The Process of Digitising Natural History Collection Specimens at Digitarium

Juha Lehtonen, Susanne Heiska, Mika Pajari, Riitta Tegelberg & Hannu Saarenmaa
Digitarium - Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland
Faculty of Science and Forestry, Joensuu Science Park
Länsikatu 15 (P.O. Box 111), FIN-80101 Joensuu
www.digitarium.fi hannu.saarenmaa@uef.fi

***Abstract*— Digitarium is a joint initiative of the Finnish Museum of Natural History and the University of Eastern Finland. It was established in 2010 as a dedicated shop for large scale digitisation of collections. The paper gives an overview of the steps of digitisation process, including tagging, imaging, data entry, georeferencing, filtering, validation, publishing, and archiving. A functional model is presented. The work at Digitarium is independent of any collection management software. Instead, the digitisation process is managed through XML-documents and versioning. All specimens are imaged and distance workers take care of the digitisation from the images. Data and images are published through Morphbank and GBIF.**

Keywords— digitisation; imaging; natural history collections; XML.

I. INTRODUCTION

Digitisation of specimens in natural history collections is a huge challenge (cf. [1]) . A total of 2-3 billion specimens has been estimated to exist in natural history museums worldwide, and less than 5% of them have been catalogued digitally until now. A much smaller proportion has been imaged. In Finland, the six largest public museums contain an estimated 22 million specimens, out of which 12% has been digitally catalogued, i.e. minimally digitised. In addition, private collections contain up to 8 million specimens.

In the national digitization strategy of natural history collections [2] it has been estimated that the required effort to digitize most of these holdings is 750-1000 person years. This estimate is based on the rate of 100 samples per day for each worker. Such rate has been reported in well-organised digitization projects [1, 2] with easy materials. However, the current efficiency in most digitisation projects still seems to be around 20-30% of this optimum. These rates need to be improved, if digitisation at all is going reach its goals. The improvement should come from moving from hand-crafting to industrial-scale assembly lines and workflows.

As a response to this challenge, Digitarium, the Digitisation Centre of the Finnish Museum of Natural History and the University of Eastern Finland was established in 2010. Digitarium implements the national digitisation strategy for natural history collections [2], and aims at speeding up of digitisation through an efficient production line and knowledge management.

This document outlines the process of digitisation as it is being implemented at Digitarium, and the related ICT support. Not all the steps are fully in place at this writing, but the production line works, and is being streamlined and tuned up.

Special features of the process at Digitarium are imaging of all material, distributed workflow that can employ distance workers, and XML (Extensible Markup Language) based data management. The process is for the first time being described here.

II. PROCESS STEPS

The below steps normally occur in sequence in this order, and are driven by the JJC tool, which is described below. They are illustrated in the functional model of Fig 1.

A. Receiving

Digitarium does not manage its own collections, but is a shop for digitising materials from "customers", i.e., museums and other institutions located elsewhere. Therefore, delivery of material is the first step. An agreement is made with the sending institute of the material that will be received, and in what detail and timeframe it will be processed. After receiving, material is subjected to deep-freezing to eliminate any pest organisms. A metadata entry is made about the received material and agreements.

B. Tagging

Each sample will be tagged with a globally unique identifier in the of form of an HTTP URI, for example http://id.luomus.fi/GA.105. This namespace is managed by the Finnish Museum of Natural History. The URI is resolvable to the specimen details. The last part of the URI will also be written in a 2-dimensional barcode. The tag will be glued to the paper sheet or pinned in the needle of an insect sample. The labels from insect specimens are removed and placed temporarily on a sheet of cardboard for imaging.

