

# Swiss Natural History Collection Network

## Development of the data aggregator for natural history collections DAGI-V0

Pia Stieger, Nils Arrigo, Lukas Vanazzi

## 1. Scope

The aim of the present project was to offer solutions to **collect** and **harmonise** heterogeneous, multilingual natural history collection data (NHC-data) from decentralised sources, and to **transfer** standardised information to the Global Biodiversity Information Facility (GBIF.org) and the datacenters of InfoSpecies, the Swiss Information Center for Species.

Specifically, the project aimed at developing a secured system to **aggregate, store** and **transmit** NHC-data and to act as a work gateway where data providers can **connect** to **upload, visualise** and ultimately **edit** and **update** their datasets (before transfer).

*Connect* - Data providers are able to register to the system in the work gateway through a login and to assign different rights to other collaborators by the owner of the dataset.

*Upload* - Different ways of uploading datasets had to be considered and the manual loading of data files csv be developed. Opportunity to upload complementary files, such as images and weblinks, had also to be developed.

*Aggregate* - This unit allows data to be compiled and configured according to stable controlled vocabularies and defined standard Darwin Core. The development of automated basic controls and formatting processes associated with a reporting service was essential and a rules generator module as a central component of the system, allowing to continually improve with new rules as needs arise, had to be incorporated.

*Update* - Data providers are able to make changes to the dataset they own and add newer information, overwrite or remove outdated datasets. To simplify the follow up of such changes, the system deals with different versions of a dataset and offers an efficient lasting solution for the user to manage these versions.

*Export* - Compiled raw data as well as standardized and enriched datasets are stored in different layers within the data aggregator. Data owners can export each layer of data separately.

*Visualise* - The work gateway provides data providers a graphic interface that allows the visualisation of a dataset through the different stages of treatment. Moreover, different filters allow the display of selected data.

*Store* - A centralised place where data can be managed and maintained in an organised and perennial way was a crucial requirement to be fulfilled. Moreover, as complementary files (e.g. images) must be stored as well, a solution had to be provided for their accessibility and distribution.

*Transmit* - The system had to be developed to be able to send out data to other relevant receiver systems as automatically as possible. Differentiated transmission of a dataset had to be foreseen, as Swiss data follow a stricter dataflow.

*Open source* - The software will be protected with the GPL-3.0 open source license and be free for use and modification for non-commercial purposes.

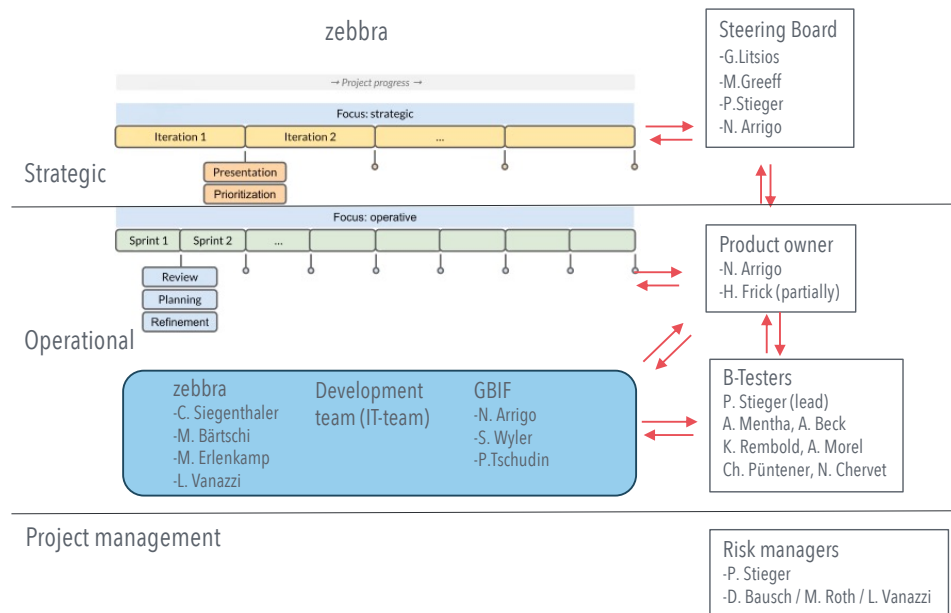
## 2. Organisation and documentation

The project was managed with an AGILE approach, based on iterative, incremental, and adaptive development cycles. A close collaboration between the IT company and the info fauna / GBIF.ch team to define and decide on the appropriate solutions to be implemented by the IT-company was established. The actual work was planned in iterations, each of six weeks or longer. Project iterations were divided into sprints of two weeks.

