Biodiversity Informatics, 9, 2014, pp. 18-29

DIGIWEB – A WORKFLOW ENVIRONMENT FOR QUALITY ASSURANCE OF TRANSCRIPTION IN DIGITIZATION OF NATURAL HISTORY COLLECTIONS

TERO MONONEN, RIITTA TEGELBERG, MIRA SÄÄSKILAHTI, MARKKU A. HUTTUNEN, MARKO TÄHTINEN AND HANNU SAARENMAA*

Digitarium, SIB Labs, University of Eastern Finland, Länsikatu 15, FI-80101 Joensuu, Finland
**corresponding author: hannu.saarenmaa@uef.fi*

Abstract – Data produced by digitization increases the scientific use of natural history collections. However, in mass digitization, attention must be paid to the flawless management of the workflows, and high quantities of end results should not be compromised by a low standard of quality. A web-based environment *DigiWeb* was created for controlling the workflow of transcribing data from images of natural history specimens. Using *DigiWeb*, it was possible to manage the workflow of transcription and data proofing, include all participants to the workflow, allow collaboration and training, and also to provide useful processing features. The data emerging from this process pass quality control standards which are supported by *DigiWeb* and based on the strict requirements of the ISO 2859 standard.

Keywords – data entry, mass digitization

INTRODUCTION

Collections conserved by natural history museums are an important source of information on taxonomy, biodiversity and environmental change. Digitization of these collections and the practices of providing open access data are expected to improve the world-wide utilization of museum data. Recent advances in technology (e.g. Schmidt et al. 2012; Heerlien et al. 2013; Tegelberg et al. 2014) offer solutions for increased efficiency in the imaging of natural history specimens. These automated imaging pipelines are now producing huge amounts of digital content in many projects, and thus there is an urgent need to develop new solutions for streamlining the transcribing of data from images. Such data entry is still mostly based on manual, labor-intensive work. However many approaches are being tried to modernize transcription. Optical character recognition (OCR) is a promising method for type-written material (e.g. Haston et al. 2012; Tulig et al. 2012). Crowdsourcing (Flemons and Berents 2012; Hill et al. 2012; Herbaria United 2014; Les Herbonautes 2014) has lately gained much attention, but is not an appropriate solution for in-house and time-bound project work.

The digitization of biological and geological collections can be described as a process or workflow, containing steps such as transportation, tagging, imaging, transcription and archiving (e.g. Dou et al. 2011, 2012; Lehtonen et al. 2011; Nelson et al. 2012; Tegelberg et al. 2012). The methods used in some of these steps may vary depending on specimen type. For example, the physical dimensions of the objects strongly affect the imaging phase (Tegelberg et al. 2014). The management of mass digitization is however particularly sensitive to abnormalities in workflows and when digitizing different specimen types, the differing steps of alternate solutions may increase the risk of error if not carefully controlled.

Efficient management of the digitization process can be achieved by use of an information system that is designed not only for controlling the workflows, but also to promote the quality assurance of results. For example, based on a survey among digitizers, automatic filtering of data by country and collector is expected to reduce mistakes made during data entry (Drinkwater et al. 2014). Such a gentle change in digitization workflow leads to familiarity with the geography or handwriting of a collector. This allows the

digitizer to concentrate on specific parts of data entry, with good results.

During the process, the collaboration of partners and specialists is required. When using a well-established imaging method for specimens, scientific expertise is particularly important for transcribing specimen data. Fast feedback and the formation of a community of specialists around a digitization project may enhance the quality of the outcomes. The transcription would also benefit from a knowledge base of the results of earlier digitization projects, available both in-house and on the web. For example, the sometimes rather cryptic handwriting of certain collectors might already have been cracked by a devoted scientist, so providing helpful material for transcribers.

Digitarium is a digitization center providing services to museums of natural history (Tegelberg et al. 2012). The basic idea is to outsource digitization from museums to a factory-like setting, where entire collections are processed, and all steps are automated as much as possible. This idea was first implemented by the commercial company Océ for the herbarium of the natural history museum in Paris (Pignal & Michiels 2012). This was a real break-through, the birth of mass-digitization, which led to increased focus and funding for digitization worldwide. Mass-digitization of herbaria is now underway also in Leiden (Heerlien et al. 2013), Helsinki (Tegelberg et al. 2014), and Oslo. Digitarium has expanded the concept by not only providing a commercial mass-digitization service, but also research and development on industrial engineering and biodiversity informatics, tackling the bottlenecks of the whole digitization process in a real production environment.

This paper discusses the management of the steps required for selecting the specimens for data entry, transcribing the labels from the images, and validating (quality control) the result. The aim was to develop a web-based environment that allows a gradual streamlining of the whole digitization process. Especially, the process is intended to be open to all participants of the digitization projects concerned, promote collaboration, and include the use of helpful features which lead to a high-quality result.

Name	Size	Type
20131111_060436.O-V2190095	4.1 kB	folder
20131111_060451.O-V2190094	4.1 kB	folder
20131111_060525.O-V2190093	4.1 kB	folder
20131111_060530.O-V2190092	4.1 kB	folder
20131111_060712.O-V2190091	4.1 kB	folder
20131111_060742.O-V2190090	4.1 kB	folder
20131111_060755.O-V2190089	4.1 kB	folder
20131111_060822.O-V2190088	4.1 kB	folder
20131111_060845.O-V2190087	4.1 kB	folder
20131111_060857.O-V2190086	4.1 kB	folder
20131111_060934.O-V2190082	4.1 kB	folder
history	4.1 kB	folder
locks1	4.1 kB	folder
20131111_060436.O-V2190095.xml	3.6 kB	XML documer
20131111_060451.O-V2190094.xml	3.1 kB	XML documer
20131111_060525.O-V2190093.xml	3.7 kB	XML documer
20131111_060530.O-V2190092.xml	3.1 kB	XML documer
20131111_060712.O-V2190091.xml	2.7 kB	XML documer
20131111_060742.O-V2190090.xml	2.3 kB	XML documer
20131111_060755.O-V2190089.xml	1.6 kB	XML documer
20131111_060822.O-V2190088.xml	1.4 kB	XML documer
20131111_060845.O-V2190087.xml	1.4 kB	XML documer
20131111_060857.O-V2190086.xml	2.1 kB	XML documer
20131111_060934.O-V2190082.xml	2.1 kB	XML documer

Figure 1. Files and folders – the DigiWeb data store.

