**Transcription Workflow v1**

|  |  |
| --- | --- |
| **Date** | 23rd September 2024 |
| **Author(s)** | Pip Brewer |
| **Purpose of document** | Outline minimum requirements for the v1 transcription workflow for Tranche 1 of DaSSCo to be used for NHMD pinned insects |
| **Status** | Draft; for discussion |
| **Audience** | Developer, DaSSCo team and Steering Group, collection managers and curators |

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision Date** | **Summary of Changes** | **Author(s)** | **Version** |
| 2024-06-12 | Document creation | PB | Draft |
| 2024-17-09 | Edits | JS | Draft |
| 2024-18-23 | Edits | PB | Draft |
|  |  |  |  |

[**Overview**](#_Toc169081562)

[**1. User stories**](#_Toc169081563)

[**2. Functional requirements**](#_Toc169081564)

[**3. Non-functional requirements**](#_Toc169081565)

[**4. Workflow components**](#_Toc169081566)

[**5. Out of scope componentsLimitations**](#_Toc169081569)

[**6. Next steps**](#_Toc169081570)

[**7. Glossary**](#_Toc169081571)

[**8. References**](#_Toc169081572)

# **Overview**

The premise of this document is to outline the requirements for a short-term tool and integrated workflow to test and gain experience with integrating transcription and automated text recognition and parsing, into the DaSSCo pipelines. This experience will be used to inform our overall approach and aid in decision-making as to the next steps in developing our transcription strategy. It is not intended to cover all transcription needs in the first instance.

The workflow should integrate a tool that is easy to use and facilitates the capturing of data from images of specimen labels. Different means of data capture should be possible (at least eventually) from automated text recognition to manual transcription. Eventual automation of as many of the processes should be the goal, although more manual methodologies are expected to be common in the first few iterations/versions. Ultimately, we want to be able to capture high-quality, validated data that can be back-populated into existing Specify records (which may already have some data in them).

Version 1 will concentrate on being able to facilitate data capture for the Dung Beetles Research Project led by Alexey Solodovnikov and Aslak Kappel Hansen at the Natural History Museum Denmark (NHMD). For v1, we will concentrate on pinned insect specimen type at NHMD. The workflow and transcription platform will then be evaluated, prior to a version 2, where its remit will be expanded.

# **1. User stories for workflow, Specify and transcription platform**

Users:

* Digitiser
* Collections manager/curator
* Researcher
* Data manager / senior digitiser (data)
* IT Lead
* Volunteer
* Project lead

MSCW:

* M = Must have
* S = Should have
* C = Could have
* W = Won’t for now

Version:

* 1 = Using core data for NHMD Dung Beetle Project digitised using DaSSCo infrastructure
* 2 = Using core data for NHMD Entomology, NHMA Entomology, NHMD Herbarium Sheets, AU Herbarium Sheets digitised using DaSSCo infrastructure
* N/A = Will be considered for a later version

Development part:

