human mental effort. After exactly one year of running the system, more than 1.2 billion reCAPTCHAs had been solved, amounting to over 440 million words correctly deciphered. Assuming 100,000 words per book (400 pages, 250 words per page), this is equivalent to over 17,600 books manually transcribed (about 25% of the words in each book had not been recognized by OCR). And popularity of the system continues to grow: Whereas in 2008 the rate of transcriptions per day was about 4 million, the number today exceeds 11 millions today (August 2010), which is equivalent to 440 books per day. Achieving this rate via conventional "key and verify" means (without aid from OCR, so every word in a text would be typed) would require a workforce of more than 2,000 people deciphering words 40 hours per week (assuming an average rate of 60 words per minute).

3.4 Herbar@Home

Herbaria@home[3] is a volunteer-based project that aims at cataloguing historical herbarium collections hosted by universities and museums.

It combines the three steps outlined in the beginning of this report without clearly separating them. This is possible because neither OCR nor any other feature detection software is used. Herbaria@home solely relies on the recognition power of its volunteers; finding text objects with metadata on a specimen sheet as well as classifying and capturing them is done by the user in one step. The project focuses on historical collections with mainly handwritten labels and annotations, often with old-fashioned writing, on which OCR algorithms usually fail.



909 Chenopodium poly
(a) acutifolium
Redbrook, West-
September 27

☒ ⊞ specimen

taxon: Chenopodium polyspermum *i*

collected by: name (□ illegible)

collection date: (to)

herbarium: person or institution | ex herb ▾ (□ illegible)

site: placename | county | country ▾
OS grid ref (□ illegible) (□ cultivated)
precise site or habitat information

notes:

fruits/flowers: ▾

user comments:

□ request feedback about this sheet

Figure 7: The herbaria@home workspace

After registering with herbaria@home, the user can log into the website and work on cataloguing specimens. He will be assigned a set of four randomly chosen specimens by

[3] http://herbariaunited.org/atHome/

default; alternatively, he can choose sheets from a particular collection or search by genus for a particular set of sheets to work on. With this feature, volunteers can focus on a collection or genus of their special interest.

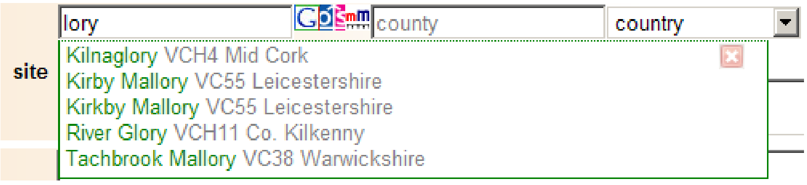
The user will be presented with a page showing the full specimen as a small thumbnail (right side) and a working area that can be dragged to any region of the sheet (see Fig. 7). The zoom level can be adjusted; the image can be rotated to account for upside-down labels. A specimen can be rejected completely because the labels are too hard to read or because the image is faulty or contains no information; in these cases, the specimen will be marked as illegible or faulty. If the user decides to skip a specimen, it will stay in the user's queue.

The fields for entering the captured data are located below the working area. The taxon name is already filled in: This piece of information is already recorded during digitisation. However, if the user finds the recorded name incorrect, he can edit this field. The form can handle multiple identifications; the user can add and remove identifications. One determination can be marked as being preferred.

The fields the user is asked to fill in are: Collector name; gathering date (with a range, if applicable); potential previous herbaria or the source of the specimen; the gathering location with place name, county, country and a precise locality or habitat description; the state of flowers of fruits on the specimen, if any; type status; and a notes and a comments field. The form can be expanded to show additional fields: The determination(s) can be append with identifier person, identification date and notes; the gathering date with a note; and the location with geo references - coordinates, grid square or grid point.

None of the fields are mandatory, so they can be left blank if no information can be found on the specimen. If the user cannot decipher the handwriting, the appropriate field can be marked as illegible. If unsure about any of the fields, the specimen can be saved and marked as still incomplete; it will stay in the user's specimen queue and can be re-edited and completed later.