Iterations were used for strategic project management. At the end of each iteration, the achieved results were presented to the Steering Board of the project. In the same meeting, the scope for the next iteration was defined. The development team (IT-team) and the product owners were submitting a corresponding proposal to the Steering Board and the goals were defined together. Sprints were

used for technical synchronisation of the development team. At the end of each sprint, the development results were presented to the team and goals for the next sprint were defined. A Beta-tester group was established during the project to test the development of new features and give inputs on missing functionalities. Beta-tester meetings were held every two weeks (**Figure 1**).

Iteration meetings were documented on MIRO-board (slide presentation) and minutes were prepared after every meeting. Conceptual work for the development of the IT-architecture was documented and stored in confluence. Technical developments were collected in a ticketing system in JIRA. The code was developed in the surroundings of the IT-company and transferred to the surroundings of info fauna for use and maintenance.



**Figure 1: Project organization.** The Steering Board was composed by representatives of info fauna and SwissCollNet to take strategic decisions, provide solutions to diminish risks and monitor progress. The product owner (info fauna), in close contact with stakeholders (representatives of collection institutions, InfoSpecies, GBIF) was planning, refining and reviewing the work together with the IT-company. The project management was carried out by representatives of the IT-company and SwissCollNet.

### 3. Milestones and results

The project was started with a kick-off meeting for stakeholders, followed by an evaluation phase to define the project organisation and major milestones. The second part of the project was divided in 9 iterations covering management, feature development, testing and infrastructure set-up.

#### Kick-off meeting (May 30<sup>th</sup>, 2023):

The concept and actual state of the project for data aggregation and visualisation of biological natural history collections were presented by P. Stieger and M. Greeff to members of natural history institutions, members of data centers of InfoSpecies and members of the Steering Board of SwissCollNet. Thereafter, collaborators from zebbra presented their view on how and with whom to realise the project, addressing challenges as well as envisaged methodologies and collaborations. In a third part, actors and roles were presented to the attendees and the content of the evaluation phase was outlined. Furthermore, reactions and reflections were collected from the attendees in absence of members of zebbra.

### Evaluation phase (May 15<sup>th</sup> to July 6<sup>th</sup>, 2023):

This iteration was dedicated to formulate and conclude a contract between SCNAT and zebbra, as well as a collaboration contract between SCNAT and InfoSpecies. Furthermore, the organisation of the project and roles of persons involved were defined and the iteration phases scheduled. Also, a tech workshop was held in Neuchâtel with the team of GBIF.ch and the team of zebbra to work on legacies of already developed systems at GBIF.ch, as well as on technical landscape and IT-architecture design for the data aggregator. And a stakeholder / data provider workshop was held to identify the needs of the stakeholders.

### Tech workshop (June 14<sup>th</sup>, 2023):

The product owner presented a compilation of technical and functional requirements gathered so far in the project. The meeting aimed to initiate a knowledge transfer from users (museum actors, GBIF.ch, and InfoSpecies) to the development team, leveraging documentation and materials collected through SwissCollNet workshops, reverse-engineering of GBIF.ch systems, data standards, and over 50 stakeholder interviews. This synthesis created a working documentation base, maintained as a project Wiki on Confluence. On June 14<sup>th</sup>, zebbra was introduced to this Confluence space, establishing collaborative tools (JIRA, Confluence, MIRO) and aligning terminology among teams. This Confluence space remains active, centralizing project designs and ideas for refinement and sharing.

Stakeholder / data provider workshop (June 28<sup>th</sup>, 2023): At the workshop, stakeholders were invited to express their wishes and needs. It was emphasized that zebbra should develop a data aggregator for natural history collection data of biological and paleontological specimens and not extend the already existing IT-tools from GBIF.ch (PICTIS), but rather connect the new data aggregator to PICTIS. The "Edit" functionality was also deprioritized to allocate resources to more critical aspects of the project. This decision was based on the realization that most museums already manage daily operations with established, and mostly functional data workflows. It became evident that widespread adoption of direct data editing within DAGI would occur later in the platform's lifecycle, as more institutions integrate it into their daily activities. While smaller institutions in need of improved data management are likely to adopt DAGI sooner as their main CMS, larger institutions are less likely to transition from their long-established databases and CMS in the short to mid-term and will use DAGI as a gateway to GBIF.org, VDC and InfoSpecies in general. As a consequence, a strong backend solution for data upload and data standardisation, filters and interfaces for exports to various data platforms, as well as an automatised publication of specimen data to GBIF.org appeared to meet the main demands expressed by the stakeholders. To have one data pool and be able to show data in different windows, interoperability with other data platforms, annotation tools for non-data owners for collaboration, metadata on specimen and collection level, be able to connect molecular, morphological and ecological data and geographical distribution of the specimens, link to literature in the future, have advanced search functionalities, tools for reporting and statistics, enable data discovery and data analyses, get identification tools and tools for georeferencing, be able to up- and download data and images have been named by the stakeholders.

