DiSSCo related output

This template collects the required metadata to reference the official Deliverables and Milestones of DiSSCo-related projects. More information on the mandatory and conditionally mandatory fields can be found in the supporting document 'Metadata for DiSSCo Knowledge base' that is shared among work package leads, and in Teamwork > Files. A short explanatory text is given for all metadata fields, thus allowing easy entry of the required information. If there are any questions, please contact us at info@dissco.eu.

Title
DiSSCo Prepare Deliverable D5.5 Construction plans for the improvement of technical infrastructure in the areas of geo-collection data and taxonomic services

Author(s)
Matt Woodburn
Wouter Addink
Olaf Bánki
Tom Dijkema
Lisa French
Falko Glöckler
Josh Humphries
Sharif Islam
Sam Leeflang
Sabine von Mering
Julia Pim Reis

Identifier of the author(s)
MW: https://orcid.org/0000-0001-6496-1423
WA: https://orcid.org/0000-0002-3090-1761
OB: https://orcid.org/0000-0001-6197-9951
TD: https://orcid.org/0000-0001-9790-9277
LF: https://orcid.org/0000-0001-7279-8582
FG: https://orcid.org/0000-0002-7127-2738
JH: https://orcid.org/0000-0001-5493-9804
SI: https://orcid.org/0000-0001-8050-0299
SL: https://orcid.org/0000-0002-5669-2769
SvM: https://orcid.org/0000-0003-2982-7792
JPR: https://orcid.org/0000-0003-4383-0357

Affiliation
MW: NHM
WA: Naturalis
OB: Catalogue of Life
TD: Naturalis
LF: NHM
FG: MfN
JH: NHM
SI: Naturalis
SL: Naturalis
SvM: MfN
JPR: MfN

Contributors
Olle Hints (TalTech)
Enar Mustonen (TalTech)

Publisher
DiSSCo Prepare

Identifier of the publisher

Resource ID
https://doi.org/10.34960/dzs0-xa94

Publication year
2022

Related identifiers
https://doi.org/10.34960/xhxw-cb79
https://doi.org/10.34960/n3dk-ds60
https://doi.org/10.34960/366d-sf49
https://doi.org/10.34960/50B9-KJ05

Relation type
Deliverable

Is it the first time you submit this outcome?
Yes

Creation date
14/11/2022

Version
1.0

Citation
Woodburn M. et al. (2022): DiSSCo Prepare Deliverable D5.5 Construction plans for the improvement of technical infrastructure in the areas of geo-collection data and taxonomic services. DiSSCo Prepare. https://doi.org/10.34960/dzs0-xa94

Abstract
DiSSCo Prepare Deliverable D5.5 describes an analysis of DiSSCo requirements for interoperability with external taxonomic and geo-collections services and provides recommendations for technical and strategic approaches to building integrations during the construct phase of DiSSCo development. In earlier DiSSCo projects and work packages within the DiSSCo Prepare project, the integration of taxonomic checklists, represented by Catalogue of Life (COL) Checklist and the ChecklistBank infrastructure, into the DiSSCo architecture as a taxonomic backbone has been identified as a key requirement for users of DiSSCo services. Geo-collection data, represented in this review by GeoCASe, have also been highly underrepresented in terms of available services for data mobilisation and publication. During the task, conversations with the Earth Sciences community also identified Mindat as a key community resource for authoritative data about geological classifications and localities, and so that platform was added to the review scope of the task.

Reviews of each platform were carried out through a combination of desk-based research, document review and direct input from developers and representatives of the platforms. In parallel, user requirements were extracted from earlier DiSSCo outputs, and technical requirements from architectural design and pilot activities carried out by the DiSSCo developer team. Further detail was added to the picture by running an event-storming workshop, which engaged wider members within the community to help to identify and prioritise those events in external systems that would need to be reflected in the data within the DiSSCo architecture. From these activities, it was possible to make a comparison between the technical capabilities and resource capacity of the reviewed platforms on one hand, and DiSSCo's requirements and technical roadmap on the other.

DiSSCo's favoured technical model for sustainable interoperability with external services is based on the event publishing approach, originally proposed for integration with institutional collections management systems. An important message that emerged is that this would need investment and collaboration to achieve with GeoCASe and Catalogue of Life. Although both are enthusiastic about exploring further integration, GeoCASe is critically under-resourced, while Catalogue of Life would expect DiSSCo to partake in the joint investment of the sustainability and maintenance of its services, and key infrastructure called ChecklistBank that will be jointly governed by international biodiversity data initiatives and infrastructures. While there are more pragmatic approaches to integration that can be explored in the short term, further exploration of the longer term relationships and opportunities for external investment form a key part of this report's recommendations.

Content keywords
technical

Project reference
DiSSCo Prepare (GA-871043)

WP number
WP5

Project output
Deliverable

Deliverable/milestone number
D5.5

Dissemination level
Public

Rights

Resource type
Text

Format
PDF

Funding Programme
H2020-INFRADEV-2019-2

Contact email
M.Woodburn@nhm.ac.uk

H2020-INFRADEV-2019-2
Grant Agreement No 871043

DiSSCo Prepare WP 5 – D 5.5

Construction plans for the improvement of technical infrastructure in the areas of geo-collection data and taxonomic services

Work Package leader: Mareike Petersen (MfN)

Authors: Matt Woodburn (NHM), Wouter Addink (Naturalis), Olaf Bánki (Catalogue of Life), Tom Dijkema (Naturalis), Lisa French (NHM), Falko Glöckler (MfN), Josh Humphries (NHM), Sharif Islam (Naturalis), Sam Leeflang (Naturalis), Sabine von Mering (MfN), Julia Pim Reis (MfN)

Contributors: Olle Hints (TalTech), Enar Mustonen (TalTech)

MUSEUM FÜR NATURKUNDE BERLIN

Abstract

DiSSCo Prepare Deliverable D5.5 describes an analysis of DiSSCo requirements for interoperability with external taxonomic and geo-collections services and provides recommendations for technical and strategic approaches to building integrations during the construct phase of DiSSCo development. In earlier DiSSCo projects and work packages within the DiSSCo Prepare project, the integration of taxonomic checklists, represented by Catalogue of Life (COL) Checklist and the ChecklistBank infrastructure, into the DiSSCo architecture as a taxonomic backbone has been identified as a key requirement for users of DiSSCo services. Geo-collection data, represented in this review by GeoCASe, have also been highly underrepresented in terms of available services for data mobilisation and publication. During the task, conversations with the Earth Sciences community also identified Mindat as a key community resource for authoritative data about geological classifications and localities, and so that platform was added to the review scope of the task.

Reviews of each platform were carried out through a combination of desk-based research, document review and direct input from developers and representatives of the platforms. In parallel, user requirements were extracted from earlier DiSSCo outputs, and technical requirements from architectural design and pilot activities carried out by the DiSSCo developer team. Further detail was added to the picture by running an event-storming workshop, which engaged wider members within the community to help to identify and prioritise those events in external systems that would need to be reflected in the data within the DiSSCo architecture. From these activities, it was possible to make a comparison between the technical capabilities and resource capacity of the reviewed platforms on one hand, and DiSSCo's requirements and technical roadmap on the other.

DiSSCo's favoured technical model for sustainable interoperability with external services is based on the event publishing approach, originally proposed for integration with institutional collections management systems. An important message that emerged is that this would need investment and collaboration to achieve with GeoCASe and Catalogue of Life. Although both are enthusiastic about exploring further integration, GeoCASe is critically under-resourced, while Catalogue of Life would expect DiSSCo to partake in the joint investment of the sustainability and maintenance of its services, and key infrastructure called ChecklistBank that will be jointly governed by international biodiversity data initiatives and infrastructures. While there are more pragmatic approaches to integration that can be explored in the short term, further exploration of the longer-term relationships and opportunities for external investment form a key part of this report's recommendations.

Contribution to DiSSCo RI

This deliverable provides the DiSSCo RI with a set of recommendations for technical and strategic approaches to inform the development of DiSSCo integrations with external taxonomic and geo-collection services. It highlights priorities and specific challenges to address in the construct phase of the RI, and provides initial resource and cost estimates to include in the DiSSCo cost book.

Keywords

DiSSCo, GeoCASe, Mindat, ChecklistBank, Catalogue of Life, taxonomy, taxonomic names, collections, life science, earth science, data, interoperability

Index

Introduction

The modernisation of key services, especially of services for data currently underdeveloped in the DiSSCo community, is of great importance for the overall improvement of the technical readiness level from the DiSSCo RI.

This task is focusing on construction plans for the improvement of technical infrastructure in the identified key areas of geo-collection data and taxonomic names services. Geo-collection data are highly underrepresented in terms of available services for data mobilisation and publication. Thus, these services need special consideration in order to significantly increase DiSSCo technical readiness in the earth scientific domain. In addition, the harmonisation of life science taxonomic checklist services (e.g. Catalogue of Life) needs construction plans for the integration into the DiSSCo architecture in order to exploit their full value as a taxonomic backbone (meaning correct representation of scientific collections according to the most recent scientific consensus, or variations thereof, on taxonomic names).