Even though there is no OCR taking away any work from the user, the fields offer very useful assistance. As soon as you start typing in a collector of field name, for example, a list of known collectors and place name pops up. It doesn't matter whether the typed-in letters are initials, given name or surname – the popup list will show all matching entries. The pop-up for the gathering site (see figure below) has even more functionality built in: fractions typed in will be used for a full-text search on place names; if the user selects an entry, county and country will be filled in automatically. For all fields that offer this suggestion tool the user can still type in any value if none of the suggestions matches the information on the specimen sheet.



site

lory | county | country

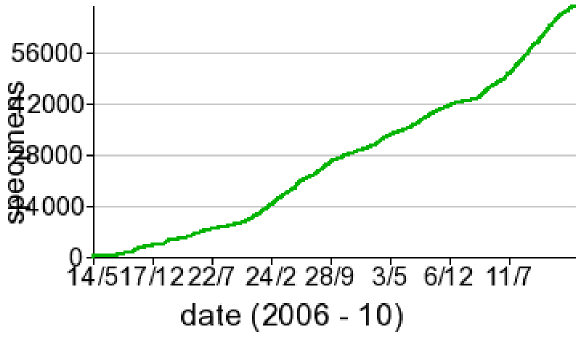
Kilnaglory VCH4 Mid Cork
Kirby Mallory VC55 Leicestershire
Kirkby Mallory VC55 Leicestershire
River Glory VCH11 Co. Kilkenny
Tachbrook Mallory VC38 Warwickshire

Once the user has completed a specimen, he will be presented the next item in his or her queue. Once the whole queue is done, the user can choose to be assigned another set of specimens (with between one and ten items) or to finish work.

The specimens catalogued can be search and viewed by anyone without registration. A form similar to the data-entry form can be used to specify the search criteria. The same assistance tools available for data entry can be used for the search criteria. However, selecting entries from the popup lists doesn't necessarily return any search hits, in case there is no specimen catalogued yet for a certain place name, for example.

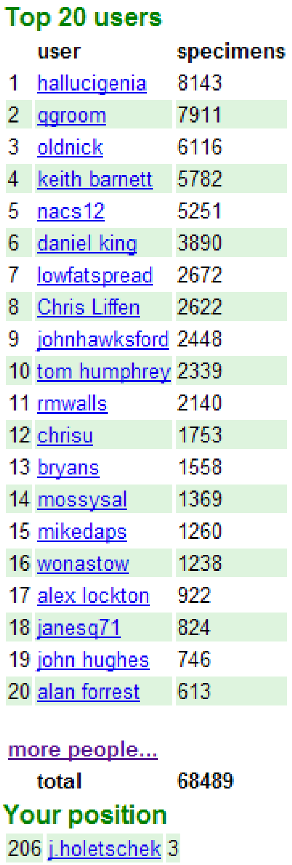
Users interested in specimens of certain taxa, from a particular collector or a specific place, county or country can set up watch lists (registration is required for this). A watch list serves as a filter for newly catalogued specimens; they offer the opportunity to keep track of new records of special interest. When logging in, the user will be presented with a list of matches; alternatively, a watch list can be configured to send a daily, weekly, fortnightly or monthly email update.

Herbaria@home seems to have quality checks built-in to validate the data captured by users; unfortunately, these are not documented. A small portion of records (around one percent) will be sent to multiple participants to cross-check the results.



As of August 27th, a total of 69.227 specimens have been documented with herbaria@home. As the graph in the figure to the left shows, the documentation rate seems to increase slowly. One reason for this might be that the number of volunteers grows steadily, while old users seem to continue their work.

Since the volunteer's work is the only asset of herbaria@home, giving incentives to users is important. Herbaria@home stimulates competition between users by keeping track of the number of documented records for each user. High score table like the one shown in figure below list the most hard-working volunteers. Complete lists of all users with their documentation efforts can be viewed; for individual users, a graph showing the number of records per day can be produced. Little pieces of statistics indicating the user with the most records in the current month or the highest number of records per day keep track of the current activities.