### Iteration 1 (July 6 to August 21, 2023):

During the first iteration, focus was put on the conceptualisation for a minimal viable product (MVP). A data model concept was nearly concluded, a process map was designed, agreements taken on the development setup and the tech stack. Furthermore, a UX concept was initiated.

### Iteration 2 (August 21 to October 2, 2023):

For operation and hosting of the infrastructure at info fauna, options and suggestions were validated, however a decision was postponed to the next iteration. A development and test environment were set up and a first minimal running application was developed with CSV file import, storage of files in S3 and persistent data input for encoding. High level flows for users were designed. The challenge of this iteration was to develop and decide on the data model, as data import was only possible to be fully implemented with a definitive data model. Also, interactions

with more potential users were needed. At least, Holger Frick could be recruited for a limited amount of time to help with data models and the description of user requirements. It was decided to quickly set-up a Beta-tester group.

At the end of this iteration, the project management within zebbra was officially handed over from Dan Bausch to Markus Roth.

### Iteration 3 (October 2 to November 13, 2023):

The data model development was terminated, a prototype for data import made, a data encoding concept designed and the user flows further specified. A MVP structure to store layered data was developed with static mapping of data to an internal data model. Also, data flows were defined resulting in a fast track to publish data on GBIF.org and an approval track with export of data to the data centers for validation and mapping to Swiss taxonomic backbone. The idea of having the possibility to review data within DAGI by external experts was dropped.

### Iteration 4 (November 13, 2023 to January 8, 2024):

Encoding services for GBIF taxonomy catalogue were implemented, data export (backend) developed, data flows within the Swiss context clarified, import of bigger datafiles made possible. A query about the number of media files had been started to understand hosting costs and be able to take a decision. The organisation of the Beta-tester group was started, but no tests have been performed in this iteration. Data flows have been conceptualised for the fast publication track to GBIF.org through PICTIS and the approval track to the data centers and back to PICTIS and the data aggregator. However, the data flow for the fast track was questioned by SwissCollNet, as it seemed more appropriate to have a direct publication from the aggregator to GBIF.org. To make sure that the concept of data flows and services were in line with the expectations of the stakeholders, a meeting was organised with the bioscience data management group.

### Meeting with bioscience data management group (December 5, 2023):

The meeting goals were to expose stakeholders of DAGI (data users, data curators and experts) to the concepts and solutions developed so far by the informatics team of GBIF.ch and the IT company zebbra and to get feedback on the coverage of the main scope and goals of the project. The concept of data import, mapping, encoding, and the different data flows were presented by Nils Arrigo and feed-back of the stakeholders was collected.

### Iteration 5 (January 8 to February 26, 2024):

Focus was on getting a commonly accepted concept of data publication and on the implementation of the User Interface concept at the frontend for the fast-track publication. Also interfaces towards PICTIS and InfoSpecies datacenters were defined, Swiss Species Catalog Encoding and Forward and Reverse Geo Encoding developed and validation via GRSciColl for collections and institutions implemented. A mid-term review was held on February 8<sup>th</sup>, in which many discussions helped to sharpen the data model and data flows (**Figure 2**).