This deliverable is intended to address these key services by providing construction plan recommendations for:

i. the mobilisation and publication of geo-collection data and integration of geological classification authorities, and

ii. the harmonisation and integration of taxonomic checklists.

These recommendations are intended to contribute to the DiSSCo architecture and services construction blueprint, the DiSSCo costbook, and the construction plans of the relevant community services.

Summary of recommendations

This section summarises the high-level recommendations resulting from the work in this task. For more detailed recommendations and discussion, please refer to the recommendation sections in the main body of the document.

1. **Apply a phased and pragmatic technical approach in working towards interoperability with external services.**

   An event-based approach represents an efficient and scalable method for interoperability in the long term, but is also dependent on third parties prioritising the required development and having resources available, which may not be easy to achieve without investment of funds and/or technical resources from DiSSCo into that development.

   In the shorter term, a more effective approach may be for the DiSSCo development team to continue to pilot more bespoke integrations with Catalogue of Life, GeoCASe and Mindat.

These quicker proofs of concept might also help to demonstrate the benefits of interoperability and strengthen the case for funding from external sources to work towards the more robust event-based architecture.

2. **Investigate strategic opportunities to address the funding and resource deficit in the geo-collections and geo-classification services.**

   There is a clear and critical resourcing issue for GeoCASe at present, with a very small amount of developer time currently available that is due to expire at the end of 2022. In the current situation, GeoCASe will have little if any capacity to implement any recommendations to meet DiSSCo integration requirements, and there is also a risk to the ongoing stability, maintenance, and support of the platform. DiSSCo may, by working in collaboration with the GeoCASe governance bodies, be able to help in addressing these issues and laying the foundations for a sustainable integration with a stable GeoCASe platform.

   Mindat also appears to have a dependency on limited developer resources, which may restrict the platform's ability to deliver the API development and server scalability that would be necessary to support interoperation with DiSSCo at scale. Although initial conversations are yet to be held, there may be opportunities for DiSSCo to work with Mindat to help to progress its development roadmap in this area, if this aligns with Mindat's strategic objectives.

3. **Engage more deeply with Catalogue of Life and the ChecklistBank programme through the Alliance for Biodiversity Knowledge.**

   Although DiSSCo may make use of COL's open tools for the community, there is greater long-term potential in developing a collaborative relationship with COL to better understand and support the needs for taxonomic services across the DiSSCo user base. There is also an expectation from Species 2000 that major users, like DiSSCo, that have dependencies on COL Checklist and ChecklistBank services will help to carry the financial sustainability for ChecklistBank and jointly maintain it as a global resource. COL will invite DiSSCo, just like GBIF and other biodiversity data infrastructures, to become part of the governance of the ChecklistBank infrastructure. Besides the strategic considerations, investment from DiSSCo is likely to be required if there is any expectation of development work in COL to pave the way towards the more robust event-based model of interoperability with ChecklistBank.

List of Abbreviations

ABCD (+EFG)	Access to Biological Collections Data (+ Extension For Geosciences)
BPS	BioCASe Provider Software
CETAF	Consortium of European Taxonomic Facilities
CETAF ESG	CETAF Earth Sciences Group
CMS	Collections Management System
COL	Catalogue of Life
ColDP	Catalogue of Life Data Package
DiSSCo	Distributed System of Scientific Collections
DwC-A	Darwin Core Archive
ECOI	European Collection Objects Index
ENA	European Nucleotide Archive
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GBIF	Global Biodiversity Information Facility
GeoCASe	Earth Science Collection Portal (Geoscience Collections Access Service)
ITIS	Integrated Taxonomic Information System
MOTU	Molecular Operational Taxonomic Unit
OTU	Operational Taxonomic Unit
PID	Persistent Identifier
RI	Research Infrastructure
UNITE	Unified system for rDNA sequences based identification of fungal species
WoRMS	World Register of Marine Species

Methodology

Requirements

Event storming activities

Event Storming[1] is a flexible workshop technique created by Alberto Brandolini for modelling and designing a process that consists of an unlimited number of **events** along a timeline (Glöckler et al. 2022). This technique allows sophisticated cross-discipline conversation between stakeholders with different backgrounds, delivering a new type of collaboration beyond silo and specialisation boundaries. The event, that could be any action, is triggered by an agent ("actor") doing a certain activity ("command"). Based on each event, one or many reactions ("responses") can be defined, which may be described as domain events as well. Thus, the process consists of a chain or network of interactions (events and responses) (**Figure 1**).

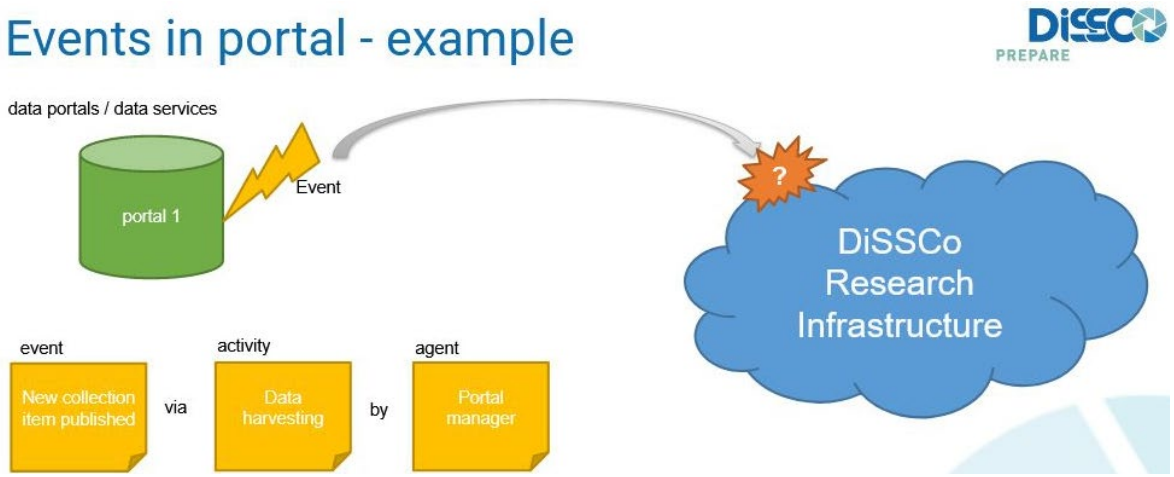

Figure 1: Example for an event that occurs in a data portal and can be considered by the DiSSCo RI for further processing.

In DiSSCo, the event storming method was already applied in the context of specific work processes with collections management systems (CMS) and the DiSSCo Research Infrastructure (RI). A description of the first event storming workshop and its results was published in Deliverable D6.1 (Glöckler et al. 2022). This method was identified in task 5.4 to be very useful to gather important events in data aggregators (e.g. GeoCASe and Catalogue of Life) and the DiSSCo RI as well.

Thus, an event storming workshop was organised for a mixed group of participants from different institutions, representatives and users of geo-collection services and taxonomic services in order to:

- brainstorm and aggregate all kinds of processes in domain-specific data portals or services of the DiSSCo RI,

[1] https://eventstorming.com/

- identify, prioritise, and document connection points between data portals, services and the DiSSCo RI,
- identify potential dependencies,
- collect events that could occur to a Digital Specimen in the CMSs and the DiSSCo RI,
- develop a process-based modelling approach, and
- derive recommendations for interoperability and technical readiness (e.g. standards, API guidelines).

The 33 participants have been asked to answer a few introductory questions. The results of this poll highlight the diversity of attendees with different roles and different knowledge about DiSSCo and the geo-collection services and taxonomic names services (**Figure 2**). No deeper knowledge on technical topics was required.