Top 20 users

	user	specimens
1	hallucigenia	8143
2	ggroom	7911
3	oldnick	6116
4	keith barnett	5782
5	nacs12	5251
6	daniel king	3890
7	lowfatspread	2672
8	Chris Liffen	2622
9	johnhawksford	2448
10	tom humphrey	2339
11	rmwalls	2140
12	chrisu	1753
13	bryans	1558
14	mossysal	1369
15	mikedaps	1260
16	wonastow	1238
17	alex lockton	922
18	janesg71	824
19	john hughes	746
20	alan forrest	613

more people...

	total	68489

Your position

206	j.holetschek	3

Currently, the system uses specimens from Birmingham University, Bolton Museum and Charterhouse School. Herbaria@home is looking for more collections to take part in the scheme and open to collaborations and offers of data. For small, under-resourced collections the offer includes arranging for someone to photograph the collection and handle the data processing. Larger institutions are invited to consider using the project to target parts of their collections that cannot attract the resources needed for documentation by traditional methods.

3.5 Helping Science

Helping Science[4] is a project developed by Silver Biology aimed at using citizen science for digitizing herbarium specimens. Even though this is a commercial company and the use of the system will probably involve a licensing fee once its development has been finished, the basic approach is worth being looked at and will therefore be outlined in this section.

Similar to herbaria@home, Helping Science covers all three involved subtasks: identifying labels on a specimen sheet, finding and classifying the pieces of metadata and their capture. However, it uses OCR algorithms on features that contain text. In contrast to processing the whole sheet, processing these features only reduces complexity to a level that can be handled by current OCR software.

Figure 8 show the basic life cycle of a herbarium label in Helping Science.

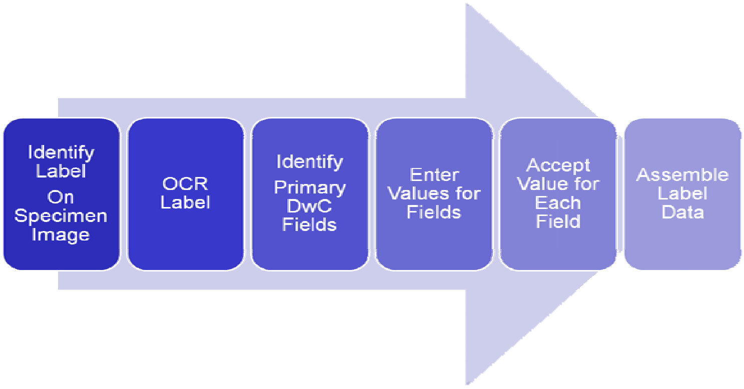

Identify Label On Specimen Image
OCR Label
Identify Primary DwC Fields
Enter Values for Fields
Accept Value for Each Field
Assemble Label Data

Figure 8: Helping Science Processing Stages

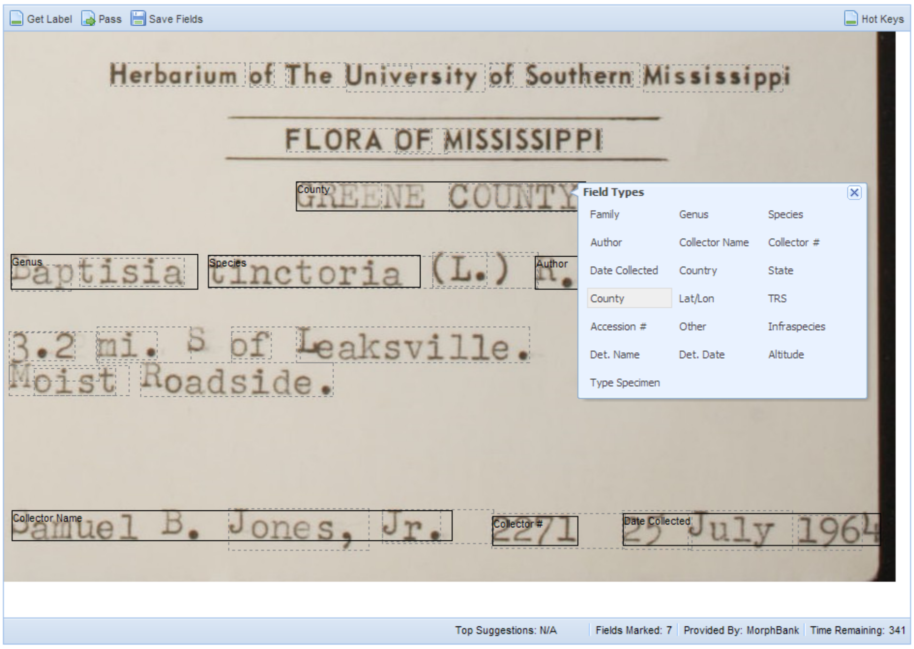
1. In a first step, the label is identified on the herbarium sheet. As mentioned before, this is a very challenging task for software, but can be easily and quickly done by a human. The user (the volunteer citizen scientist) is presented with a herbarium sheet and asked to cut out the herbarium label. Faulty or undoable specimens can be passed. As a result, the label will be detached from the specimen sheet and saved in a separate JPG file.