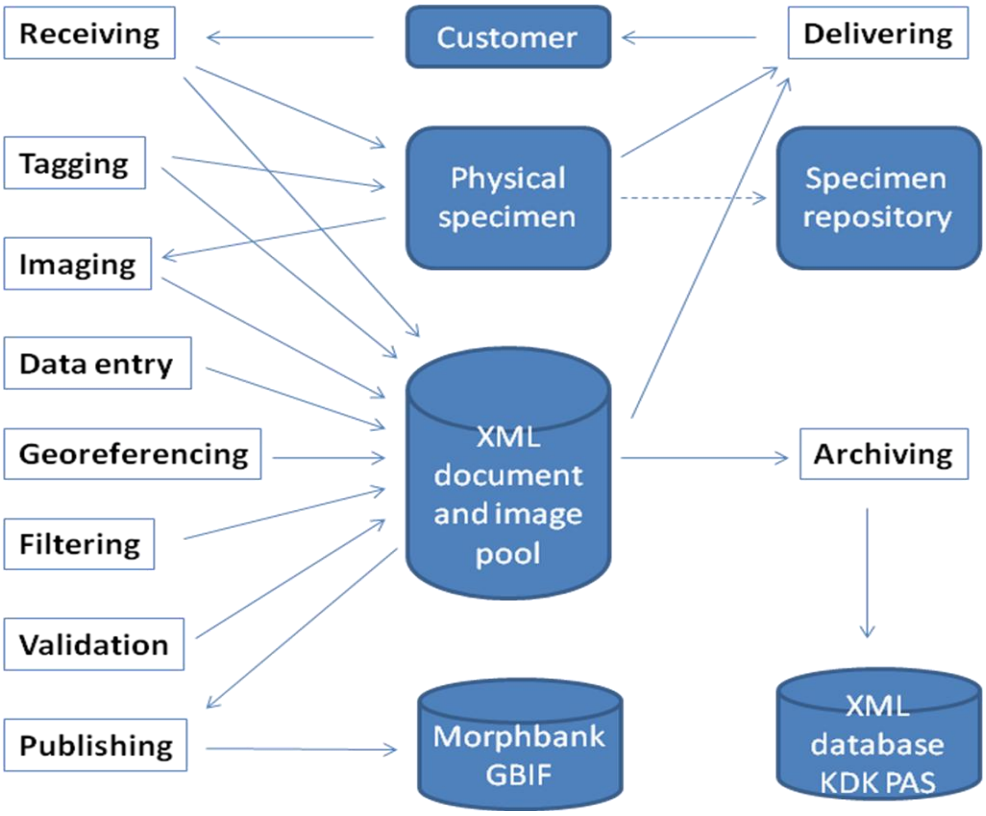

Figure 1. Functional model of the digitisation process.

Figure 2. An example of the result of imaging the lower half of a plant sheet, which is the source of the data in Fig. 3.