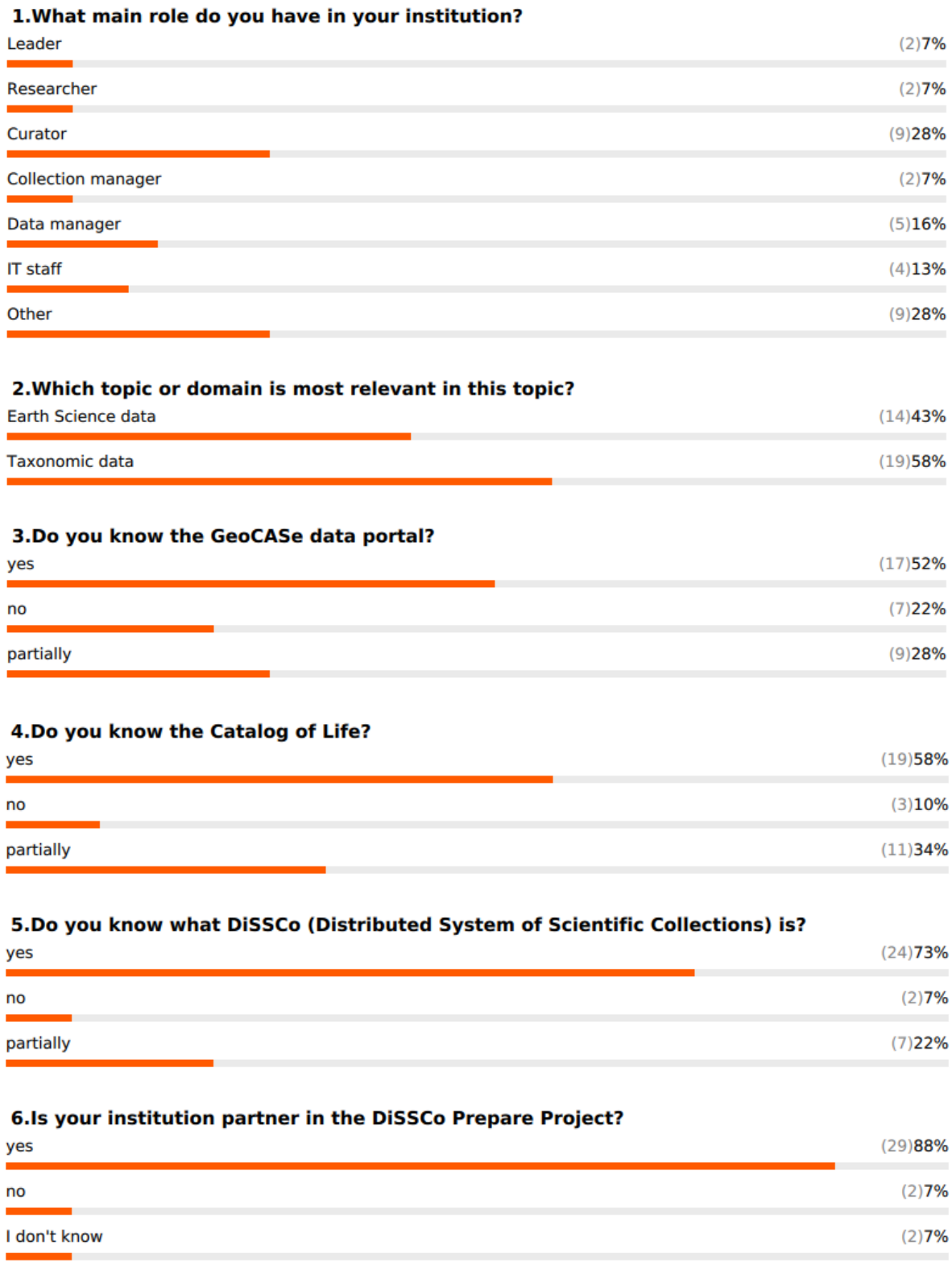

Figure 2: Results of an introductory poll for participants of the task 5.4 event storming workshop.

After a series of introductory talks on DiSSCo, GeoCASe, Catalogue of Life and the method used, the 33 participants were divided into four breakout groups. Two groups on events in taxonomic names

services and two groups on geo-collection specific events, respectively. The allocation of the people to the groups was leveraged by the participant's individual interests. Facilitated by dedicated moderators (members of the 5.4 task group team), the participants were asked to use a virtual whiteboard in the Miro online tool[2] that contained a prepared structure and a summary of instructions for the collaborative work. They were asked to list the events they considered relevant, based on their related work routines.

After the breakout sessions, the groups presented a summary of their results to the plenary. At the end of the workshop the participants were asked to assign stars (5 stars per person) in order to indicate their personal preference to the collected events. The individual results of the three breakout groups showed some overlap in the events listed, because people in different groups had similar ideas. In the post-processing of the workshop results, equal or very similar events have been aggregated to a unique list of events across the groups (for details see Appendix 1 in Glöckler et al. 2022). The stars assigned to these events have been added to the aggregated events as well. In the final and aggregated outcome, the events with the most stars represent a ranking that can be considered as a priority list for the DiSSCo pilot.

DiSSCo Prepare Work Package 1 user story analysis

Within DiSSCo Prepare Work Package 1 ('User needs and socioeconomic impact'), Tasks 1.1 ('Analyse life sciences use cases and user stories') and 1.2 ('Analyse earth sciences use cases and user stories') involved the aggregation and analysis of user stories from DiSSCo stakeholders within the Life Sciences and Earth Sciences domains.

After quality filtering and de-duplication, 317 user stories were identified for Life Sciences. These were analysed and linked to a categorised set of functional demands, and the frequency of each demand reported. As **Figure 3** below demonstrates, the top five frequencies identified were for tools for data discovery (100 user stories), distribution data (75) and morphological data (59), and metadata at the collection (58) and record (54) level.

[2] https://miro.com

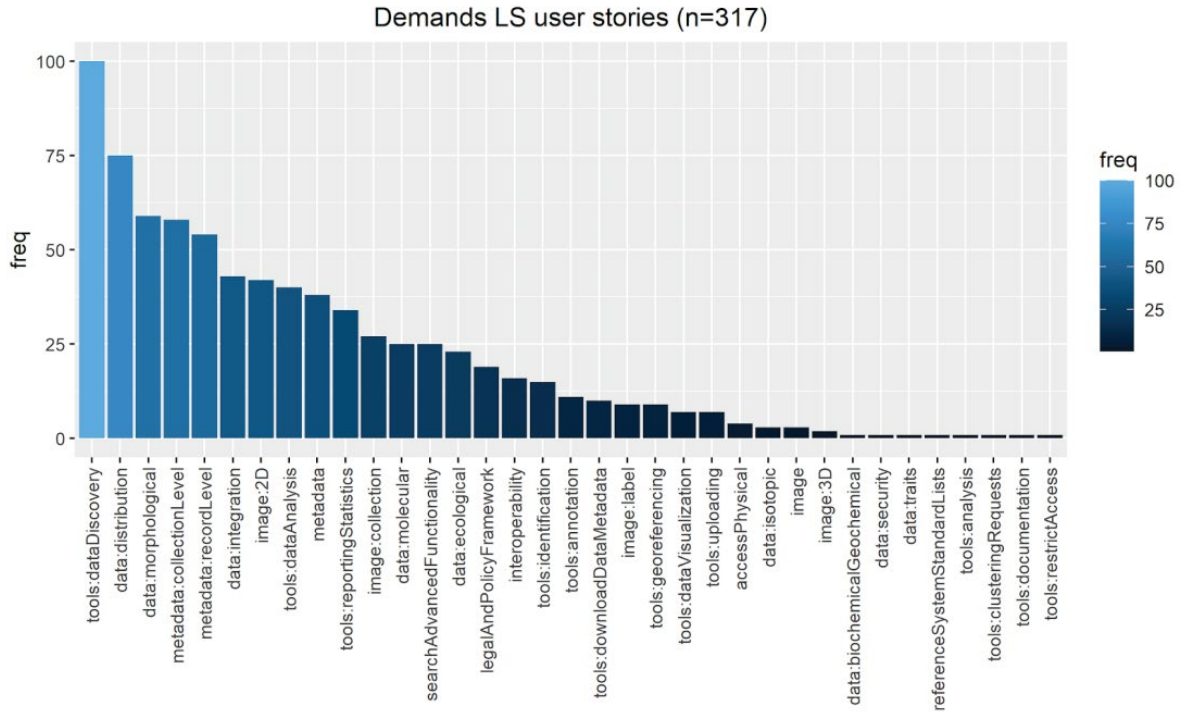

Figure 3: Number of times each of the 35 demands was mentioned in Life Sciences user stories (from Fitzgerald et al. 2021).

For Earth Sciences, 128 user stories were identified and analysed using the same methodology (**Figure 4**). In this set, the five most frequent demands were for collection-level metadata (42 user stories), advanced search functionality (31), data integration (30), tools for reporting & statistics (29) and record-level metadata (23).