2. The image cut out in the first step will be processed by OCR. This is an internal step transparent to the user. It makes use of the Evernote[5] web services, which produce a JSON document (JavaScript Object Notation) for each label holding a list of identified words – each with a bounding box and the captured word, or, in case it cannot be identified without doubt, a list of potential words. Helping Science uses this information for making educated guesses and to assist the user in the next step.

3. In the Helping Science WorkCenter (Fig. 9), volunteers will be presented with the result of the OCR: The image label with bounding boxes for each identified word. The bounding boxes can be moved and resized; in case the OCR missed text elements, more bounding boxes can be added. The volunteer is asked to classify each of the boxes by either choosing from a pop-up list or by using hot keys. Currently, Helping Science is restricted to data fields defined in the DarwinCore standard. Moreover, only the fields shown in the figure are captured, namely

[4] http://www.helpingscience.org/
[5] http://www.evernote.com/

collection information, determination (person, date, result) and some data on geography. Detailed habitat and locality descriptions are neglected.



Herbarium of The University of Southern Mississippi

FLORA OF MISSISSIPPI

GREENE COUNTY

Baptisia tinctoria (L.) R.

3.2 mi. S of Leaksville.
Moist Roadside.

Samuel B. Jones, Jr. 2271 25 July 1964

Figure 9: Tagging metadata elements in the Helping Science WorkCenter

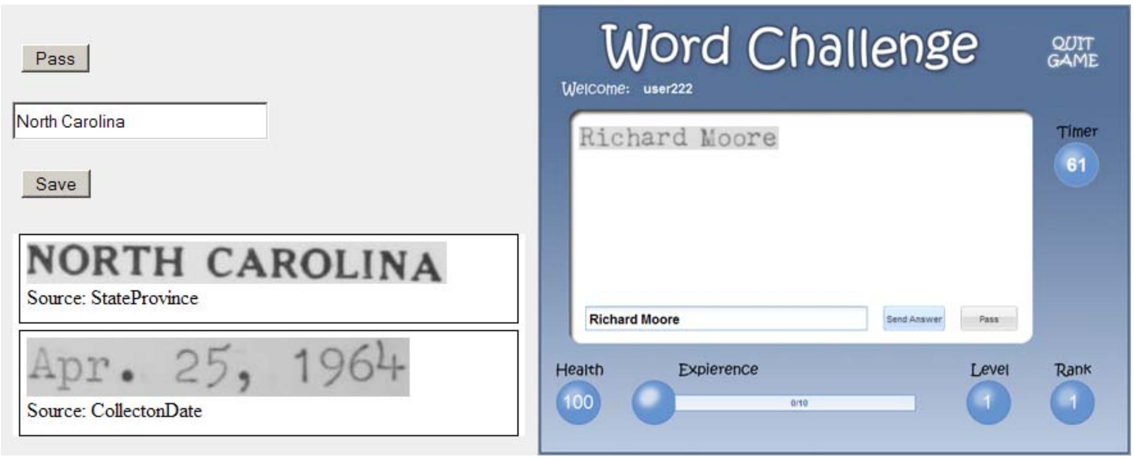
4. In the next step, all the identified and marked fields will be presented to volunteers for being key stroked. Per default, the user will be presented with all types of fields. However, if he or she is specialised on a certain collection or geographic locations or interested in capturing a certain type of field, the queue of fields can be filtered accordingly. By this, specialised volunteers can focus on the tasks that need a particular knowledge, whereas other volunteers can take care of the remaining tasks.