* P = Platform
* S = Specify
* W = Workflow

The following user stories were captured during a workshop on 07/11/2023, feedback from an email distributed on 16/01/2024, an online pinned insects meeting on 22/02/2024 and an online herbarium meeting on 29/02/2024. These discussions, captured in the following user stories relate to transcription generally, and not just to v1 (the subject of the rest of this document).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **User stories** | **MSCW** | **Version** | **Develop-ment Part** |
|  | As a collection manager, I want to focus on capturing data from specific fields and making that data available as quickly as possible (for example to compare collector name and numbers with other databases)  As a researcher, some data is more important than others (e.g., locality data) for specific research projects (and in general) and this data should be prioritised in any transcription projects  As a project manager, I want to be able to prioritise the capturing of some data over others (e.g., prioritising collector name) in order to design specific transcription projects around the results (e.g., projects transcribing data from specific collectors so transcribers get to know the collectors handwriting which should improve the overall speed and quality of the results)  As an IT Lead, I would want the platform designed to be able to accommodate (choose between different data fields)different work flows. | M | 1 | P |
|  | As a project manager, I want to automate as much as possible so that we avoid spending a lot of time and resources on manual processes (people are expensive). This will make the transcription process sustainable in the long-term.  As an IT Lead, I would want the automation to be iterative based on the complexity of automation and require the iterative process to lay down the ground work for automation where ever possible, for example, during manual processing when the experts are validating data, it should be categorised and annotated, so as to help prepare the data for future automation and so on. | M | 1-2 (in part) | P, W |
|  | As a collection manager, it must be possible for people to do the transcriptions, not a machine, so that the risk of incorrect annotations is limited (e.g., for handwritten labels)  As IT Lead, I would like to set up quantifiable process tagging and data tagging to be able to automate and optimize the transcription process | M | 1 | P |
|  | As a researcher, I want to always keep verbatim data in a designated field so that it is searchable and available for specific research purposes  As a curator and researcher, I want the data to be transcribed as "verbatim text" along with the interpretation, so that both are available in case the interpretation is incorrect  As a curator and researcher, I want transcribers to do verbatim transcription instead of interpretation so that we limit the risk of incorrect interpretations as much as possible  As IT Lead, I would like the data to categorised as verbatim and interpreted and have history (versioning) of ammendments of them that are easily viewable in the UI | M | 1 | P, S |
|  | As a data manager, I want a lot of standardisation and a limitation of what you can enter so that the cleaning and validation of data require as little time and resources as possible  As IT Lead, I would want the data standardisation to be designed and developed in a way that, incase of multiple standards for same data exist, the view can facilitate to view the data in the chosen standard. | S | 1 | P |
|  | As a collection manager, I want to capture the information on who prepared the specimen so that it is available in case it is different from the collector as slides, for example, are stored under collector name  As IT Lead, I would want the platform to be able to capture the data about the transcription process like, timestamps of validations, Booleans to keep track of process and ready to publish data, and other parts of the work flow. Specific of these can be defined once the work flow and the design has been finalised. | S TBC | 2 TBC | P |
|  | As a data manager, I want to have tools, tips and/or explanations for every field so that it is clear to transcribers which type of data belongs to which fields. This should be part of the UI (e.g., a question mark near the field containing the information).  As IT Lead, I would recommend to have short vidoes and similar field value provided as drop down suggestions to help transcribe data that already exist in the platform/database and avoid errors like typos. This can also be gamified. | S | 1, 2 | P |
|  | As a collection manager, I want a way to indicate which fields have not yet been populated so that it is clear if the data is missing or if it has just not been captured yet  As a project manager, it is essential to know which fields have no data in them because it has not yet been captured versus those for which the data is not available, otherwise digitisers will spend time perpetually trying to capture unavailable data. I also need it to estimate record completeness (e.g., MIDS levels), design new projects for transcription and to report on progress.  As IT Lead, I would like a way to show if the data has been validated or confidence level of the data in the fields. I would also like to be able to collaborate with users(expert users) to increase the confidence leve of the data. | S | 1 | S |
|  | As a project manager, I want to be able to update existing records in Specify so that we can split up the mass digitisation process into smaller steps (incrementally adding more data to records)  As IT Lead, I would like the platform to conform to a certain format, that might help in the synchronisation between platform and specify. | M | 1 | P, W |