C. *Imaging*

Several pictures are made of the sample with a high-end digital camera. The cameras have resolution of 24 megapixels and produce TIFF images of 75 MB. A plant sheet is imaged in two pieces (Fig 2.), which gives a resolution of 450 dpi over the entire sheet. The two pieces are later joined programmatically. In the case of insect samples, the specimen and the labels are imaged separately. Scanners are not employed. Details of the imaging event and results are stored in an XML document automatically.

```
<?xml version="1.0" encoding="UTF-8"?>
<dwr:DarwinRecordSet
xmlns:xsi="http://www.w3.org/2001/XMLSchema"
xsi:schemaLocation="http://rs.tdwg.org/dwc/dwcrecord/
http://rs.tdwg.org/dwc/xsd/tdwg_dwc_classes.xsd"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
xmlns:dwr="http://rs.tdwg.org/dwc/dwcrecord/">
<dwc:Occurrence>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
    <dwc:associatedMedia>./EA9.00C/Image001.tif;
./EA9.00C/Image002.tif; ./EA9.00C/Preview001.jpg;
./EA9.00C/Preview002.jpg</dwc:associatedMedia>
    <dwc:recordedBy>Laine Matti</dwc:recordedBy>
    <dwc:preparations>dry</dwc:preparations>
    <dwc:individualCount>1</dwc:individualCount>
    <dwc:disposition>in collection</dwc:disposition>
    <dcterms:type>PhysicalObject</dcterms:type>
    <dcterms:modified>2011-05-04; 2011-03-31</dcterms:modified>
    <dcterms:creator>Pennanen, Marja (d); Lemmetyinen, Juha
(i)</dcterms:creator>
    <dcterms:contributor>Digitarium</dcterms:contributor>
    <dcterms:language>FI</dcterms:language>
    <dwc:basisOfRecord>PreservedSpecimen</dwc:basisOfRecord>
</dwc:Occurrence>
<dwc:Event>
    <dwc:eventID>http://id.luomus.fi/EA9.00C</dwc:eventID>
    <dwc:fieldNumber>G333</dwc:fieldNumber>
    <dwc:eventDate>1986-8-31</dwc:eventDate>
    <dwc:habitat>harjun S-rinteen puoliväli</dwc:habitat>
    <dwc:year>1986</dwc:year>
    <dwc:month>8</dwc:month>
    <dwc:day>31</dwc:day>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Event>
<dwc:Identification>

<dwc:identificationID>http://id.luomus.fi/EA9.00C</dwc:identificatio
nID>
    <dwc:identifiedBy>Laine Matti</dwc:identifiedBy>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Identification>
<dcterms:Location>
    <dwc:locationID>http://id.luomus.fi/EA9.00C</dwc:locationID>
    <dwc:continent>Europe</dwc:continent>
    <dwc:countryCode>FI</dwc:countryCode>
    <dwc:stateProvince>V;Ab</dwc:stateProvince>
    <dwc:municipality>Lohjan kunta</dwc:municipality>
    <dwc:locality>Perttilä; Unionin huoltoaseman E-puolella kantatie
53:n vierellä</dwc:locality>
    <dwc:verbatimCoordinates>668  :33</dwc:verbatimCoordinates>
    <dwc:verbatimLatitude>668</dwc:verbatimLatitude>
    <dwc:verbatimlongitude>333</dwc:verbatimlongitude>
    <dwc:verbatimCoordinateSystem>YKJ</dwc:verbatimCoordinateSystem>
    <dwc:locationRemarks>"Kantatie 53" is changed in the year 1996
to "valtatie 25".</dwc:locationRemarks>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dcterms:Location>
<dwc:Taxon>
    <dwc:taxonID>http://id.luomus.fi/EA9.00C</dwc:taxonID>
    <dwc:taxonRank>variety</dwc:taxonRank>
    <dwc:scientificName>Diphasiastrum complanatum ssp.  ×
zeilleri</dwc:scientificName>
    <dwc:scientificNameAuthorship>(L.) J. Holub;  (Roy)
Kukk.</dwc:scientificNameAuthorship>
    <dwc:genus>Diphasiastrum</dwc:genus>
    <dwc:specificEpithet>complanatum</dwc:specificEpithet>
    <dwc:infraspecificEpithet>zeilleri</dwc:infraspecificEpithet>
    <dwc:taxonRemarks>ssp. means subspecies but zeilleri is a
hybrid. Author names are mispelled.</dwc:taxonRemarks>
    <dwc:occurrenceID>http://id.luomus.fi/EA9.00C</dwc:occurrenceID>
</dwc:Taxon>
</dwr:DarwinRecordSet>

Image001.tif
Image002.tif
Preview001.jpg
Preview002.jpg
```

Figure 3. Example of a sample in XML-document after the data entry phase, describing the plant sheet in Fig 2.

D. *Data entry*

The data from labels is entered manually from images using the JJC tool. The data are stored as written, including any misspellings, abbreviations, etc. into the "Verbatim" fields of the Darwin Core data exchange standard, see http://rs.tdwg.org/dwc/. A new version of the XML-document is generated, and the old version from the previous step is kept. Here, like in all other steps of the process a separate document version is retained.