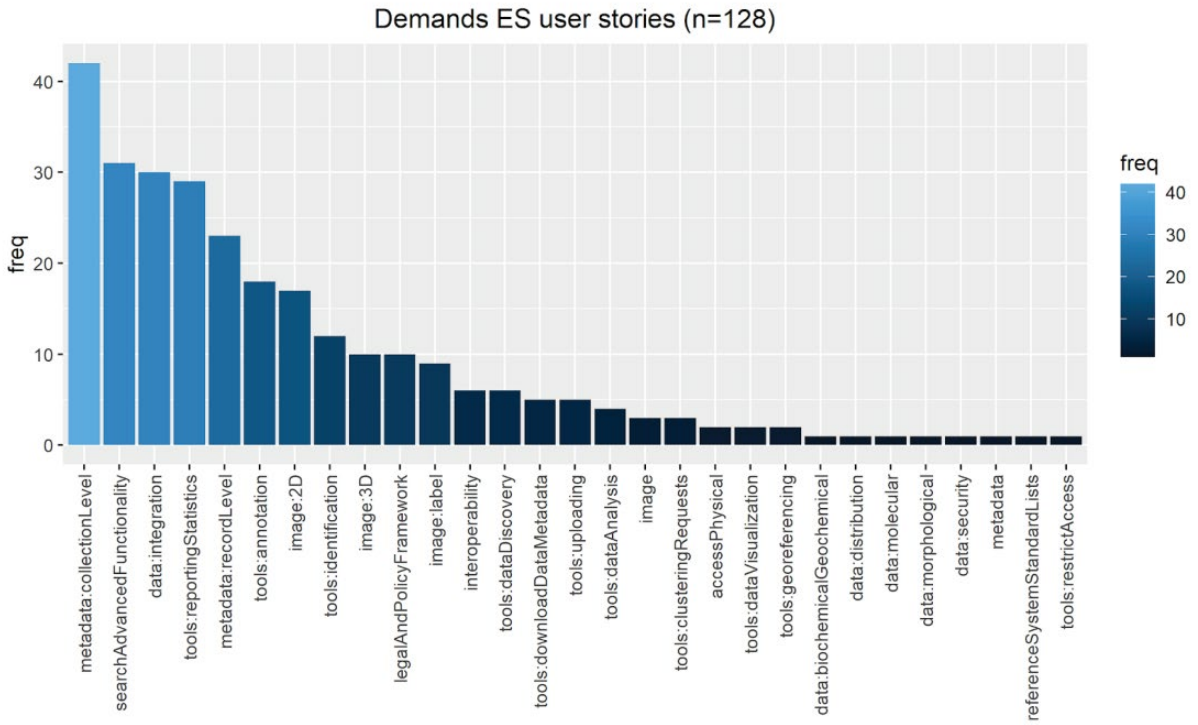

Figure 4: Number of times that each of the 29 functional demand (sub-)categories was mentioned in Earth Sciences (ES) user stories (from von Mering et al. 2021).

Detailed information on the methodology and results can be found in the DiSSCo Prepare deliverable reports D1.1 (Fitzgerald et al. 2021) and D1.2 (von Mering et al. 2021).

Research

Methods used to research the various platforms involved in the assessment included:

- reviewing online material, including websites and online documentation, published papers and presentations,
- exploring GeoCASe and Catalogue of Life API capabilities using Google Colab notebooks, and
- virtual meetings with representatives of GeoCASe and Catalogue of Life.

These activities were used to make an assessment of the capabilities and status of each platform, and used in conjunction with requirements to develop diagrams of potential interoperability between DiSSCo and the external services.

Several initial technical pilots were also carried out by the DiSSCo core development team at Naturalis in alignment with the task, which helped to explore the technical elements and potential challenges around integration with third party platforms and services.

This task also drew upon several other DiSSCo Prepare deliverables, in particular:

- D1.1 Report on life sciences use cases and user stories (Fitzgerald et al. 2021)

- D1.2 Report on Earth sciences use cases and user stories (von Mering et al. 2021)
- D6.1 Harmonization and migration plan for the integration of CMSs into the coherent DiSSCo Research Infrastructure (Glöckler et al. 2022)
- D6.2 Implementation and construction plan of the DiSSCo core architecture (Leeflang et al. 2022)

DiSSCo model for interoperability with external services

A technical design for the DiSSCo data infrastructure has been described in DiSSCo Prepare deliverable D6.2 (Leeflang et al. 2022). Within this infrastructure, there are three main areas of the proposed architecture that intersect with external platforms and services.

1. **Data ingestion**

   DiSSCo's primary data will be sourced from collections management systems (CMSs), digitisation pipelines and digitally born specimens of DiSSCo facilities (natural science/history collections and related third party organisations). These data are expected to pass through a translator service in the DiSSCo architecture in order to be translated into the required structure for DiSSCo Digital Specimens.

   This component is anticipated to source the primary data from collections-holding organisations that would ultimately be provided by DiSSCo to GeoCASe.

2. **Data enrichment**

   The data enrichment component of the architecture will contain services that enrich existing Digital Specimens by attaching additional data as annotations. These services may be fully automated (potentially involving AI components), manual processes or some combination of the two.

   It is anticipated that this component of the DiSSCo architecture will be the most relevant to integration with external taxonomy or geological classification services.

3. **DiSSCo services**

   The DiSSCo services component contains key elements of the architecture for interoperability (**Figure 5**). These include the Application Programming Interface (API) that exposes DiSSCo data to external users, and the event publisher (described further below).

   The API and event publisher are ultimately intended to be the primary mechanism for exposing both data and metadata about data changes to GeoCASe and the Catalogue of Life.

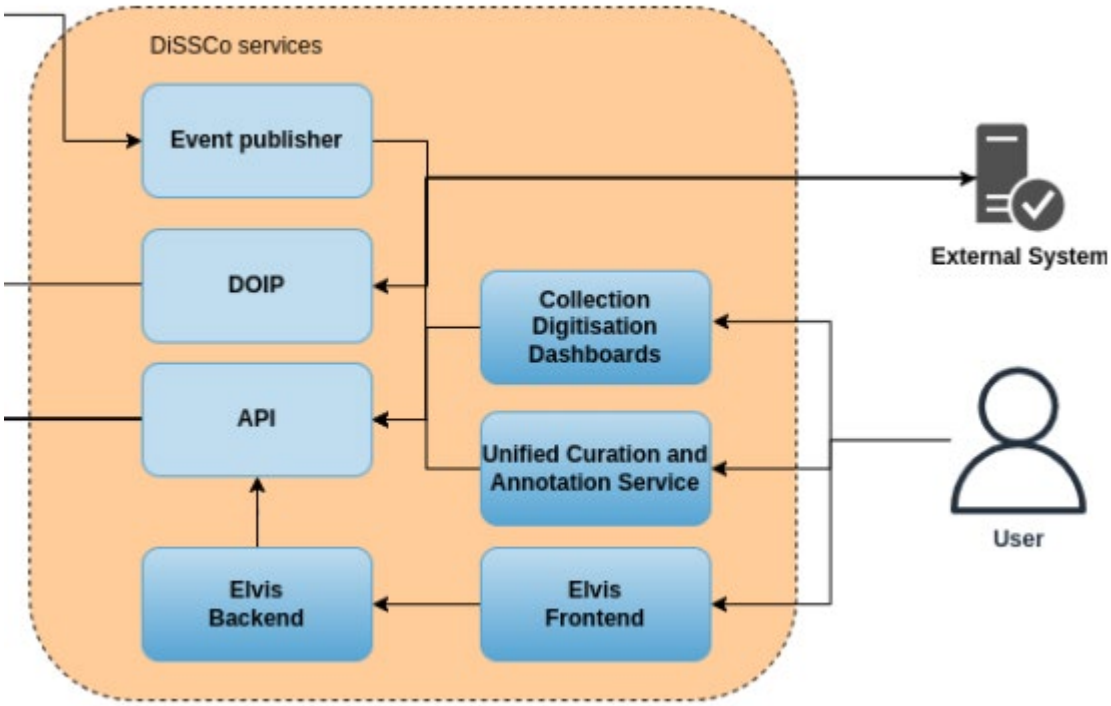

Figure 5: Extract from the DiSSCo architectural overview showing key areas (event publisher and API) for the integration of external services (from Leeflang et al. 2022).

Event-driven interoperability

Under DiSSCo Prepare Task 6.1 ('CMS systems interoperability and harmonisation'), an event-driven approach has been proposed to support the synchronisation of data between the DiSSCo core architecture and collections management systems of data providers. This included the use of the CloudEvents specification to expose metadata about data changes relating to both the DiSSCo Digital Specimens and the records in the source systems. Within the CloudEvents wrapper, there are various options currently under assessment for structuring and serialising the event metadata itself, including representing each event as a Fair Digital Object. The event metadata can then be referenced to determine which data operations (create, update, delete) are required against which records to maintain appropriate data synchronisation between platforms. Further details are available in the DiSSCo Prepare D6.1 deliverable report (Glöckler et al. 2022).

The event publisher (mentioned in the previous section and shown in **Figure 5**) is the component of the DiSSCo architecture design intended to support this event-driven approach to interoperability. Essentially, the event publisher will provide information on changes that have happened to a record (identified by a unique persistent identifier), which can be used to make the appropriate calls to the API to access the modified data and process the data changes locally.

This approach has potential for wider application than the CMS integration use case, also forming the basis for a general model of interoperability between DiSSCo and external platforms (**Figure 6**). This model has been used to inform the requirements and recommendations in this document, but it

must also be recognised that there would be related development required on the side of the external platforms as well as DiSSCo to implement this approach, and so other pragmatic approaches should be taken into account.