|  | As a project manager, I want to test automatic validation, e.g. by point systems, for simple scenarios, in order to free up resources (manual validation is time consuming and expensive)  As IT Lead, I would like the automatic validation to be generated and prompted as potential option of data fields along with the image to help transcribe the data easier and faster. This can also be gamified. | M | 1 | P |
|  | As a collection manager, I want to keep the transcription process as simple as possible so that the data entry is very easy for transcribers and they make as few mistakes as possible.  As IT Lead, I would want to conduct several UI tests to capture the manual process of transcription to help facilitate this requirement. | M | 1 | P, W |
|  | As a researcher, I want exact coordinates or a description of the locality for geolocation so that the exact locality is available for my research.  As IT Lead, I would want to standardise the capturing of geolocation data. It can be categorised into exact, coordination system, name of location at time of specimen capturing, name of location at time of transcribing etx and these fields should become available/viewable based on relevance and reference. | M | 1 | W |
|  | As someone who works with data validation, I want to authorise experienced transcribers/volunteers to validate the work of other transcribers so that I will have to spend less time and resources on data validation.  AS IT Lead, I would like to tag the data validated by experts and have them as suggestions with higher confidence for trancribers with auto prompts and drop down lists. | W\* | N/A | N/A |
|  | As a collection manager, I want a standard for the order in which locality data is captured (biggest to smallest or vice versa) so that the locality data becomes standardized across collections.  As IT Lead, I would like this to be customisable so that, it can be chosen to be displayed in ascending or descending order. | M | 1 | W, P |
|  | As a collection manager, I want to capture the history of the specimen so that I can contact the institution/collection it originated from and possibly get already digitised data on the specimen.  As IT Lead, I would want the capability in the platform to invite users to view specific records, and had data to the records via the platform. | M | 2 TBC | P |
|  | As a project manager, I want a solution for when data is unreadable so that transcribers and data managers will know exactly how to handle it when they come across an unreadable label  As IT Lead, I would like the unreadable data tagged and be assignable to experts. So that it s easy for experts to log in to the system and prioritise their work accordingly. | M | 1 | P, W |
|  | As a data manager, I want to keep a log of every submission from the transcription platform so that it is possible to keep track of any changes that volunteers make to their submissions. | W | N/A | N/A |
|  | As a user of digitised data, I want it to be indicated on records if data has been derived from citizen science so that it is clear what method was used to capture the data  As a collection manager, I want to save the name of the validator so that it is clear that the data has been validated by a person and not a machine  As a person who works with collection data, I want to register and save the name of the validator so that it is always possible to go back and see who validated the data in case you have questions or comments | W | N/A | N/A |
|  | As a user of the digitised data, it is essential that data accepted into Specify following transcription is of as high quality as possible | M | 1 | P, S, W |
|  | As a data manager, I want to split up the transcription and focus on fewer fields/smaller part of the label at a time so that it is very clear what data from the label belongs in each data field, and so that transcribers are less likely to make mistakes | M | 1 | P |
|  | As a project manager, I want to have a platform for volunteers to help each other so that the members of my team have to spend less time on support of volunteers | W | N/A | N/A |
|  | As a project manager, I want to find a way to implement gamification in the transcription process so that the process is more interesting to transcribers and they stay motivated and engaged | S | 1 | P |
|  | As a project leader/digitisation co-ordinator, I would like to be able to easily compare how different users perform over time and on different collections. | M | 1, 2 | P |
|  | It should be necessary to log into the UI and the projects are restricted to specific authorised users | M | 1 | P |
|  | As a project manager, I would like software that can be used from any computer with internet access, so we can expand and test different types of users based in a variety of locations.  As IT Lead, I would like the platform to be mobile friendly or have an corresponding app for easier access. | S | 2 | P |
|  | As a project manager, I would like digitisers to be able to continue to transcribe even when the internet is not accessible so there are no bottlenecks due to periods of downtime | S | 1 TBC | P |
|  | As IT Lead, I would want the data ingestion and data extraction from the platform to be flexible and in various formats to facilitate easy integration with other systems. |  |  |  |
|  | As IT Lead, I would like the copyright, Data license mentioned in the platformed and policies established to adhere to them. For example, Terms and condition, setting page for users to give permissions etc. |  |  |  |
|  | As IT Lead, I would want the data recording to be done in consulatation with and AI/ML expert to help us capture the data in a way that can be later fed to a AI model. |  |  |  |
|  | As IT Lead, I would want to have the possibility to add more fields and give them labels, split fields depending on collections and provide a way to club fields in case that is required to send the data to Specify. |  |  |  |