E. *Georeferencing*

Most specimens do not come with geographic coordinates, and candidates for these will be found automatically using web services such as GEOLocate and those of the Finnish National Survey. This results normally in several choices, which are ranked and stored in the XML file in the Darwin Core field georeferenceRemarks. In case grid coordinates have been given in the sample using the Finnish national system (called "YKJ"), these are automatically converted into geographic coordinates already in the previous phase on data entry, and no further candidates are searched.

F. *Filtering*

Certain details of datasets sometimes need to be filtered out before publishing, because of reasons such as endangered species or the customer requiring an embargo of the material. Such filtering is done automatically based on species name or an agreement stored in the metadata of the dataset. Detailed instructions, but still in draft form, exist for this step [3]. Two versions of the XML file are retained: filtered and unfiltered.

G. *Validation*

After all the preparation above, some steps of which being automatic, a final check is made by an experienced staff member. Any errors in data entry are corrected. Out of the georeferenced location candidates one is chosen, or a new manual search is made, and data is stored in the decimalLatitude, decimalLongitude, and precision fields of Darwin Core. The result of any filtering is checked and masked versions of the images will manually be created.

H. *Publishing*

The data from the latest XML document version and images will be imported to Digitarium's Morphbank database instance and Digitarium's GBIF IPT service. From there they will be published, as agreed with the customer, or if publication has not been agreed, retained for Digitarium's internal use.

I. *Delivery*

The data will be delivered to the customer in the agreed format. The samples will be sent back, unless it has been agreed to keep and curate them at Digitarium's repository.