### Mid-term review (February 8, 2024):

The achievements of the project to that day were the stable implementation of the data model, which went to many rounds of complex concepts and drafts, functional data import, adaptation of the scope during every iteration to reach a common understanding of data flows, aggregation and publication and a closer collaboration with the product owners and backlog update. Two major decisions were taken and validated: i) to directly publish from the aggregator to GBIF.org in the fast track and ii) to host media files at SWITCH and the aggregator in ORACLE. It was also discussed and decided that data providers should not create collection units within the data aggregator, but that metadata of collection units have to be announced and published in the Global Registry of Scientific Collections (GRSciColl). Only registered collection units are selectable in the data aggregator for data import. Focus points for the future development were to quickly solve bottle necks (especially for fast publication to GBIF.org and for collection registration on GRSciColl), to get principles for picture upload and publication, get fast-track rules ready (red list filtering) and



(ASH) was made. It was decided to go live at the end of the next iteration. Also, propositions were made by zebbra on what to focus for further iterations after the go live to get the DAGI-V1. 25% to 50% of the time were estimated to be necessary for bug fixing and performance improvements. Priorities were set in picture upload and handling, publication of data and metadata, georeferencing and user management. Sofia Wyler (GBIF.ch) was welcome in the function of the new product owner.

### Iteration 9 (September 2 to October 14, 2024):

To onboard users, tutorial sessions were held by Anne Morel and instructions published on the SwissNatColl-staging page. The infrastructure to run DAGI-V0 at info fauna was established, OCI and SWITCH production and staging machines up, tested and running. The backlog was cleaned up, automapping developed, import fixings and changes in history made and terms and agreement pop-ups for publication of data integrated. A first draft of Terms of Use was drafted. Picture upload was still not possible to be implemented and had to be postponed to iteration 10.

### GO LIVE (October 22, 2024):

DAGI-prod was opened on October 22<sup>nd</sup> in the info fauna environment and grantees informed for data upload.

### 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 (for results see chapter 4 Highlights and challenges). For the future of the project, the participants revealed important points to address (for results see chapter 5 Further steps).

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.

## **4. Highlights and challenges**

Highlights and challenges of the project were collected at the RETRO-FUTURE workshop and divided into four categories:

### a) What went well

- Collaboration and problem solving within the core group of the project as well as interactions with the Beta-testers, despite the complexity of the project organisation and the many stakeholders involved.
- The user-oriented approach and consistency of the project content; major functionalities could be developed in time
- Good work environment created with the different collaboration tools (MIRO, JIRA, Confluence, Jotforms)
- Very valuable returns by Beta-testers

### b) What was difficult or not so smooth:

- Lack of product owner from the collection side and slow B-tester onboarding
- Communication with so many partners and stakeholders and transmission of information to

newcomers in the project

- Alignment of requests, urgency to fulfil requests, no sudden visibility of requested changes
- Struggling of the engineers, as project was much more complex and needed much more time than what was budgeted by zebbra
- Finding enough persons with expertise to test the system
- Urgent treatment of many questions, lack of priority setting
- Misalignment of some decisions with the starting concepts and ideas

c) What was personally challenging:

- Wide audience of users, alignment between experts and users, handling of conflicts of interest, chain of command and decisions, new kind of work process for zebbra
- Drive project with very limited human resources
- Provide high quality information for development team in time
- Having three different cohorts to be integrated within the project with different agendas and priorities
- Pushing people to provide the information needed to advance
- Design the IT-architecture that respects and integrates the needs of all partners involved
- Take over of responsibilities (product owner, Beta-tester management)

d) Success stories and highlights:

- Integration of DAGI into data infrastructure of InfoSpecies and collaboration with GBIF
- Connection of DAGI to GRSciColl, the choice of the hosted portal at GBIF.org
- Optimal interactions between product owner and zebbra from the very beginning
- First publications of datasets on GBIF.org staging
- Great team spirit, arrival of new members in the project
- Positive feedbacks by the users
- Visit of zebbra in the Herbarium of the University of Bern and see the enthusiasm of collectors at work

## 5. Further steps

The project DAGI-extensions has started on October 22, 2024. It covers additional iterations to enhance functionalities and the stability of DAGI, to become DAGI-V1 (5 iterations). Three iterations are planned for optimisation of the features already conceptualized and partially implemented in DAGI-V0, one iteration for urgent new features and/or urgent optimization of already realised features and one roll out iteration for the going-live process, open source licensing and code transfer to info fauna.

In the RETRO-FUTURE workshop, the following important points to be addressed have been detected and will be prioritised by the steering group of the project and be addressed as much as possible with the remaining resources:

- The communication within the project group and towards stakeholders had to be adopted and optimised, health check of the communication tools, clear view on responsibilities to help zebbra with setting the focus on most important implementations
- The design of a success matrix would help to measure the success of the project
- Set priorities in the features to be developed for DAGI-V1 and realistic planning
- Extension of data model with GBIF extensions
- Merging information of different data layers
- Integrate one to n relationships
- Have API for data upload
- Solve problem of coordinates