METHODS

Preparatory Steps

The first steps in Digitarium's digitization process are specimen labeling and imaging (Lehtonen et al. 2011). For each specimen, a unique identifier (ID) is generated, and a physical label with the ID is attached to the specimen. The labeling process enables different types and sizes of labels, with variable contents to be generated. All of the information concerning the generated IDs and their hierarchy are stored in a MySQL database, which contains all the necessary data to control the process. The labeling and imaging methodology used at Digitarium in the mass-digitization of herbarium sheets and insect specimens has been explained in more detail (Tegelberg et al. 2014).

The actual digitized data is not stored in the MySQL database, but as image files and XML documents. This allows version control through the

various phases of the digitization process better than a relational database (Fig. 1). The XML file has a unique ID and also a corresponding web address created for the digital object corresponding to the specimen, to which data can be later uploaded. The folder contains all the images which relate to the same ID, and the metadata of the specimen is contained in the XML file according to the Darwin Core standard.

The DigiWeb Environment

DigiWeb is a web application created for the browsing of produced images and data, and for transcribing and verifying the images metadata. It relies on the Digitarium file system hierarchy of imaged material, and also on a database which contains the hierarchy of labels and multiple assistant data sets. *DigiWeb* is a Java Enterprise Edition (EE) application, and the user interface utilizes Java ServerFaces 2 (JSF2) and PrimeFaces frameworks to implement rich standalone software-like features in a browser. *DigiWeb* has a built-in version control system of saved metadata where the newest version of the metadata is considered as the primary one. Therefore, all versions are available to trace the accumulation of data, and a rollback function to a specific version is available.

DigiWeb is divided into three main views: browsing, transcribing and administration. On the browsing page, a user can view the collections and specimens. With permission, the user may reserve specimens to form a queue of data to be transcribed. On the transcribing page, the user is able to transcribe the data from the queue of specimens. The administration page contains the administrative functions of *DigiWeb*, e.g. user management, user monitoring and a function that enables all the reports generated by the system to be viewed. In user management, users are allocated a role which defines the views they can access and the functions they can use. The roles are:

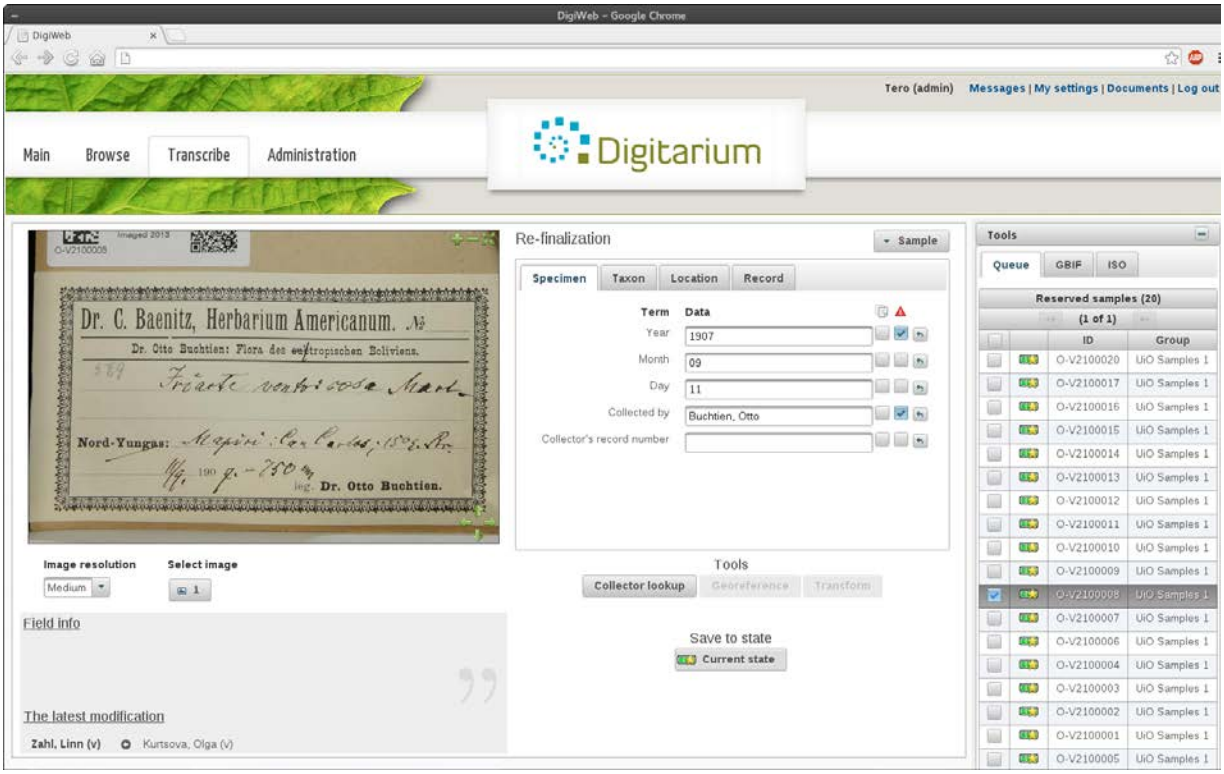

Figure 2. The user interface of DigiWeb.

administrator, in-house participant and client. Users with an administrator role have access to all collections and all functionalities. The in-house role is for those users who are involved in the production and entering of metadata. Each in-house user has customized access rights based on their skills and experience. As a result, users can be divided into groups of standard staff and experienced staff. The client role is for those not participating in everyday data entry work but who need to monitor the process.