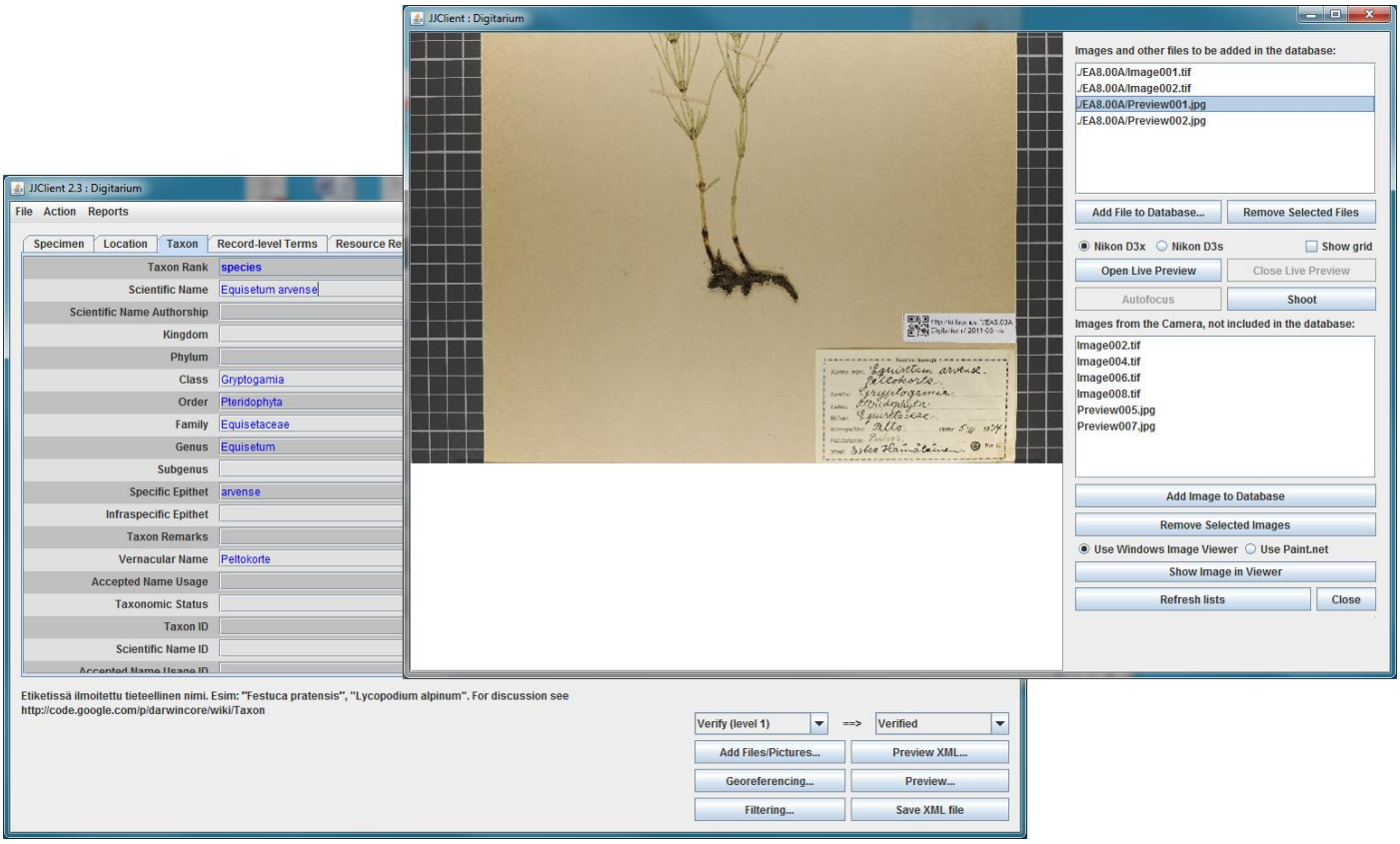

Figure 4. Snapshot of JJC showing some workflow functions.

J. *Archiving*

All the XML documents and images will be retained indefinitely on Digitarium's Metacat service and eventually at the long-term archival service of the National Digital Library (KDK PAS).

III. THE ICT SYSTEM AND WORKFLOW

The above steps are being supported by an ICT system that implements the workflow. Its main components are described below.

A. *XML and Darwin Core*

All original data is being stored in XML documents. They contain only terms from the Darwin Core and Dublin Core (http://dublincore.org/documents/dces/) standards. A detailed guide for their application at Digitarium has been written [4]. An example is given in Fig. 3. Typically a new version of the XML document is generated at each step of the process. For the time being, the all the documents and images are managed just on the file system. Metadata describing datasets (i.e. groups of Darwin Core XML documents, and orders by customers) will also be stored in XML files using the EML (Ecological Metadata Language) standard.

B. *Digitisation workbench*

A dedicated tool has been written for digitisation and automation of workflow. This tool, called JJC, has been written in Java at Digitarium by the first author. The tool runs in Windows and provides for data entry into the XML documents, and driving of Nikon cameras for imaging. It can retrieve and write the XML documents pertinent to each step in the workflow. See Fig. 4.

C. *Morphbank*

This service is available at http://morphbank.digitarium.fi/ and it is part of the global and Nordic collaboration. Morphbank is a database tool designed in particular for natural history specimens, the locations where they have been collected, images of them and their parts, image views, taxonomy, and annotations [5]. Morphbank is a publishing platform in the sense that after publishing date, the objects in principle cannot be removed from the service anymore and all objects have stable short URIs that can be reused elsewhere.

D. *XML database*

In order to keep track of all the XML documents and their versions, a document repository needs to be used. Ideally, it will support search within the XML documents. Metacat, the XML database tool from the LTER network is has been used elsewhere for long term archival and search of material and

related metadata [6]. Metacat is not yet in production at Digitarium, but is being tested. The National Digital Library of Finland (KDK) is building a long term archival systemn (PAS), which also will be used when it becomes operational in 2016.