\* Initially we will look at consensus/automatic validation and evaluate that via wider discussions. We will make further decisions based on that evaluation.

# **2. Business requirements**

There should be no restrictions or conflicts as a result of DaSSCo using, adapting, distributing and promoting the software and it should be released under an open licence.

V1 should be operational by November 2024 at the latest in order to meet the needs of the Dung Beetle Project.

# **3. Functional requirements for v1**

|  |  |
| --- | --- |
| **#** | **Requirement** |
|  | The transcription platform can be used by multiple institutions and collections (and collection types).  That means being able to conform their data to platform to be able to input/output? the data into the platform. |
|  | Object (specimen) records for transcribing can be grouped into projects. Based on ?grouped hin a heriarchy ? |
|  | Projects can be different sizes (made up of different numbers of object records for transcribing), should it be possible to add/remove more object records to these groups, should this be userbased? |
|  | Multiple projects can be available at any time on the transcription platform per user? |
|  | It is possible to specify which fields appear in the UI for individual projects by an administrator or project owner |
|  | There should be a one to many (and possible to have a many to one, but less common) relationship between non-object records (e.g., localities) and object records (should not be one to one) |
|  | It is possible for administrators to easily add projects to the platform and specify what fields to show |
|  | Data from Specify object records and associated data for loading into the platform can be imported into the transcription platform via csv or tsvs??? The data must include associated image URLs |
|  | For non-object records imported to transcription platform, only those Specify records tagged as “Validated” will be used |
|  | The landing page for the transcription platform is a log in (and registration, if not already an authorized user), and only users have access to data upon login. There should be a terms and condition for data that user should accept. |
|  | The log on/ landing page for the transcription platform shows DaSSCo logo and name of transcription platform and a short description of platform |
|  | There are contact details on the landing page in case a user has difficulty logging in |
|  | A project owner or administrator needs to assign users to projects, otherwise they will not be visible |
|  | In the project page in the UI, users who have access to more than one project can chose the project they wish to work on and can see a basic description and some statistics on the projects |
|  | There is a point score system for different users which can be improved with successful acceptance of their transcriptions. The categories are as follows:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Role** | **Category** | **Description** | **Accepted submissions required** | **Points attributed to each submission** | | CONTRIBUTOR | Basic | First time user. Fields displayed are restricted. | 0 | 10 | | Beginner | More fields are displayed, but some are restricted. | 10 | 20 | | Competent | More fields are displayed, but some are restricted. | 50 | 30 | | Advanced | More fields are displayed, but some are restricted. | 100 | 40 | | Expert | Can submit all fields. | 500 | 50 | | ADMINISTRATOR | Administrator | Can perform all tasks above, data management (submission approval) | - | 60 | | ROOT | Root | Can perform all tasks above, administrator management, specimen management. | - | 60 | |
|  | It is possible for users to see their progress and improvement in the platform for the user, admin, others?? |
|  | It is possible for some users (e.g., administrators) to be able to see how individuals are performing across different projects and to view their progress |
|  | The platform is available in English and in Danish |
|  | The total number of records left to transcribe in a project is visible in the UI during transcription |
|  | It is possible to easily navigate through records for transcribing using forward and back arrows |
|  | Data that already exists in Specify in the fields to be transcribed as part of the project, can be viewed in the transcription platform for each object (specimen) record as editable data or vieable data? |
|  | It is possible to see values for a field that others have submitted as part of the transcription process and have the option to change your own, in light of them |
|  | It is possible to skip records or part of records with ease |
|  | Records that have been full transcribed and validated can no longer be seen in the transcription UI |
|  | There should be field-level help in the form of question mark icons by fields to help users, it should also contain examples |
|  | If 2 expert users are unable to agree on the contents of a field it is automatically flagged as being undecipherable |
|  | The value “unknown:missing” should be available in transcription UI when data is relevant data is not shown on image of label (e.g., the locality is not known) |
|  | Multiple images can be accessed for any one specimen record (not necessarily in one view) and it is possible to move between them |
|  | It is possible to zoom in and out on images in the transcription UI |
|  | Images should take up the left hand half of the transcription UI, with the right hand half displaying the fields for filling in |
|  | Fields for inputting data in the transcription UI can be of a variety of types including text strings, dropdown boxes, query-combo boxes, dates, numeric, check boxes, radio buttons and can be specified for each field. |
|  | Start collection date and end collection date as DD/MM/YYYY are included in the transcription UI. It is possible to enter the full dates, or partial dates (e.g., MM/YYYY or YYYY). |
|  | Locality is split into multiple fields (TBD/TBA: Country, Region, Kommune, specific locality written smallest to largest components and Broad Geographic Region) in the transcription UI. |
|  | For existing localities in the transcription UI – the components making up a single locality (divided up into fields in the UI) are linked, so that if you search for the locality in the top field and select one, the other fields are automatically filled in |
|  | It is not possible to update an existing locality record in the transcription UI, but it is possible to create a new one in the UI. |
|  | New locality records created in the transcription UI are available for subsequent selection and associating with other object records as soon as they are created in the transcription platform |
|  | There is a remarks field with a related remarks source field, included in the transcription UI for each set of fields (e.g., localilty and date). This enables digitisers to add verbatim text. These go into the Remarks table in Specify and the source will say “Verbatim locality from specimen label” and “Verbatim date from specimen label” for v1. |
|  | Initially, the data in Specify will only be updated once all records are transcribed and accepted by the transcription platform |
|  | The transcription platform uses the Specify Operations API to update object records following transcription |
|  | It is possible to add to and replace existing data in Specify in bulk as part of the data ingest from the transcription platform |
|  | When updating Specify via the API from the transcription platform, the update should be recorded in the Specify audit log in a meaningful way so it is obvious that the update was due to data from the transcription platform, the transcription platform should also have a way to record was replaced and which was appended. IT might also be a good idea to have the replaced text saved for some duration |
|  | It should not be possible to add new object records to Specify during the ingest from the transcription platform |
|  | Object records should be capable of being updated in Specify during ingest from the transcription platform |
|  | It should be possible to add non-object records (e.g., localities) to Specify during the ingest from the transcription platform |
|  | New non-object records (e.g., localities) added to Specify during the ingest from the transcription platform should be flagged as “Requires validation” in Specify. |
|  | It should not be possible to update existing non-object records (e.g., localities) in Specify as a result of data ingest from the transcription platform. |
|  | Values in object records with “unknown:missing” or “unknown:indecipherable” should link to Specify records in the relevant tables through the Specify Operations API and not directly using Specify Database API (e.g., in v1, these are existing records in the locality table in Specify) |