On the transcribing page, the view is divided into two parts (Fig. 2). First are the specimen images which can be zoomed and panned. The image viewer uses standard JPEG images as an input, which are served by the Digitarium API (Application Programming Interface). The second part of the view contains all of the input fields grouped into tabs. The users required that the image and the input text fields must be visible simultaneously. The data entry region is dynamically generated according to the structure and components defined in the separate XML file. It supports a full list of Darwin Core terms (Taxonomic Databases Working Group 2014) but can also be customized case specifically by hiding terms that are not needed.

External Lookups and Web Services

In *DigiWeb*, for every Darwin Core term there is a configured input component in the user interface. Supported components are the input text, input text area, auto-complete input text and a drop-down menu. For example, the entry of plant species names is implemented by an auto-complete input text component, including a suitable name suggestion feature. Finalization of the name can be done by choosing the right name from the list, which is then automatically added to the text area. The taxonomic list of species names is taken from The Plant List (http://www.theplantlist.org/) and contains genera that represent families which are currently being digitized. In future, also authority lists of names of other organism groups will be included. In a similar way, the country (and in some cases a smaller geographical area) representing the location of the collection effort can be chosen from a selection list, after inputting the first letters of the location. This look-up list originates from the University of Oslo. A look-up function for a collector's name from a list showing historical and present plant collectors around the world is also supported by *DigiWeb*. The name list has been created by Harvard University, and in *DigiWeb* it aims to help the transcriber when the handwriting of the collector is difficult to translate.

DigiWeb uses the collecting locality service (Tähtinen et al. 2014) developed by the BioVeL project (http://www.biovel.eu). This is a RESTful web service using a JSON data transfer format with fields based on the Darwin Core standard. This service has been published in the Biodiversity catalogue (Tähtinen 2014) for general use. The idea of this service is to provide an easy way to access and reuse existing locality data, which may also include geo-reference information. As a source of information, it can serve the local database of already digitized specimens, and also GBIF's global data portal where there are currently over 500 million records. Algorithms such as fuzzy text search can be used to find the most probable collecting locality of a specimen, by giving the collector's name and additional information, in particular, the collection date. Possible collecting localities are shown on the *DigiWeb* interface and the user may then choose from the options displayed. Localities cannot be saved manually in separation of the specimens, but once the specimen data has been saved by DigiWeb, any new collecting locality information becomes available for transcribing labels of further specimens.

Test Case Oslo

In 2013-2014, *DigiWeb* was used in a project to transcribe data from the Herbarium of the University of Oslo. In this project, the management of workflows by *DigiWeb* was developed for a large number of specimens and additionally, methods for the quality assurance of data were created. During the project, a total of 168,527 herbarium specimens were imaged. Transcription began during the imaging process and it will continue until the end of 2014. In practice, an image became available for transcription within