IV. Conclusion

The process and tools described above have been designed in 2010-2011. They are well known and have been described earlier by GBIF [7] and others. However, there are many unique features in the Digitarum process. First, digitisation and collection management have been separated. This should simplify the production, and has led to the choice of using just XML-documents for data management, as the solution and process are independent of the collection management software used by customers. Indeed, a relational database might not be an ideal solution at all for natural history collections, as there are few transactions. On the other hand, there is a need to keep track of the history of specimen curation, identifications, georeferencing, publishing events, which can be easily handled with a version control system such as SVN (Apache Subversion; http://subversion.apache.org/), but is not a typical feature of a relational database. XML-based document management makes it possible to easily go back to original material and retain all older versions, if needed.

Second, all material is imaged. With the low cost of storage this is becoming increasingly popular also elsewhere. This reduces the need for handling the specimens.

Thirdly, and more importantly, comprehensive imaging makes it possible to distribute data entry and subsequent steps of the process to distance workers. This way costs can be reduced and access to remote experts gained for purposes such as handwriting recognition, languages, and species identification.

Fourth, comprehensive digitisation may reduce the need to access the specimens physically. This is still somewhat controversial, but many studies can be carried out just using the digital copy of the specimen. When digitised, the specimens and entire collections can be stored in a remote repository in a less expensive town and building than the big museums in city centres typically can provide. Digitarium offers this repository service for the material it digitises.

Implementation of the process described here is by no means completed. The process is being tested and refined, but is not yet very effective for large scale production. Scaling up of capacity will happen gradually, but it is too early to estimate what level of efficiency will be achieved.

Acknowledgments

We thank Mikko Heikkinen for support for the URI tagging and Tapani Lahti for the ideas concerning XML-based data management. We are grateful of the help and cooperation of Greg Riccardi, Deb Paul, and others of the Morphbank team at Florida State University. We also thank an unknown referee for improvements of the text. This work has been financed by the European Social Fund and European Regional Development Fund.

References

[1] Berendsohn, W.G., Chavan, V. & Macklin, J. 2010. Summary of Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Digitisation of Natural History Collections. Biodiversity Informatics, Vol 7, No. 2.

[2] Pelkonen, V-P., Saarenmaa, H. & Laurenne, N. (editors) 2009. Luonnontieteellisten museokokoelmien digitointi. Strategia ja toimintasuunnitelma 2010-2015. Helsingin yliopisto, Luonnontieteellinen keskusmuseo 31.12.2009.

[3] Saarenmaa, H. 2009. Luonnontieteellisen keskusmuseon tietoaineistojen avoimuuspolitiikan toteuttaminen Hatikassa ja muissa aineistoissa. Luonnos, versio 4.3 http://gbif.fi/fi/node/32

[4] Haapala, J. & Lehtonen, J. Darwin Core 2009-09-23: Käyttöohje biologisten aineistojen tallentamiseen. Manuscript 2011-05-24, 26 pages.

[5] Morphbank :: Biological Imaging (http://www.morphbank.net/, 31 May 2011). Florida State University, Department of Scientific Computing, Tallahassee, FL 32306-4026 USA.

[6] Berkley, C.; Jones, M.; Bojilova, J.; Higgins, D.; 2001. Metacat: a schema-independent XML database system. Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on , vol., no., pp.171-179, 2001.

[7] Global Biodiversity Information Facility. 2008. GBIF Training Manual 1: Digitisation of History Collections Data, version 1.0. Copenhagen: Global Biodiversity Information Facility.

This paper can be cited as follows:
Lehtonen, J., Heiska, S., Pajari, M., Tegelberg, R. & Saarenmaa, H. 2011. The process of digitising natural history collection specimens at Digitarium. Pages 87-91. In: Jones, M.B. & Gries, C. (editors) Proceedings of the Environmental Information Management Conference 2011 (EIM 2011). September 28-29, 2011. Santa Barbara, CA. University of California. doi:10.5060/D2NC5Z4X
https://eim.ecoinformatics.org/eim2011/eim-proceedings-2011