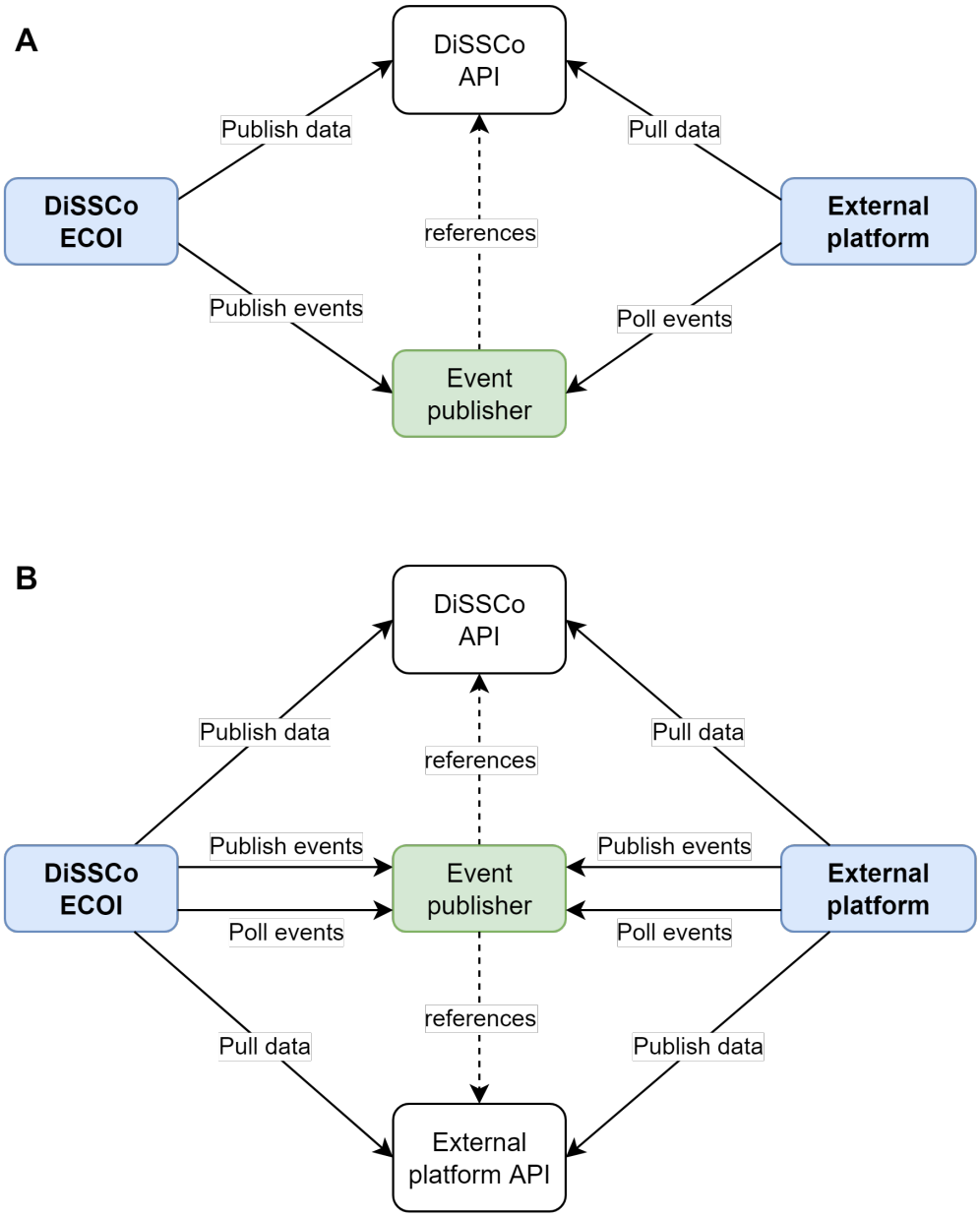

Figure 6: A summary of the event publishing approaches for A. uni-directional (from DiSSCo) and B. bi-directional data flow between DiSSCo and external platforms.

Business cases for interoperability

Catalogue of Life

Catalogue of Life (COL) is an international collaboration bringing together the effort and contributions of taxonomists and informaticians from around the world. COL aims to address the needs of researchers, policymakers, environmental managers and the wider public for a consistent and up-to-date listing of all the world's known species and their higher taxa. In December 2020 COL launched (in collaboration with GBIF) its new infrastructure, consisting of a public portal[3], ChecklistBank[4] and a set of APIs[5]. A more detailed description of COL and ChecklistBank can be found in **Appendix A**.

The recommendations from the DiSSCo Prepare analysis of Life Sciences use cases (Fitzgerald et al. 2021) highlighted the importance of taxonomic names in one of the primary demands, data integration:

"Data integrity and interoperability would be improved by incorporating existing, discipline-specific digital services in the development of further digital collection systems. For example, incorporating vocabularies derived from taxonomic name lookup and resolution services such as the GBIF Species API would make reporting on life sciences collection access requirements more efficient, granular, and repeatable. It would also facilitate identification of data gaps and feed into transnational digitisation prioritisation and planning."

While the categorisation within the analysis didn't focus specifically on taxonomic names services, this is arguably because taxonomic classification of specimens is a fundamental part of many of the user stories articulated. The most common demand identified (**Figure 3**) was for tools for data discovery, and taxonomic names are a (if not the) primary descriptor used for most use cases about search and discovery of biological collection specimens. Taxonomy is also implicit in some of the most common demands relating to data (morphological, distribution), where the value of those data are tied to the classification of the specimens to which they apply.

For DiSSCo, taxonomic names services provide taxonomic and nomenclatural data which are an essential part of the Digital Specimen. DiSSCo needs to match and look up species names attached to the specimen received from the source system (often a collections management system) against one or more taxonomic checklists. Taxonomic names services also provide stable taxonomic name identifiers, and these are the route for linking to many other elements of an enriched Digital Specimen.

[3] https://catalogueoflife.org/
[4] https://www.checklistbank.org/
[5] https://api.checklistbank.org/

GeoCASe

The Earth Science Collection Portal, commonly known as GeoCASe (Geoscience Collections Access Service), is a data network and web portal designed to make collections of minerals, rocks, meteorites and fossils held in museums and research institutions universally available online, in order to foster scientific research and collaboration internationally (GeoCASe: The Earth Science Collections Portal 2022). The platform has a global scope, and offers users functionality for searching across institutional datasets, downloading data and accessing data programmatically through an API. A more detailed overview of the platform can be found in **Appendix B**.

Within the DiSSCo strategy (in development at the time of publication), integration with geo-collection services falls under the 'Access' programme. From a DiSSCo perspective, the business case for integrating DiSSCo with GeoCASe appears to be predominantly strategic, to contribute to the global aggregation and discovery of Earth Science collections rather than creating a new silo of collection data for participating European institutions. As a platform specifically targeted to the Earth Sciences domain, GeoCASe has existing credibility within the Earth Science collections community.

As a substantive requirement has not yet been identified for DiSSCo to ingest data from GeoCASe to supplement Digital Specimens within the DiSSCo repository, data would be intended to flow in one direction from DiSSCo to GeoCASe. At the current stage of development, GeoCASe includes a clear and fairly standard set of search, view and download user interface functionality, but doesn't yet offer richer end-user functionality and data services beyond the functionality that is likely to be developed by DiSSCo for discovery and access to Digital Specimens.

There is also an understandable reluctance among data providers to have to set up multiple processes to provide the same or similar data to multiple aggregators. Providing a single avenue for publishing to both platforms via DiSSCo is therefore another strategic imperative and avoids providers potentially needing to make a choice between the two, which could undermine the business case of both platforms and lead to a more fragmented data ecosystem. As GeoCASe will continue to have data providers that are both within and outside of the DiSSCo membership, and so the technical option will remain for DiSSCo member institutions to publish to GeoCASe directly, an agreement with GeoCASe and clear communication to DiSSCo members may be needed to ensure that the appropriate path for data publication is taken.

It should be noted that there is duplication of scope between GeoCASe and GBIF in terms of the publication of palaeontology specimen data. Under the assumption that DiSSCo will also contribute biological and palaeontological specimen data to GBIF, it would therefore be only the geological (minerals, rocks and meteorites) specimen data from DiSSCo that would not be published to a global aggregator, should DiSSCo be unsuccessful in setting up an effective integration with GeoCASe.

Mindat

Mindat offers a comprehensive resource of mineral and rock classifications (including images and supplementary data including composition, physical and chemical properties etc.), localities and

occurrences at those localities. The data are provided and curated by a large global resource of expert contributors.

Mindat has significant credibility within the Earth Sciences community, lending a strategic angle to the business case for DiSSCo integration. While there haven't been formal requirements captured in previous DiSSCo activities for integration with geological classification services, remarks from a number of members of the geosciences community have suggested that the data provided by Mindat are a key component of their activities. GeoCASe also already maintains links to classifications in Mindat for many of its mineral and rock specimen records. These factors indicate that there may be user requirements in this area that haven't been exposed in previous requirements capture activities in the ICEDIG and DiSSCo Prepare projects, possibly due to underrepresentation from the geosciences part of the collections community.

There are likely to be some similar use cases in linking specimens to geological classifications and localities to those identified around linking biological specimens to taxonomy and taxon distributions. This applies to both enriching the extended specimen (with reference to Minimum Information for a Digital Specimen (MIDS) levels), and improving data quality through validation and verification.