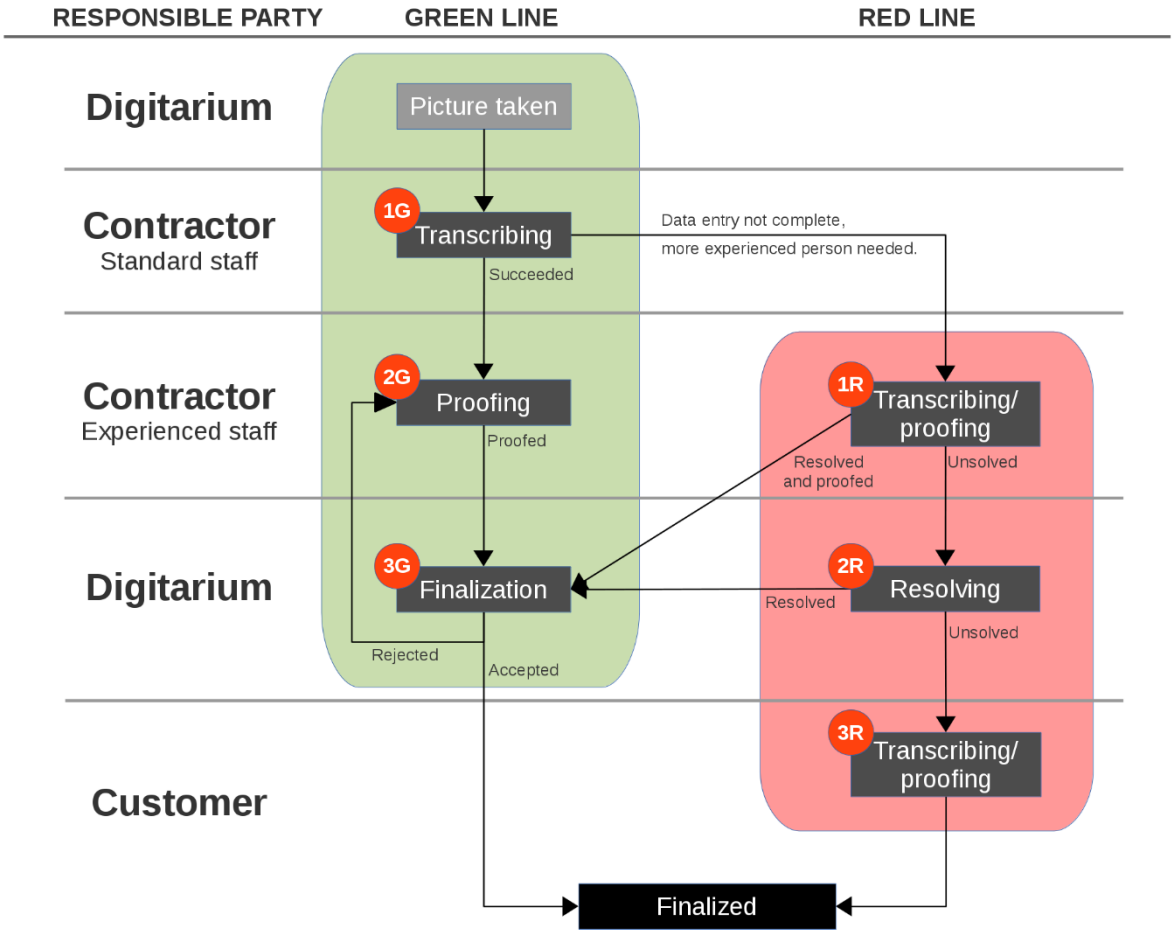

Figure 3. The workflow developed for quality control.

five minutes of its creation. The contracting company DigForsk AS performs transcription in five digitization centers in northern Norway. Up to 30 distance workers simultaneously use *DigiWeb* to transcribe the data from the images produced at Digitarium.

The transcribing process began by conducting a training session at the DigForsk premises. Aimed at the leaders of data entry centers, the training was conducted by the project manager, a botanical expert and a software developer working at Digitarium. In addition, additional training was offered by taxonomists from the University of Oslo. The training covered the use of the web-based transcription tool, reading and understanding old labels of botanical samples, and the accurate and uniform way to transcribe the data. After training, there was a practice period of two-four weeks, based at the digitization centers. During that time, transcribers in Norway worked in a test environment of *DigiWeb*, and specialists at Digitarium offered feedback on the quality of their work through *DigiWeb* and by e-mails. A person was only allowed access to the production environment of *DigiWeb* after approval by specialists based at Digitarium.

RESULTS

The Workflow

The functionality of the program in managing workflows was not found to be dependent on the type of specimens or the imaging methods used. Thus, the program can use any data that follow the file system structure and format used at Digitarium.

The developed workflow for transcription and verification using *DigiWeb* is shown in Fig. 3. The workflow contains altogether seven specimen states in two lines. The state of the specimen contains two types of information: difficulty and step. Difficulty is indicated by color where green (G) represents an easy specimen and red (R) represents a complex specimen. The step in the transcribing process is indicated by the number of spots in the symbol - for example 1G means one green spot, i.e., initial transcribing. The state transitions of a specimen are finalized when no modifications need to be made and data is ready to be delivered to the customer.

The default route in transcribing workflow is 1G, 2G, 3G, and then finalized. In specimen state 1G, the initial transcription is done by non-professional, possibly inexperienced staff, who pass the specimen to the next level. In specimen state 2G, more experienced staff proof the specimens. They pass the specimen to state 3G, which indicates that transcription is finished and ready for the final quality check.

In the initial transcription (1G), a specimen may be forwarded to the red line due to a complex label. Then more experienced staff will check the specimen indicated as 1R. From that state, a resolved specimen may be forwarded back to the green line but if it stays unsolved, it will be passed to 2R. Samples in state 2R are checked by Digitarium's experts. If it seems almost impossible to transcribe, it will continue to state 3R and await inspection by specialists of the client organization.

In a large collection, there may be tens, hundreds or occasionally thousands of specimens collected by the same collector. As a result, the same information will be transcribed several times, by persons with different skill levels. This may create different spellings of the same locality, unnecessary duplication of records, and results in poorly usable data. This can be avoided in DigiWeb by phasing the transcription of different fields. That is, the user does not need to transcribe all fields at once. Instead, the locality data of specimens can be processed in a later phase as a bulk operation after the collector's name has already been entered. In this case, the transcriber already knows the specific handwriting style and is familiar with the visited locations. Consequently the produced data will be more uniform and of higher quality. This idea was preliminarily tested by the scientific curator of JOE (Herbarium of the University of Eastern Finland). During the first round of data entry, the scientific name of the taxon, the collector's name, collector's record number, and the date of collection were transcribed, but locality was skipped. During the second round, the locality information was transcribed. Between the two rounds, the data were re-organized according to the collector and date to facilitate bulk operations. Both the duration of the data entry and quality of the end results were assessed. Results showed that doubling the sample size from 374 to 748 items reduced the portion of collectors' first samples from 54 % to 41 %. It was also found that when the collector's name was successfully transcribed, the average total time spent on the transcription of a specimen using the re-organized data dropped from original 5 minutes to 2 minutes. Thus, the presented workflow could in future be changed to having two 1G states - 1G and 1Gfinal. From a quality assurance point of view, concentrating on the repeating localities increased the validity and uniformity of the metadata produced.

