

## Development of the data aggregator for natural history collections DAGI-V1

Pia Stieger, Sofia Wyler, Lukas Vanazzi



## 1. Scope

The major goal of this project was the enhancement of an initial version of the Data Aggregator (DAGI-V0), which was developed between May 2023 and October 2024 (for details see: Development of the data aggregator for natural history collections DAGI-V0). Building on the feedback and modifications collected during the previous project, this project aimed to bring DAGI-V0 to the next level by expanding its capabilities and optimizing its performance by enhancing the system's overall functionality and stability. Features such as optimized image upload and management, the application of publication rules aligned with the InfoSpecies-GBIF.ch deontology and an extended metadata publication were in focus for the development of DAGI-V1. The extension of the data model to deal with additional information, the improvement of the encoding process and its use during publication with a more customizable approach (accept a certain modularity instead of all or nothing) remained on a wish list, do to restricted financial resources.

In the first contract phase (May 2023–October 2024), the following specific project objectives were defined:

- The developed solution aggregates digitally available data on collection objects and collections from public and non-profit Swiss organisations, as well as data on collection objects collected in Switzerland and curated abroad.
- Minimum goal: consolidation of all data produced within the framework of the SwissCollNet initiative.
- The developed solution applies the FAIR principles to data structuring and uses international data models to ensure interoperability with national and international data infrastructures.
- The developed solution enables the publication (digital accessibility) of data in national and international data infrastructures and on the SwissNatColl portal.
- The developed solution is documented as an open source project and made selectively or generally available at the request of SCNAT or info fauna.
- The developed solution can be adapted to future requirements in terms of data structure and data content and ensures the continuous expansion of the data collection via the data aggregator as well as the updating of existing data (future-proofing of the data collection).
- The developed solution is based on a modern infrastructure that enables efficient maintenance, updating and further development (future-proof technical database).

These project objectives have also been applied as specific requirements in the expansion phase of DAGI, unless they were adapted or expanded by the client within the scope of the project. The results achieved were compared with these objectives.

## 2. Organisation and documentation

Based on the experiences gained from the project, an iterative approach was followed, ensuring close and collaborative partnerships with the teams from SCNAT, GBIF.ch, and info fauna. To ensure thoughtful incorporation and alignment with the project goals of new functionalities, flexibility and refinement of each iteration were foreseen. This collaboration was key to aligning the development efforts with the needs of all stakeholders, fostering a productive and synergistic working environment.

The agile approach was composed of iterations and sprints. The following setting was applied:

- After testing DAGI-V0, the parties agreed on defined project goals as well as the time frame and budget required to implement these project goals.
- The tasks necessary to implement the project goals were structured according to subject areas (epics). Both epics and tasks could be adapted to requirements during the project.

- The epics were processed in 6-7-week iterations, each with 2-week sprints.
- Each iteration started and ended with a coordination meeting. At this meeting:
  - a) The results of the iteration were presented and approved by the client.
  - b) The project progress was reported and any outstanding issues were recorded.
  - c) The project goals were evaluated and supplemented or adjusted, if necessary.
  - d) The tasks were planned for the next iteration and assigned to the responsible parties.

This approach has enabled efficient and planned project progress, ensuring that the developed solution meets the specific needs of the customer.

The iterative development process was divided into 5 cycles (iterations). Each iteration has involved the following workflow and responsible persons (Figure 1):

**Planning:** Prior to the start of an iteration cycle, the steering board composed of representatives from SCNAT, GBIF.ch and info fauna has identified priorities for each iteration cycle in collaboration with stakeholders and decisions taken by the steering group were documented. Thereafter, development goals were defined for the upcoming iteration cycle in a meeting between representatives of zebbra and the steering group (iteration meeting). Minutes of the meeting were taken and approved by the persons present in the meeting.

**Development:** Features defined during the planning stage were implemented by the engineers of the IT-company zebbra in collaboration with the team of GBIF.ch. Implementations were planned, reviewed and refined in regular sprint meetings (coordinated and documented by the coordinators of the two organisations).

**Reporting of iteration cycle achievements:** At the end of an iteration cycle, achievements were presented by the IT-company and approved by the steering board. If needed, refinements were planned for the next iteration cycle.

**Testing and reviewing:** After every iteration cycle, the system was tested by the members of the testing group, feedback collected, evaluated and if appropriate, integrated in the next planning steps.