Taxonomic services

Catalogue of Life and ChecklistBank

Overview

Catalogue of Life (COL) is an international collaboration bringing together the efforts and contributions of taxonomists and informaticians from around the world. COL aims to address the needs of researchers, policymakers, environmental managers and the wider public for a consistent and up-to-date listing of all the world's known species and their higher taxa (the COL Checklist). While the production of the COL Checklist is governed by Species 2000, the organisation that also represents the taxonomic community underpinning the COL Checklist, the joint ChecklistBank infrastructure will be governed by the Catalogue of Life organisation. Both organisations, Species 2000 and Catalogue of Life are strongly tied together.

Requirements

The user stories collected and analysed in DiSSCo Prepare deliverable D1.1 (Fitzgerald et al. 2021) highlighted the need for the verification and linkage of taxonomic (name) information as part of the enriched Digital Specimen. The event storming workshop identified in more detail the events in COL that should be reflected by actions within the DiSSCo data architecture (**Table 1**). Within these events, the highest priorities were related to changes to names within the COL Checklist and ensuring that these were reflected in the information attached to the Digital Specimen.

Table 1: Events identified in the Task 5.4 event storming workshop relating to taxonomic names services (as part of the COL Checklist). Actions highlighted in green show the events that received the most votes from workshop participants as highest priority.

Subject	Action	Trigger	Example
taxonomic name	changed	revision / publication	
	created	revision / publication	
	not found	does not match COL	fossil name matching or different taxon tree / reference concept or unpublished name or misspelt
	merged		
taxonomic checklist	added		
	cited	publication / digitization	
	removed		not participating in COL any longer
determination	added		
	changed		
type designation	has conflicts	check sources / literature	
common name	requested to add	Wikidata linking	

For DiSSCo taxonomic names services provide taxonomic and nomenclatural data which are an essential part of the Digital Specimen. DiSSCo requires a function to match and look up species names attached to the specimen received from the source system (often CMS) against one or more checklists. This could be done either on ingestion, update or manually triggered by the user. DiSSCo does not intend to build these services itself but to use and build upon existing services. Part of the matching and harmonisation of taxonomic checklists will take place outside the DiSSCo architecture, and only the results of the harmonisation are integrated. The COL Checklist and ChecklistBank may provide necessary taxonomic names services. It needs further exploration as to which other taxonomic checklists DiSSCo requires for a proper representation of its mediated specimen collection data.

DiSSCo also requires taxonomic names services to be accompanied with stable taxonomic names identifiers. The retrieval of the taxonomic name identifiers can be an automated annotation service which makes a request to a taxonomic name service, retrieves the information and stores this in an annotation. This annotation can then be accepted by a user with sufficient rights and added to the Digital Specimen data record. The harmonisation between different taxonomic name identifiers is done outside the DiSSCo architecture, and only the harmonisation is integrated. The COL Checklist provides stable taxonomic name identifiers and is looking into a scalable approach for taxonomic concept identifiers (still early days). ChecklistBank may be an avenue for the harmonisation of taxonomic name identifiers coming from different data sources.

Having one consensus classification for DiSSCo, such as the COL Checklist, to supply discovery and access services for scientific specimen collections based on scientific names will enable users' quick access to data, both for science and policy. Additional taxonomic names services to deviate from the consensus classification can also be provided.

ChecklistBank will enable users to quickly compare taxonomic checklists. COL and GBIF work towards a semi-automated part of the COL Checklist that in future can replace the GBIF BackBone Taxonomy. This also involves the integration of taxonomic name information from digitally published literature mediated through Plazi (and including articles from Pensoft Publishers and the European Journal of Taxonomy), and (M)OTU data from the International Barcode of Life (iBOL), ENA, and UNITE. COL/GBIF and ENA currently collaborate to have a more encompassing mapping between the COL Checklist and the NCBI Taxonomy. One of the main issues is that the latter does not take authorship into account (snapping an accepted species binomial with an authorship to a synonym; e.g. *Quercus robur* L. becomes a synonym of *Quercus robur*). NCBI Taxonomy also does not show a full classification, with all potential synonyms associated as children to an accepted species parent. In other words, for the use case '*give me all sequences associated with an accepted scientific name*' a user needs to put in a series of queries involving all known synonyms of that accepted species separately. Through the ENA and COL collaboration a more user-friendly service will be investigated.

DiSSCo risks a similar situation as described in the example above if it does not provide a discovery and access system for taxonomic names based on a consensus classification. Enabling searches through higher taxonomic ranks, e.g. genus and family level, seems also advisable for DiSSCo.

Species 2000 envisions various interactions at various levels between COL/ChecklistBank and DiSSCo, for example:

1. At Natural History Collection institute level,
2. At the level of the specimen data refinery,
3. At the DiSSCo central level.

At the level of Natural History Collections, several examples of the potential interactions could be given. For example, the COL hierarchy may be used in collection management systems (CMS). At the same time, Natural History Collections may use their own taxonomic lists to make their holdings of scientific specimen collections accessible. Such taxonomic lists could be published to ChecklistBank to

enable better mapping with the COL Checklist as well as other taxonomic lists or policy relevant species lists. Natural History Collections also house some of the taxonomic communities that aid in maintaining a global authoritative database of a specific taxonomic sector (e.g. also in the COL Checklist). These institutions also maintain valuable information, such as type specimen indexes and published references.

A specimen data refinery as developed by the SYNTHESYS+[6] project and envisioned within the DiSSCo technical plan, provides interesting interactions between specimens and scientific names. On the one hand, correct scientific names can enhance the information held in specimen collections. In turn, a specimen data refinery is also likely to deliver information that may become part of ChecklistBank as well as a semi-automated part of the COL Checklist.

Interactions at the central DiSSCo level could be varied. For example, the COL Checklist and ChecklistBank may provide taxonomic names services to enable a DiSSCo discovery and access service to users. A DiSSCo central service may also be a central information channel to ChecklistBank / COL for validated information, such as type specimens, taxonomic species lists, person information, references, distribution etc. A DiSSCo central organisation also could promote that the taxonomic effort of maintaining a specific taxonomic sector in ChecklistBank or the COL Checklist by a Natural History Collection or a country is seen as a key contribution to DiSSCo.

The requirements below use a MUST/COULD/SHOULD prioritisation approach loosely based on the MoSCoW method[7].

1. DiSSCo MUST make Life Sciences Digital Specimen data, including taxonomic information from the source data providers, openly available to external consumers and aggregators

    1.1. Changes to the institutional data ingested by DiSSCo MUST be reflected in the event publisher and API within an appropriate time frame

    1.2. Appropriate metadata SHOULD be provided via the event publisher to enable data consumers to detect when a record has changed

    1.3. Appropriate metadata COULD be provided via the event publisher to enable external consumers to detect which data items have changed

    1.4. DiSSCo MUST include clear type designations in the Digital Specimen data model

    1.5. DiSSCo SHOULD include events specific to type specimens in the event publisher

2. DiSSCo MUST support taxonomic properties and storage of taxonomic name identifiers and links in the Digital Specimen data model

[6] https://www.synthesys.info/
[7] https://en.wikipedia.org/wiki/MoSCoW_method

3. DiSSCo SHOULD develop an enrichment service to resolve classification strings against COL APIs, and add links and other appropriate annotations to Digital Specimens

4. DiSSCo SHOULD provide a discovery and access system for taxonomic names based on a consensus classification, including searching on higher taxonomic ranks

5. DiSSCo SHOULD leverage taxonomic services for classification at the collection level as well as the Digital Specimen level

Interactions with DiSSCo

Data flows, workflows and integrations

Specimen data is ingested by DiSSCo from the source system. Data administrators can select which automated annotation service they want to run on data ingestion. If the data administrator selected one or more taxonomical services these services will run after the specimen has received a PID and has been added to DiSSCo.

The triggered automated annotation services will make a call towards the taxonomic names service requesting the taxonomic name identifier. It will wrap this response into an annotation which will be attached to the specimen object. If the annotation has been accepted the taxonomic name identifier will become part of the actual Digital Specimen.

Additionally, users with a specific role will be able to request a run of an automated annotation service against their selected specimen or set of specimens. Through this workflow a user can request information for one or multiple taxonomic names services.

Technical and architectural approaches

Figure 7 below summarises a potential approach to COL integration following the event-based model. An automated annotation service needs to be created which runs for the taxonomic service. It needs to make a request with the taxonomic information of the specimen. The response will be wrapped in the correct annotation type and added to the specimen.

The annotation approval workflow is a generic part of the DiSSCo core infrastructure and would not require a specific implementation.