Acceptance procedure

The procedure for quality assurance was developed at the beginning of the test project. In practice, before the data is delivered to the customer, a final quality control check takes place (Fig. 3). When at least 1000 (test batch size) specimens have been marked as finished (state 3G), they are subjected to quality control. The quality check and reports are created by using the *DigiWeb* acceptance test tool which is based on the statistical ISO 2859 standard. In the check, a batch

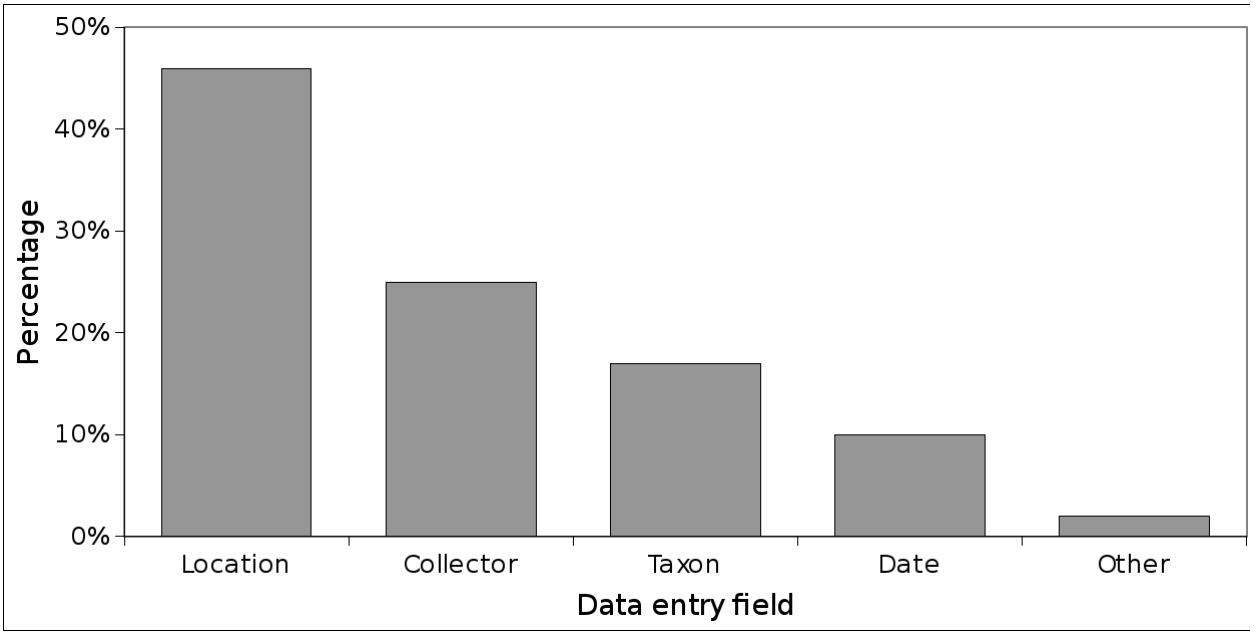

Figure 4. Errors found in transcribed data during the first four months of the test case project. In total, there were 187 rejected specimens.

of 1000 samples is created and 80 randomly selected samples are checked by an expert. If there are ≤3 rejected samples, the whole batch of 1000 samples is accepted. However, in the case of >3 rejected samples, the whole batch (excluding the accepted samples) is returned for proofing (state 2G). The acceptance criteria may vary, depending on the customer demands. Typically, however, at least the country, collector name, scientific name, and catalog number must be free of error, because without them, the specimen cannot be found in a search. In general, small errors in other fields may be tolerated as they can be checked directly from the image.

The sets of specimens accepted in the quality-control process are considered as finalized and can be delivered to the collection owner.

Quality and quantity of transcription

During the test project, excluding the initial practice weeks, the data of around 5000 specimens were transcribed every week. There were on the average 24 different people at work each week who transcribed on the average 42 samples in a 3-hour part-time working day. As transcribing progressed, this allowed proofing of the metadata to start. On average, about 8000 specimens were proofed per month. This resulted in some degree of backlog; however, this will be cleared at the end of the project by people gathering experience on transcription. On average, the transcription and proofing of the label data of a specimen lasted about five minutes, however this varied significantly between specimens. Typed and printed labels were relatively fast to transcribe but some of the hand written labels went through the whole of the 'red line' of the workflow (Fig. 3), thus expanding the time spend on that particular specimen.

At Digitarium, the quality control of data entry was centralized to a specialist to ensure the uniformity of decisions of acceptance. The requirements in quality control were created in co-operation with the experts in the customer organization. Reports made by the quality controller by using *DigiWeb* showed the main errors that led to rejection of the data were missing information (especially some or all of the locality