# **4. Non-functional requirements for v1**

|  |  |
| --- | --- |
| **#** | **Requirement** |
|  | The UI should be simple and easy to use including for non-technical people |
|  | Images in the UI should be high enough quality that it is possible to clearly see any handwriting |
|  | Images should load quickly in the transcription UI when navigating between records |

# **5.** **Workflow components**

In v1, data will be extracted manually from Specify and loaded into the transcription platform as a project. The set up of individual projects can be specified in a user interface. Attachments to the specimen records in Specify that are exported will need to be viewed in a viewer in the transcription platform. Validation of the transcriptions will be via an automated points system. The Specify API will be used to merge the changes with Specify. V1 will deal solely with NHMD pinned insects and capture only location as a query-combo box and as a verbatim text string, as well as collection date (to and from).

Preparation stage

A diagram of a flowchart

Description automatically generated

Transcription

A diagram of a diagram

Description automatically generated

Validation

[being decided: <https://github.com/NHMDenmark/DaSSCo-Transcription/issues/18>]

# **6. Out of scope components**

The user stories captured in part 1 of this document and not included in v1 will be considered for later versions, pending evaluations and discussions of earlier versions and features.

The following are out of scope of v1:

* Changing the platform to Python
* Citizen science activities are not considered here. The expectation is that early versions will be tailored for and used by DaSSCo digitisers and collection managers.
* Collections and institutions other than NHMD pinned insects and fields/data/records other than those required for an initial test of the Dung Beetle Project are out of scope of v1. Version 2 is expected to expand the remit of the platform and workflow to cover additional institutions (Natural History Museum Arhus and Arhus University Herbarium) and additional specimen types (herbarium sheets).
* Synchronisation with Specify to constantly update the transcription platform so it reflects the latest changes in Specify at all times
* Developing an automated georeferencing pipeline (it will be primarily a manual process in v1)
* Integrating automated text recognition of the label text

# **7. Limitations**

Version 1 will be limited to developing basic features to facilitate the Dung Beetle Project. Future versions will expand its remit. It is expected that there will be increasing automation of processes due to each development cycle.

# **8. Next steps**

Following finalisation of this document, requirements will be tracked via a spreadsheet. Testing and refinement of the transcription platform and preparation of the data and project (see preparation in part 5 of this document) is expected to happen in October 2024. V1 should be ready and in use in November 2024. A workshop to evaluate v1 is scheduled for February 2025. Finalisation of requirements for v2 will follow on from this. Version 2 will expand the remit of the platform and workflow to cover additional institutions (Natural History Museum Aarhus and Aarhus University Herbarium) and additional specimen types (herbarium sheets).

# **9. Evaluation**

During testing, the core functionalities and requirements will be tested.

Following release of v1, there should be a period of evaluation by stakeholders, IT team and users culminating in a workshop. This testing and feedback will look at the primary purpose of the tool, the features and functionality, but also the sustainability of this piece of software and whether it is likely to meet future needs of the programme (and how much investment will be required).

# **10. Glossary**

# **11. References**