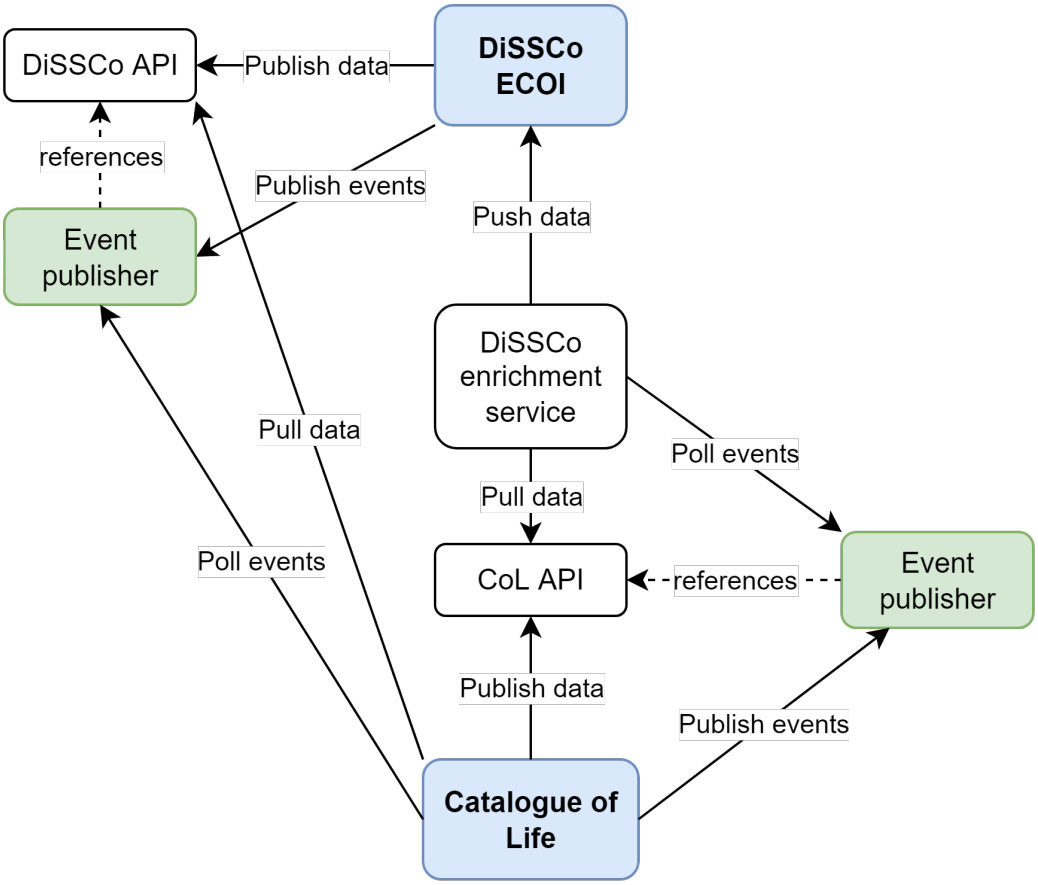

Figure 7: Summary of a potential interoperability model with COL taxonomic names services, based on an event-driven architecture. Alternative models are also in discussion that might involve COL actively pushing data changes to DiSSCo.

Gap analysis against requirements

Data interoperability

Substantial development in the last few years on the new ChecklistBank infrastructure and integrated tools means that COL is in a very strong position technically to deliver on DiSSCo requirements. However, a full implementation of the event-based model of interoperability would still require some development on the part of COL to generate appropriate event metadata, push it to an event publisher and link that information back to the appropriate data in the ChecklistBank API.

Support and collaboration

COL's data and tools are openly available for community use. However, Species 2000 has made clear that, for an infrastructure of DiSSCo's scope, a more collaborative approach and consideration of mutual benefits would be required if there were any expectation of additional support or development required to meet DiSSCo's needs.

Fossil taxonomy

A recent review by members of the palaeontology collections community highlighted some areas where fossil taxonomy can be problematic in a consensus classification of living organisms such as used by the COL Checklist (Little et al. 2022). With some groups a good fit can be made with an existing tree of life but in other palaeontological groups this proves to be challenging. The same issues would therefore be faced in the use of COL taxonomic names services for palaeontological specimens in DiSSCo. But there is currently no global solution for this problem. COL and GBIF are keen to discuss with the palaeontological community a sustainable way forward through ChecklistBank.

Recommendations

Recommendations for DiSSCo Construct

1. Focus on articulating more detailed requirements for taxonomic names services.

    - Species 2000 would recommend a further in-depth exploration with DiSSCo into the requirements for taxonomic names services. At present, although it's clear that DiSSCo would offer users discovery and access to specimen collection information through taxonomic names (for specimens that have taxonomic names attached), the details of what these services would look like are not yet clearly defined. A more detailed base of documented use cases for taxonomic names services as discovery and access mechanism for specimen collection information in DiSSCo is desired. This should really target users of specimen collection information as well as taxonomists.

    - Engage with COL to further explore the benefits that COL and ChecklistBank might leverage from interoperability with DiSSCo, and resource implications of supporting those requirements in the DiSSCo architecture.

2. Further engagement with GBIF and COL in the context of the Alliance for Biodiversity Knowledge.

    - Continue to support the international effort by GBIF and COL to build one common infrastructure for taxonomic names through ChecklistBank.

    - Engage with GBIF and COL to support the implementation of PIDs for taxonomic names and concepts.

3. Assess and refine the potential technical approaches to interoperability with ChecklistBank.

    - Use technical pilots to assess the possibilities and limitations of using COL's public services and open APIs to meet DiSSCo integration requirements.

- Engage with COL to explore the potential implementation of the event-based interoperability model, and any associated investment that might be required into COL development.

4. Engage with the Paleo Data Working Group (Krimmel et al. 2021) on the representation of fossil taxonomy in the Digital Specimen data model and consultation with COL on improvements in ChecklistBank. It is likely this effort needs proper global resourcing, especially when other products than the current COL Checklist in ChecklistBank are considered.

Geological collection and classification services

GeoCASe

Overview

The Earth Science Collection Portal, commonly known as GeoCASe (Geoscience Collections Access Service), is a data network and web portal designed to make collections of minerals, rocks, meteorites and fossils held in museums and research institutions universally available online, in order to foster scientific research and collaboration internationally (GeoCASe: The Earth Science Collection Portal 2022).

GeoCASe was initially built by the Museum für Naturkunde (MfN), Berlin in 2007 as part of the EC-funded SYNTHESYS programme, and currently contains just over 1.7 million specimen records contributed by 10 institutions. While at present these contributors are all European institutions, GeoCASe is intended to have a global scope.

More detailed information about GeoCASe can be found in **Appendix B**.

Requirements

The event storming workshop identified in more detail the events in integrated geo-collection services that should be reflected by actions within the DiSSCo data architecture (**Table 2**). Within these events, the highest priorities were given to notification of references to records representing the same specimen in other platforms, and additions or updates to geographic localities.

The focus of the event storming workshop was on the events in external services that would need to be detected and acted upon by DiSSCo. The expectation is that (at least initially) data would flow in a single direction from DiSSCo to GeoCASe, so many are not relevant for the specific GeoCASe use case at this point in time. However, they remain relevant for geo-collection services in general, and possible future evolution of a relationship with GeoCASe. Many also work in reverse - highlighting events that could occur in DiSSCo Digital Specimens that should be exposed for external consumers.

Table 2: Events identified in the Task 5.4 event storming workshop relating to geo-collection services. Actions highlighted in green show the events that received the most votes from workshop participants as highest priority.

subject	action	trigger	example
specimen record	added		
	referenced		links to the same record in other portals
	linked		links between different specimens
	has different versions		different versions in different places
	annotated		
physical specimen	removed		removed from collection or destroyed
	status changed		broken
specimen image	updated		
	added	digitization	
	cited	revision / publication	
	annotated		
specimen identifier / reference	changed	system architecture change	
	added		e.g. IGSN
loan	requested		via the portal and/or ELViS
institutional metadata	changed		e.g. contact data updated

geographic distribution	changed	revision	
geographic locality	changed		filling data gaps
	added		filling data gaps
stratigraphy	changed		filling data gaps
	added		filling data gaps
(web) service	not available		
	replaced by new service		name providing service
data standard	changed		e.g. new fields or new version in ABCDEFG

The requirements below use a MUST/COULD/SHOULD prioritisation approach loosely based on the MoSCoW method[8].

1. DiSSCo MUST make Earth Sciences Digital Specimen data openly available to external consumers and aggregators

    1.1. Data MUST be in a format and structure that can be mapped to ABCD(EFG) for GeoCASe ingestion

    1.2. Data MUST be made available through a mechanism that GeoCASe can access for automated ingestion

    1.3. Appropriate metadata MUST be provided to enable GeoCASe to identify the earth science records that are within that platform's scope

    1.4. Appropriate metadata SHOULD be provided to allow GeoCASe to determine which records are of a suitable level of completeness and quality to ingest

    1.5. Changes to the institutional data ingested by DiSSCo MUST be reflected in the event publisher and API within an appropriate time frame

[8] https://en.wikipedia.org/wiki/MoSCoW_method

1.6. Appropriate metadata SHOULD be provided via the event publisher to enable data consumers to detect when a record has changed

1.7. Appropriate metadata COULD be provided via the event publisher to enable external consumers to detect which data items have changed

2. DiSSCo SHOULD make Digital Specimen media and images available to GeoCASe through persistent resolvable links

3. DiSSCo MUST apply appropriate filters to prevent GeoCASe from ingesting any data that is sensitive and subject to access restrictions

4. DiSSCo MUST provide PIDs to enable GeoCASe to link back to DiSSCo Digital Specimens and organisations

5. DiSSCo MUST supply metadata on licensing and copyright for GeoCASe to decide which data and images to include

Interactions with DiSSCo

Data will flow from the source system, often the CMS, into DiSSCo. To enable this DiSSCo will start with a regular harvest of the data in the source system. In a later step DiSSCo hopes to integrate with the source system through the event-based structure proposed in DiSSCo Prepare deliverable 6.1.