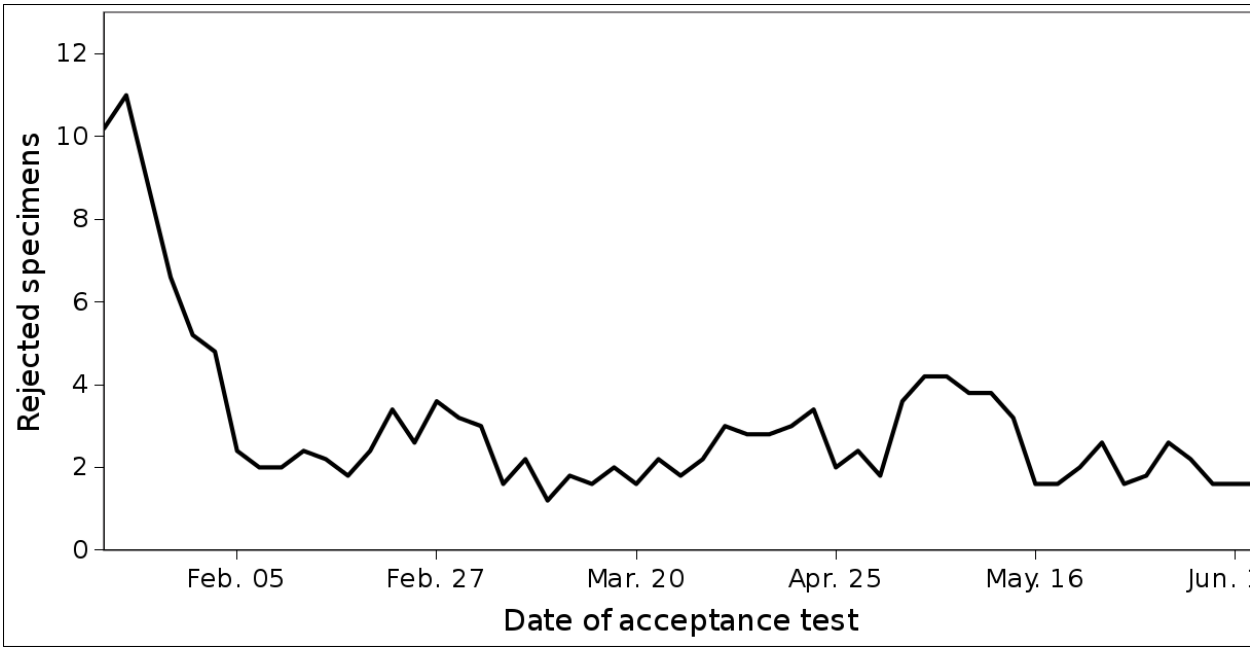

Figure 5. Numbers of specimens rejected by the quality controller during the first four months of the test case project.

information) and incorrect collector name (Fig. 4). Problems with taxa were also found, often connected to markings expressing hybrids and determiner's, or other expert's, doubts (e.g. "cf.") about the identification of the taxon. The quality of transcription was followed from the beginning of the project, and problems were identified in applying DwC terms and syntax correctly. Therefore, a new feature was introduced into *DigiWeb*. Any deviation in the presentation style of year, month, day, or collector's name automatically caused the input text to turn red. The color did not change to black until the term was presented according to the style defined by DwC. This feature was especially helpful in producing uniform data concerning the dates of records.

In general, of the batches of 1000 specimens, about 50% were accepted during the first round of quality control. This percentage increased gradually over the first six months. Fig. 5 presents the amounts of rejected samples in all batches checked by the quality controller during the first months of the test case project. The clear decrease in errors found can be explained by the increasing experience of data entry personnel, as well as thanks to use of the help features provided in the system.

Communication

In order to enhance rapid communication between digitization centers which were spatially distributed in northern Norway, Digitarium (Joensuu, Finland) and the customer (Oslo, Norway), and to lower the transcriber's threshold to ask questions from the experts, a built-in messaging system was implemented in *DigiWeb*. Such an integrated communication system was found to be an easy way to send questions and answers, without knowing individual contact information like e-mail or Skype addresses. General information was delivered to all *DigiWeb* users by using the news section in the main page, and included information and guidance on new *DigiWeb* features. To ensure that all transcribers had easy access to the most recent versions of all documents (e.g., data entry manual, guidelines and learning material), a document directory was also integrated into the system.

DISCUSSION

In *DigiWeb*, attention has been paid to the support of managing the digitization of large collections. Based on the test project, the management of large scale data entry work is possible using *DigiWeb*, and is not dependent on the whereabouts of the participants of the project. In addition, *DigiWeb* can be used as a training tool and an information source when pursuing a high level of quality in produced metadata. Recent improvements in *DigiWeb*, combined with temporary spreadsheet assistance, made it possible to divide the current transcription workflow into two partial rounds. This enabled sequential entry of repeating information, with the aim of increasing the quality of locality data and improving the results of geo-referencing for end users. It is acknowledged, however, that further testing of the proposed workflow is still needed, and the ability to use *DigiWeb* efficiently in phased transcription needs still some refinement.

The workflow developed here has special characteristics which derive from the requirement of the customer to support ISO 2859 based quality control, and from the use of large number of (initially) inexperienced transcribers. This necessitates two validation steps, first by the team leader at DigForsk AS and then at Digitarium. This is relatively expensive, and can only be defended by the fact that the basic transcribers' work cost is subsidized by the employment office. Compared with workflows developed elsewhere, there are two human data curation steps. The Kurator workflow (Dou et al. 2011, 2012) only contains one, and includes more automation. In crowdsourcing projects, the human curation steps have entirely been replaced by repeated transcription of the same samples (Flemons et al. 2012; Hill et al. 2012): When there are enough repetitions that match, the result is automatically accepted. What is "enough" repetition has been studied by Shah (2014) who developed a consensus model for aligning the differing transcriptions. A digitization workflow tool such as DigiWeb would ideally support several quality assurance methods.

The labels of specimens stored at natural history museums contain a set of basic information. However, the specimen data may be presented in a personalized way. For example, old handwriting, invalid locality names, and the variable uses of Latin will cause problems to the transcribers and translators of the label information. Therefore it is important that the tool used for transcription is easy to learn, easy to use, and provides help when possible to the user. In the case of *DigiWeb*, the functionality of the system has been tested with both academics and also those persons without any academic education but with reasonable IT-skills. This work is on-going, with new ideas constantly emerging.

A fundamental question for quality assurance is whether we want a literally accurate transcription of what is written in the label, or its interpretation, where differing spellings, ancient locality and taxon names are harmonized. If we can afford it, both would be nice. Literally accurate transcription would ideally need to be saved because the interpretation can go wrong, and because of its cultural history value. On the other hand, when images are available, they serve as the literally accurate information to fall back when in doubt. Multiple, slightly differing transcriptions showed to be problematic in this project, and our opinion now is to avoid them, and only save modern interpretations. However, such interpretations can only be made by relatively experienced staff, and not inexperienced workers or volunteers. In conclusion, aligning repeated transcriptions needs to be studied more.

