**Transcription software v1**

Minimum viable product for the transcription layer v1 of the data pipelines for DaSSCo

|  |  |
| --- | --- |
| **Date** | 23rd July 2024 |
| **Author(s)** | Pip Brewer |
| **Purpose of document** | Outline minimal requirements for the v1-2 transcription layer for Tranche 1 of DaSSCo to be used for pinned insects and herbarium pipelines. |
| **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 | 0.1 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

[**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 product is to provide a short-term tool 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 purpose of this product is to provide an easy to use tool that can be used with little training and that captures specimen data from images of labels and delivers high quality validated data to Specify quickly and effectively. These are the criteria from which it will be evaluated.

This document covers two versions?? Check with Joaquim and Bhupjit.

# **1. User stories**

Users:

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

|  |  |  |
| --- | --- | --- |
| **#** | **User stories** | **MoSCoW** |
|  | 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) | M |
|  | 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 | M |
|  | 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) | M |
|  | 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 process sustainable in the long-term. | M |
|  | 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) | M |
|  | 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 [Christian]  As a curator and researcher, I want the data to be transcribed as "verbatim text" [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 to that we limit the risk of incorrect interpretations as much as possible [Alexey] Specify implication |  |
|  | 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 Need to resolve verbatim issue above at same time to prioritise this |  |
|  | 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 [Anders] |  |
|  | 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). | S (v2) |
|  | 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 Specify implication | M |
|  | 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. Specify implication | M |
|  | 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) | M |
|  | As a project manager, I want to test automatic validation, e.g. by point systems, for simple scenarios in order to free up resources | M |
|  | 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 | M |
|  | 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 | M\* |
|  | 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 | S? v2? |
|  | As a collection manager, I want a standard for the order in which locality data is written (biggest to smallest or vice versa) so that the locality data becomes standardized across collections [Is this a Specify issue? Need more details on this] |  |
|  | 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 [this was Birgitte and Jens. Need more info here. Is this just collector name and number? Is the requirement to actually compare data and update our records with external information?] |  |
|  | 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 [discuss with Joaquim. Is this something software-related or just training/protocol development?] |  |
|  | As a data manager, I want to keep a log of every submission from crowdsourcing so that it is possible to keep track of any changes that volunteers make to their submissions [I think it was Joaquim who mentioned this] |  |
|  | As a user of digitised data, I want it to be indicated on records if data have 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 Specify implication |  |
|  | 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 |
|  | 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 [ Not 100% sure what this means. I think this was Joaquim] |  |
|  | 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 |
|  | 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 | C |
|  | 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\*\* |
|  | Logging in/security |  |
|  | 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 | M |
|  | 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/C |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

For discussion:

\* Geolocations may not necessarily need to be part of the initial transcription software at least as part of v1. This may need to be part of a second process or bulk data enhancement.

\*\* This might be easiest to do (at least in v1) in Specify (but depends on what data is accessible from Specify from this process).

# **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.

# **3. Functional requirements**

Add v1 versus v2?

|  |  |
| --- | --- |
| **#** | **Requirement** |
|  | The software is developed for use by multiple institutions and collections (and collections types) |
|  | Records for transcribing can be grouped into projects |
|  | It is possible to specify which fields appear in the UI for individual projects |
|  | Projects can be different sizes |
|  | Multiple projects can be available at any time |
|  | What is the requirement for manually adding data to the UI? Verbatim typed text, drop down lists, look up lists, a mix? Presumably start typing |
|  | It is possible to update, append or replace existing data in Specify in bulk as part of the data ingest from transcription |
|  | V1 should capture locality data for NHMD pinned insects (verbatim and controlled list) |
|  | New localities can be added to the controlled list accessible in the UI, which are later validated and the look-up lists can be regularly updated to reflect the new validated entries |
|  | There is a point score system for different users which can be improved with successful acceptance of their transcriptions |
|  | It is possible for users to easily see their progress and improvement |
|  | It is possible to have an automated text recognition “user” to test |
|  | Can see values for a field that others have submitted and have the option to change your own, in light of them |
|  | The XXX uses the Specify web API to XXXX |
|  | Multiple images can be accessed for any one specimen record (not necessarily in one view), and it is possible to zoom in. |
|  | Images in the UI should be high enough quality that it is possible to clearly see any handwriting |
|  | The name of the transcriber and/or validator of a field are recorded in Specify and can be audited |
|  |  |

For discussion:

# **4. Non-functional requirements**

|  |  |
| --- | --- |
| **#** | **Requirement** |
|  | Automate as much as possible |
|  | Minimise time between extraction of records from Specify and updating them with results of transcription |
|  | The UI should be simple and easy to use including for non-technical people |
|  |  |

For discussion: No open discussion items.

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

Input and minimum outputs for each component in the MVP.

# **6. Out of scope components**

Citizen science activities, training and workflow development that does not include the software are not considered here.

The following tasks are not included in the MVP, but should be feasible in the architecture and can be prioritised for future inclusion at a later date.

Collections and institutions other than NHMD pinned insects, NHMA pinned insects, NHMD vascular plants, AU vascular plants are out of scope of the MVP

Customisation of which projects are visible to certain individuals.

Change to Python

# **7. Limitations**

# **8. Next steps**

Working version for pinned insects NHMD (priority), then NHMA pinned insects, then NHMD herbarium, then AU herbarium

Finish off finalising the requirements (working with Pip)

Start identifying the relevant API endpoints, just to do some preliminary tests on how your software will be integrated, and to see what is possible. The new Specify 7 API is quite nicely documented but is huge.

# **9. Evaluation**

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

Following release of v2, there should be a period of evaluation by stakeholders, IT team and potential users. 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**