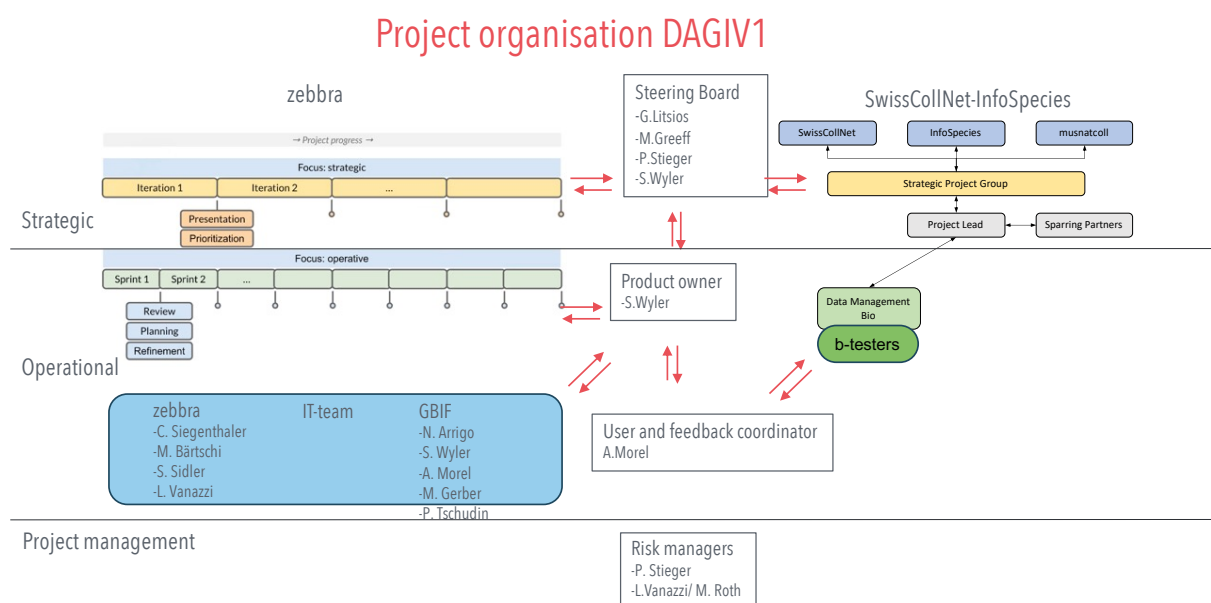


Figure 1: Project organisation for the development of DAGI-V1:

### 3. Milestones and results

Enhancements and new features were planned and divided into five iterations of 6 weeks each. Each iteration allowed for flexibility and refinement, ensuring that new functionalities were thoughtfully incorporated and aligned with project goals.

#### Iteration 10 (October 15 to December 9, 2024):

The first milestone of this iteration was the “go live” of DAGI-V0 on October 22, 2024. For this step, a product environment was set-up at info fauna, where the code was transferred to and data upload by the data providers was started. In addition, a staging environment was set-up at info fauna, where b-testers could test the newly developed features. The code was regularly updated in the product environment upon approval of new features by the steering board.

Also, the issue of image upload and management was addressed, allowing users to upload archives of images (e.g. zip-files) and map them to existing records within collections. The system also informs the user about success and failure of image upload and mapping.

Another achievement was the enhancement of information in the side panel, which contains the information for each individual specimen. Furthermore, the data model has been extended for Swiss coordinates.

The system was improved to include a switch for admins, to reduce job timeouts and eliminate ambiguous georeferencing.

Anne Morel has prepared and held tutorial sessions to coach users for data preparation and onboard them for data upload to DAGI.

This iteration also included a RETRO-FUTURE workshop to evaluate achievements and important issues to be addressed during the development of DAGI-V1.

#### RETRO-FUTURE workshop (November 20, 2024)

Representatives of the steering board, the product owners and members from zebbra (Nils, Arrigo, Sofia Wyler, Anne Morel, Michael Greeff, Pia Stieger, Lukas Vanazzi, Markus Roth) analysed the progression of the project. After a round of formulating important topics to be addressed by the participants, L. Vanazzi summarised the project steps, milestones and pitfalls as well as some statistics. So far, 164 features were implemented, 22 features still open, 37 changes made, 23 changes still open, 48 bugs fixed, 12 bugs still open. After the presentation, feedback on the project by the participants was collected to help prioritize developments for DAGI-V1 and highlight important points that would need to be addressed in the future outside this particular project.