DigiWeb uses Darwin Core as the standard and basis of data terms. The aim of DwC is to facilitate the sharing and alignment (integration across records) of information found in the specimen labels. DwC still has some shortcomings in support of digitization. There are separate fields for verbatim data which facilitates data entry of the labels literally. However, the verbatim fields do not cover everything. Therefore, a new field for all label information, e.g., "verbatimLabel", might be necessary. In *DigiWeb*, the DwC-standard is being followed precisely and instructions for each term are easily available, with different languages

available if needed. Thus, *DigiWeb* can be used as a training tool and it gently guides the users towards conformance in data entry work.

Tools such as taxonomic lists, geographic hierarchies, and search facilities for collectors' names helped data entry workers to produce data efficiently. In addition, in long series of the same taxon, the ability to select quickly the previously written taxon name with a mouse click made the workflow easier and faster. However, for example, the taxon name lists accepted by the scientific communities are not available for all taxon groups. Indeed, with the exception of *The Plant List*, they appear frequently not to be publicly available: open access to such lists should be promoted. One possibility is that they should be available from the major nomenclators involved in their compilation, which would provide high quality and flawless information. In practice, incorporating tools such as lists in DigiWeb is quick and easy: it is possible to use such tools through the servers at Digitarium or remotely through available APIs.

The interpretation of locality names (especially when presented in Latin) was proven to be a difficult task. Based on a request by the customer, the country was considered as the most important locality information. However, in the labels, country is not always mentioned, and specifying the country was demanding when for example, the only information given was a name of a mountain. According to the results, the most common errors found by quality control often concerned actions for which help was not readily available in the DigiWeb environment. Therefore, the features aimed at helping the transcribers were deemed to be found useful. For geographic locations, gazetteers are publicly available. For geo-referencing, efficient services have been created by GeoLocate (Rios and Bart 2008). However, we must point out that collecting localities are not just any localities, but rather place-collector combinations, repeatedly visited by the same collectors. Therefore, we developed a collection locality service, which also takes into account the collector's name and the time of collecting. By tracing the movements of collectors, we can get more accurate information for geo-referencing. Furthermore, the geographic hierarchies obtained from gazetteers reflect the situation today, and not that of the past. Thus there is a need for establishing historical names and also their periods of validity. If data is pooled and made available, with each new collection that is digitized, the supporting tools become more efficient.

Our experience in streamlining transcription is that much can be gained by using shared lookup services and big pools of data that are already available. Samples must be distributed to the best available agent with regard to language, handwriting, taxonomy, and geography. This distribution should happen automatically in the contemporary electronic marketplace of digitization services. Transcription in isolation is waste of time, and thus more web services are needed. The pooling of data is making this degree of distribution and access possible.

The possibility to share knowledge and solve problems in a community is a feature expected to enhance the levels of motivation and skills of those involved. On the other hand, spreading important information to all users at the same time may encroach upon working time. For the "*DigiWeb* community," releasing news for example about new features was important. *DigiWeb* was also used for showing the results of final validation, and for commenting on mistakes in data entry. This allowed all partners to recognize the specific problems in transcription and work together to find solutions to them.

The data delivered by such workflows needs to be reliable enough to meet the standards of science. In the validation of data, human resources are needed. A person may have specialized in the taxonomy of certain families or in the handwriting of a collector from the 19th century; however, whether such persons exist in every organization is questionable. Therefore access to the databases of earlier digitized collections might prove helpful. In this project, embedding access to other digital contents was worthwhile, especially when the handwriting of the collectors was poor. In these cases, after resolving the collector's name, locations could often be discerned by tracking down the collector's path by following the dates of

the collecting events. Such "natural history intelligence," where new facts are derived from examining pooled information, is actually common practice in museums.

When we understand that transcription of the entire specimen data in one pass may not be the right thing to do, we are close to making a fundamental conclusion: digitization is annotation. The digitization of scientific objects will never be finished, since new facts and measurements which relate to them are always emerging. An information system for digitization must therefore support annotations. Ideally, annotations can be inserted into distributed databases, wherever the specimen data may be found. These issues have been explored by the Annosys (Tschöpe et al. 2013) and Filtered-Push (Morris et al. 2009) projects. Digitization can also be seen as an asynchronous workflow that can span over decades. For instance, when important details such as geographic coordinates or a new identification have been annotated to a specimen, this may trigger workflows that push related data to other related specimens. Another workflow can then notify the curators to validate such annotations.

The development of *DigiWeb* will continue in co-operation with its users. As more lookup services emerge, these will be included as new features. Finally, automatic data entry and geo-referencing based on previously resolved label information will also be introduced as part of the workflows.

ACKNOWLEDGEMENTS

The development of DigiWeb has been financed by the European Social Fund (grant no. 703990) and European Regional Development Fund (grant no. 806527), and by the BioVeL project (the European Union's Seventh Framework Programme for research, technological development and demonstration, grant no. 283359). We thank the staff of the University of Oslo and DigForsk AS for fruitful collaboration.