Helping Science has developed two applications for capturing the texts (Fig. 10). Each tagged field will be examined by at least three different users. The simple field entry (left) will just show the image cut-out, the field type and an entry field. The word challenge is disguising the task as a game: the time required for entering each word will be measured and the number of presumably correct or wrong answers will be counted. This is done by comparing the user's guess with guesses of precious users. For a correct answer, the user will win experience points, for wrong ones he will loose health. High scores are available to stimulate competition between users.

On both applications, illegible items can be passed. The interfaces are available on a regular browser, for iPhone, iPad, iPod Touch, and as a Facebook app. An application programming interface allows for 3rd parties to create custom interfaces.

5. The fifth, internal step is about selecting the correct value for each field. Once a predetermined level of accuracy has been reached by being entered by different users, a value is accepted for a field. Fields that were typed in differently by users will be presented to more users; however, no user will see the same field image twice.

6. Once all fields are captured and verified for a given label, the DarwinCore record will be assembled. All processed data runs through a series of taxonomic and geographic validations, and any issues found are reported to the collection manager for review.



North Carolina

NORTH CAROLINA
Source: StateProvince

Apr. 25, 1964
Source: CollectonDate

Word Challenge

Welcome: user222

Richard Moore

Richard Moore

Figure 10: Helping Science keying applications

The results of the process can be viewed and downloaded by the collection manager in different formats: Simple CSV files (comma separated values) are good for reviewing the data in Excel or importing into a database or collection management system such as Specify or BRAHMS. For applications capable of consuming XML (extensible mark-up language) documents, the DarwinCore records can be downloaded as such. Finally, the data can be downloaded as a DarwinCore Archive that was introduced by GBIF's Integrated Publishing Toolkit (IPT), which is a zip file of both the XML and CSV data. Such an archive can be loaded directly into an IPT instance or posted online for being harvested by GBIF.

Currently, the Helping Science system is still in beta status. Testing is done with about 10,000 specimen images from six different collections. Specimen range from the early 1800's to the present, encompassing different layers of difficulties. Helping Science is still looking for collections to provide sample images for the testing phase.

4. Comparison

This chapter tries to juxtapose the approaches/projects outlined in the previous chapters. This is possible only to a certain extent, since they differ considerably and even have different goals. Some of the criteria do not even make sense for all approaches; however, this is done for the sake of clarity and to get a quick impression.

	Herbar Digital	**Workflow Optimization**	**ReCAPTCHA**	**Herbaria@home**	**Helping Science**
What is done?	Feature Detection: - Recognition of labels on specimens; - Localisation of text on specimens, whether handwritten or typewritten	Dispatching the task of metadata capture to the most appropriate agent	Capturing data from scanned text than cannot be recognized by OCR software	Capturing metadata from herbarium specimens	Capturing metadata from herbarium specimens
Covering the task of					
I: Identification of objects /labels	Yes	No	No	Yes (implicit)	Yes
II: Separating/classifying pieces of metadata	Yes	No	No	Yes (implicit)	Yes
III: Metadata capture	No	No	Yes	Yes	Yes
IV: Managing the work-flow between tasks	No	Yes	No	Yes	Yes
Type of Knowledge	Algorithms	Approach	Software (web service)	Software (web application)	Software (web application)
Status	In development	Not started yet	In production	In production	Beta status
Academic/Commercial	Academic research	Academic research	Initially Academic (Carnegie Mellon University), now commercial (Google)	Academic	Commercial (Silver Biology)
License Fee	(not applicable)	(not applicable)	No	No	Not in current stage; Later probably Yes
Registration required	(not applicable)	(not applicable)	No	Yes	Yes