The second part of the workshop was held together with the B-testers (Christian Püntener, Anouk Mentha, Noémie Chervet, Kamil Dobosz, Anne Morel). At first, the B-testers described their understanding of the functionalities of DAGI and missing information was added by N. Arrigo and L. Vanazzi. Thereafter, B-testers were asked to give feedback on the product and formulate wishes for additional features. To conclude, tasks to be addressed for the development of DAGI V1 were defined and attributed to project members.

#### Iteration 11 (December 10, 2024 to January 27, 2025):

In this iteration, the project turned from a norming phase into a performing phase: major conceptual mistakes were removed and the code adopted in consequence. A major issue was that while updating a dataset at GBIF.org, the entire file was overwritten and therefore all data that was not contained in the updated file, was deleted at GBIF.org.

Publication rules to ensure the protection of sensitive data were developed and implemented. The coordinates of the specimens collected in Switzerland and whose taxa are registered in the Swiss Species Registry (taxa treated by InfoSpecies data centres) are blurred (rounding WGS84 decimal coordinates by limiting the number of decimal places to 2).

Dataset combination of metadata was technically not possible to be implemented. Combination of already imported data into DAGI only made possible for persons with the administration rights. Publication of a partial dataset was implemented.

A lot of adjustments and bug fixings were also part of the iteration. The data model was extended with unique record identifiers (occurrence ID, gbif ID, catalog number).

Conceptual mistakes in the validation layer were corrected.

#### Iteration 12 (January 28 to March 10, 2025):

This iteration was especially dedicated to solve some important issues related to the use of the Swiss Species Registry by the DAGI. A lot of time and energy was dedicated to solve a conceptual problem and finally a compromise has been reached for most taxa in the Registry. A more stable and satisfactory solution will be object of further developments (funded by GBIF.ch/info fauna).

The aim to develop new filters for data filtering had to be moved to iteration 13.

#### Iteration 13 (March 11 to April 28, 2025):

Encoding of the attribute "date" was implemented, bugs for image upload fixed (refactoring of image upload was necessary), feedback adaptations such as renaming improvements were made and the appearance of the terms and conditions upon log-in of the user was implemented.

A big achievement was the GO-Live on GBIF.org, meaning that datasets in DAGI, when published are shown on the GBIF.org page and no more in a staging environment.

The system was also refactored to separate the earlier called fast track for publication of specimen data from the validation track through the InfoSpecies data centres.

Security enhancements were made by API Authentication.

#### Iteration 14 (April 29 to June 17, 2025):

This last iteration was foreseen for bug fixing, system stabilisation, documentation of the code (open-source) and to finalise the contract for maintenance in the future. Also, the terms of use and a privacy policy declaration were finalised.

A final meeting was held on June 17<sup>th</sup>, 2025, to terminate this project phase and launch the Go live of DAGI-V1. Nonetheless, a few points remained open, namely some testing and last bug fixing and the completing of making the code open source on GitHub.

#### **4. Conclusions and outlook**

The project of building a data aggregator for natural history collection data was very ambitious and demanding for several reasons. Project goals had to be defined to satisfy multiple stakeholders and be translated into technical solutions in short time and with limited human and financial resources. A high level of complexity had to be addressed, which led to a number of conceptual refinements. The involvement of many persons together with changes in the equipes added to the complexity and made alignments of project goals sometimes to be an act of balance. Furthermore, testing of the system was done on a voluntary basis by a very limited number of persons. These difficulties were overcome through an enormous teamwork, close collaboration and great motivation of all the people involved in the project. The participants all had a big sense of duty and respect, always stayed transparent and had big trust in each other, with a good spirit and strong sense for collaboration. The testers were very motivated and much engaged and have provided valuable and precious feedback to increase the performance of the system. The complex project was handled with flexibility, which has led to a very satisfactory product and a lot of positive feedback from the user community.

DAGI-V1 is now hosted and maintained by the team of GBIF.ch. Further developments for connecting DAGI-V1 with the databases of the data centres trough PICTIS will be made and are orchestrated and financed by GBIF.ch/info fauna.

The code of DAGI-V1 will be published on GitHub with an open source licence GPLv3.