REFERENCES

Dou, L., D. Zinn, T. McPhillips, S. Köhler, S. Riddle, S. Bowers and B. Ludäscher. 2011. Scientific workflow design 2.0: Demonstrating streaming data collections in Kepler. International Conference on Data Engineering 2011. doi.ieeecomputersociety.org/10.1109/ICDE.2011.5767938

Dou, L., G. Cao, P.J. Morris, R.A. Morris, B. Ludäscher, J.A. Macklin and J. Hanken. 2012. Kurator: A Kepler package for data curation workflows. Procedia Computer Science 9: 1614-1619. doi: 10.1016/j.procs.2012.04.177

Drinkwater, R.E., R.W.N. Cubey and E.M. Haston. 2014. The use of optical character recognition (OCR) in the digitization of herbarium specimen labels. PhytoKeys 38: 15-30. doi: 10.3897/phytokeys.38.7168

Flemons, P. and P. Berents. 2012. Image based digitization of entomology collections: Leveraging volunteers to increase digitization capacity. ZooKeys 209: 203-217. doi: 10.3897/zookeys.209.3146

Haston, E., R. Cubey, M. Pullan, H. Atkins and D.J. Harris. 2012. Developing integrated workflows for the digitization of herbarium specimens using a modular and scalable approach. ZooKeys 209: 93-102. doi: 10.3897/zookeys.209.3121

Heerlien, M., J. van Leusen, S. Schnörr and K. van Hulsen. 2013. The natural history production line. In: Digital Heritage International Congress (DigitalHeritage), Oct. 28 – Nov. 1. Vol. 2, pp. 289-294. Marseille, France. doi: 10.1109/DigitalHeritage.2013.6744766

Herbaria United. 2014. Accessed at http://herbariaunited.org/atHome 2.6.2014.

Hill, A., R. Guralnick., A. Smith, A. Sallans, R. Gillespie, M. Denslow, J. Gross, Z. Murrell, T. Conyers, P. Oboyski, J. Ball, A. Thomer, R. Prys-Jones, J. de la Torre, P. Kociolek and L. Fortson. 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. ZooKeys 209: 219-233. doi: 10.3897/zookeys.209.3472

Lehtonen, J., S. Heiska, M. Pajari, R. Tegelberg and H. Saarenmaa. 2011. The process of digitizing natural history collection specimens at Digitarium. In: Jones M. B. and C. Gries (eds) Proceedings of the Environmental Information Management Conference 2011 (EIM 2011). September 28-29, 2011. Santa Barbara, CA. University of California, pp. 87-91.

https://eim.ecoinformatics.org/eim2011/eim-prodeedings-2011. doi: 10.5060/D2NC5Z4X

Les Herbonautes. 2014. Accessed at http://lesherbonautes.mnhn.fr/ 2.6.2014.

Pignal, M. and H. Michiels. 2012. Switching to the fast track: Rapid digitization of the world's largest herbarium. Botany 2011 – Colombus, Ohio Marc Pignal, Henri Michiels 11th of July, 2012. http://collections.mnhn.fr/wiki/attach/Visit_October2012/Paris-Herbarium-Digitization_2012-07-12.pdf

Morris, P.J., M.A. Kelly, D.B. Lowery, J.A. Macklin, R.A. Morris, D. Tremonte and Z. Wang. 2009. Filtered Push: Annotating distributed data for quality control and fitness for use analysis. American Geophysical Union, Fall Meeting 2009, abstract #IN34B-08. http://adsabs.harvard.edu/abs/2009AGUFMIN34B..08M

Nelson, G., D. Paul, G. Riccardi and A.R. Mast. 2012. Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19-45. doi: 10.3897/zookeys.209.3135

Rios, N. and H.L. Bart. 2008. Community Building and Collaborative Georeferencing using GeoLocate. In: Weitzman, A.L., and L. Belbin, (eds). Proceedings of TDWG (2008), Fremantle, Australia. http://www.tdwg.org/fileadmin/2008conference/documents/Proceedings2008.pdf#page=46

Schmidt, S., M. Balke and S. Lafogler. 2012. DScan – a high-performance digital scanning system for entomological collections. ZooKeys 209: 183-191. doi: 10.3897/zookeys.209.3115

Shah, M. 2014. Accuracy assessment of crowdsourced data in biological specimen transcription. Master's Thesis 58 p., 2 appendix (13 p.), University of Eastern Finland, School of Computing, Joensuu. http://epublications.uef.fi/pub/urn_nbn_fi_uef-20140793/urn_nbn_fi_uef-20140793.pdf

Tähtinen, M. 2014. Collecting locality web service in BiodiversityCatalogue. Accessed at https://www.biodiversitycatalogue.org/services/62 16.5.2014.

Tähtinen, M., T. Mononen and H. Saarenmaa. 2014. Workflows for automation of digitisation of biological collections. 57-59. In: Saarenmaa H (ed.). Use cases, workflows, benchmarking, and related sprints in BioVeL June 2013 - February 2014. Capacities Programme of Framework 7: EC e-Infrastructure Programme, e-Science Environments - INFRA-2011-1.2.1 BioVeL - Biodiversity Virtual e-Laboratory. Deliverable Report D2.4. 78 p. European Commission. http://www.biovel.eu/images/publications/InternalDocuments/D2.4-ReportAndDocumentationOfSprints-Final-28February2014.pdf

Taxonomic Databases Working Group. Darwin Core. 2014. Accessed at http://rs.tdwg.org/dwc/ 17.4.2014.

Tegelberg, R., J. Haapala, T. Mononen, M. Pajari and H. Saarenmaa. 2012. The development of a digitising service centre for natural history collections. ZooKeys 209: 75-86. doi: 10.3897/zookeys.209.3119

Tegelberg, R., T. Mononen and H. Saarenmaa. 2014. High performance digitization of natural history collections: automated imaging lines for herbarium and insect specimens. Taxon.

Tschöpe, O., L. Suhrbier, A. Güntsch and W.G. Berendsohn. 2013. Annotating biodiversity data via the Internet. Taxon 62: 1248-1258. doi: http://dx.doi.org/10.12705/626.4

Tulig, M., N. Tarnowsky, M. Bevans, A. Kirchgessner and B.M. Thiers. 2012. Increasing the efficiency of digitization workflows for herbarium specimens. ZooKeys 209: 103-113. doi: 10.3897/zookeys.209.3125