	Herbar Digital	**Workflow Optimization**	**ReCAPTCHA**	**Herbaria@home**	**Helping Science**
Current number of volunteers (August 2010)	(not applicable)	(not applicable)	millions	274	(not applicable yet)
Documents captured so far	(not applicable)	(not applicable)	1,2b CAPTCHAs ≈ 440m words ≈ 17.600 books	69.227 herbarium sheets	(not applicable yet)
Usable without development efforts?	No	No	Yes	Yes	Yes
Applicable for own collections?	No	No	No	Yes	Yes

5. Applicability for Current Digitisation efforts

The first two initiatives outlined in chapter 3 describe basic approaches for setting up an optimised workflow for metadata capture or algorithms that can be used to solve problems connected to the automated metadata capture. There is no software that can be used, nor will there be in the near future. They were included in this report just for documenting the underlying ideas.

ReCAPTCHA is a very smart approach to use human deciphering efforts that would be otherwise wasted. However, just the basic idea could be used by other projects and initiatives. Currently, reCAPTCHA is used for digitising the archive of the New York Times and books from Google Books. Even though it would be theoretically possible to use reCAPTCHA for capturing at least parts of the label information for herbarium collections, this would require Google to cooperate. Most probably, only a very large project with sufficient public support could achieve this; realistically, it cannot be used for typical digitisation initiatives.

Herbaria@home and Helping Science are designed directly for capturing metadata of herbarium sheets. It is easy to join Herbaria@home, the project is still looking for collections providing digitised herbarium sheets. There seems to be a number of devoted volunteers, and the rate of specimens documented per months seems to increase slowly; still, it would take years to document medium-sized collections. Helping Science is still in beta status, and after its completion it will likely cost a licensing fee to use the system. Moreover, similar to herbaria@home, there would be a need to mobilise volunteers who do the work.

However, since some of the herbaria are connected to botanic gardens, which often fancy support groups of aficionados, there are a good number of potential volunteer workers. It might be just a question of good public relations work to mobilise them; the use and prominent publication of high score tables could be a good incentive. Yet, there are no experiences for this; a project going this way would definitely have to forge new paths.

References

Steinke, Karl-Heinz; Dzido, Robert; Gehrke, Martin; Prätel, Klaus: „Entwicklung und Untersuchung von Erkennungssoftware für den Einsatz im Projekt Herbar Digital.“ Hannover, 2009. http://nbn-resolving.de/urn:nbn:de:bsz:960-opus-2520

Steinke, Karl-Heinz: “Lokalisierung von Schrift in komplexer Umgebung”. Hannover, 2009. http://nbn-resolving.de/urn:nbn:de:bsz:960-opus-2838

Steinke, Karl-Heinz; Dzido, Robert; Gehrke, Martin; Prätel, Klaus: „Untersuchung von kommerzieller Software für den Einsatz im Projekt Herbar Digital.“ Hannover, 2009. http://nbn-resolving.de/urn:nbn:de:bsz:960-opus-2532

Steinke, Karl-Heinz; Dzido, Robert; Gehrke, Martin; Prätel, Klaus: „Feature recognition for herbarium specimens (Herbar-Digital).“ Erschienen in: Proceedings of TDWG, Perth 2008. http://nbn-resolving.de/urn:nbn:de:bsz:960-opus-2888

Ahn, Luis von , Benjamin Maurer, Colin McMillen, David Abraham, Manuel Blum: „reCAPTCHA: Human-Based Character Recognition via Web Security Measures“. Science Magazine, Issue 321, 12 September 2008, pp. 1465-1468. Published online 14 August 2008 [DOI: 10.1126/science.1160379]

Giddens, Michael: “Using Citizen Science to Process Digital Herbarium Labels”. In: Proceedings of TDWG, Montpellier, 2009. http://www.tdwg.org/proceedings/article/view/511

ReCAPTCHA website: http://recaptcha.net/

Herbaria@home: http://herbariaunited.org/atHome/

Helping Science: http://www.helpingscience.org/