After ingestion DiSSCo will mint PIDs, add the data to its database and indexing engine after which we will send out a CreateUpdateDeleteEvent and trigger the requested automated annotation services. The CreateUpdateDeleteEvent is stored to create a change log, which will form the basis for data provenance, traceability, versioning, reversibility and restartability within the DiSSCo data architecture. Further details of this structure will be available in the DiSSCo Data Management Plan (DiSSCo Prepare deliverable D6.4). The change log will be stored within the DiSSCo infrastructure, but CreateUpdateDeleteEvents will also be published to external systems.

One of these external systems could be GeoCASe. GeoCASe can receive the CreateUpdateDelete from DiSSCo and respectively create, update, or delete the specimen data in their data storage layer. This event will also have the PID of the object which enables GeoCASe to display the specimen PID in their frontend.

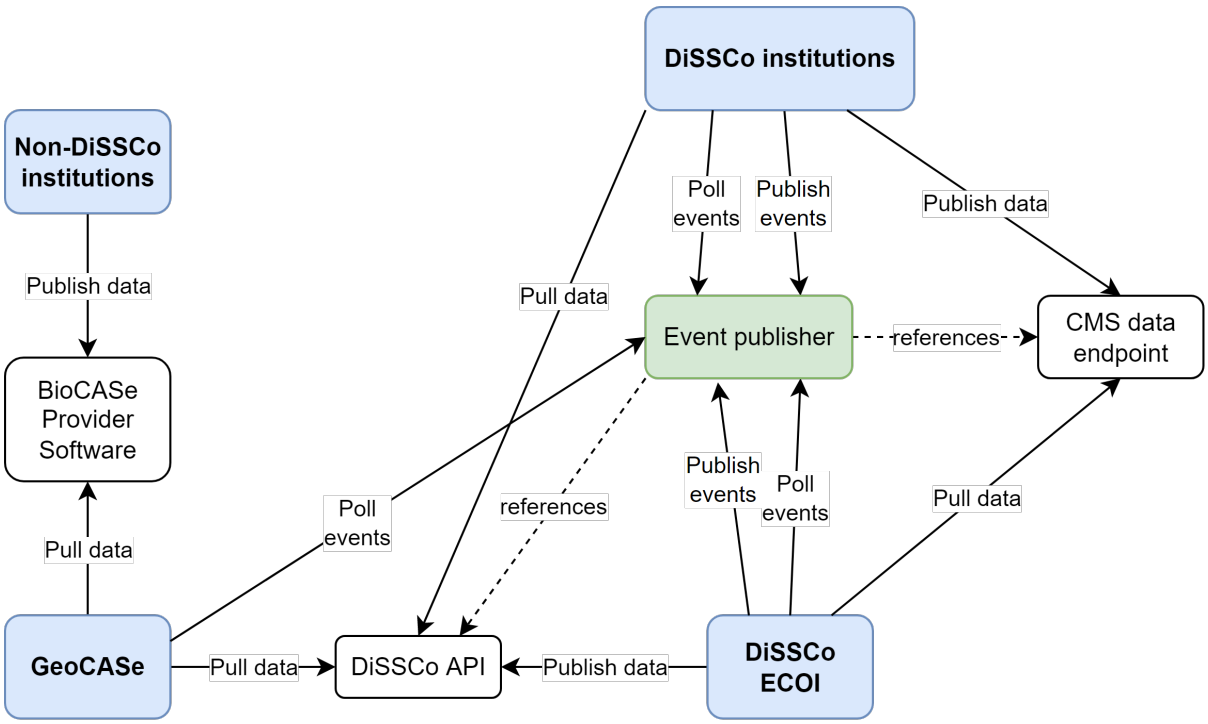

Figure 8: Summary of a proposed interoperability model with GeoCASe based on an event-driven architecture, incorporating DiSSCo interoperability with institutional CMS platforms. Alternative models are also in discussion that might involve DiSSCo actively pushing data to GeoCASe.

Within the DiSSCo core architecture we view GeoCASe as an external system which will make use of the information in DiSSCo. This means that DiSSCo will be one of the data suppliers for GeoCASe. DiSSCo itself will gather data at the institution level through the specific endpoint that exposes Geoscience Collections. DiSSCo provides the infrastructure which creates the unique digital identifier (PID), enables data harmonisation, enhances data quality, and further enriches the specimen information. This extended Digital Specimen can then be provided to GeoCASe which will receive harmonised high-quality data.

The benefits for GeoCASe will be at data ingestion level. GeoCASe can expect high quality, harmonised data from DiSSCo. Additionally, DiSSCo will provide enrichments created through the automated annotation services and the Unified Curation and Annotation Service (UCAS).

For DiSSCo it would be beneficial if we can create an easy integration with GeoCASe. Within DiSSCo we will work with an eventing structure as proposed in deliverable 6.1. This means that instead of doing regular compute heavy full dataset checks we want to receive and produce notifications when there are actual changes. Changes in our own infrastructure will be stored internally creating a change log but will also be published to external systems. This will significantly reduce the compute power needed to check if specimen data has changed. It will also ensure that changes in DiSSCo will quickly propagate to GeoCASe, keeping both systems in sync and up to date to the latest version of the data. It would be beneficial for both DiSSCo and GeoCASe if we could work together with GeoCASe to implement this event-based communication.

If GeoCASe was unable to implement an eventing structure, DiSSCo would need to evaluate the impact of setting up a BioCASe endpoint with DiSSCo for GeoCASe. This will require both extra work in man hours, extra maintenance, and extra infrastructure costs. Data will be out of sync as there will be time between the creation in DiSSCo and the data harvest from GeoCASe.

Pilot and development activities

A pilot for the ingestion of ABCD(+EFG) data is running within the DiSSCo infrastructure. The DiSSCo infrastructure can now receive and parse the geological data in ABCDEFG format. During development some issues were encountered with regards to the use of the "recordBasis" field in the ABCDEFG data.

Within the DiSSCo infrastructure we are only interested in certain record basis types, such as Preserved Specimen and Fossil Specimen, other types such as HumanObservations fall outside of the scope of DiSSCo. Within the ingested data the recordBasis used did not comply with the values accepted in the ABCD standard, examples of these are Rock Specimen or Meteorite Specimen.

This indicated that the current ABCD enumeration for recordBasis is not in line with the needs of the geoscience community. This issue has been addressed in the ABCD community and will be discussed further.[9]

Gap analysis against requirements

Data interoperability

The implementation of the event driven approach is part of the DiSSCo core architecture. If the communication with GeoCASe could be based on the events, then no specific implementation for GeoCASe would be needed in DiSSCo. However, this approach would mean a significant change in the backend of GeoCASe. GeoCASe would need to develop new functionality to:

1. poll the DiSSCo event publisher for events;
2. interpret new events to determine the required operations;
3. pull data from the DiSSCo API; and
4. carry out the appropriate create, update, and delete operations in the GeoCASe database.

Not all of GeoCASe's contributing institutions will be part of the DiSSCo network, so GeoCASe would need to retain its existing data ingestion framework in addition to implementing this new approach for DiSSCo.

If it isn't viable for GeoCASe to implement the event driven approach, an alternative would be to set up a BioCASe Provider Software instance and GeoCASe-specific data storage within DiSSCo, so that GeoCASe could harvest data from DiSSCo through its existing mechanism. In this scenario, the burden of the work would be on DiSSCo and GeoCASe would not have the benefits of the event

[9] https://github.com/tdwg/abcd/issues/15

driven approach. Additionally, the current manual triggering of data harvesting from GeoCASe would potentially cause risks in the quality and timeliness of the data synchronisation.

Data standards

The BioCASe Provider Software pilot (described above) uncovered some issues with the ABCD compliance in the XML produced, including namespaces that deviate from the official TDWG namespaces. These issues will need to be addressed if the BioCASe Provider Software approach were to be pursued into a production environment.

Data model

To support minimum requirements for DiSSCo interoperability, GeoCASe would need to make some minor changes to its data model and schema. This would include support for storing PIDs for DiSSCo Digital Specimens and ROR IDs for organisations, to ensure that persistent, unambiguous links can be made between the data and their providers in the two infrastructures.

Beyond this, there is scope for GeoCASe to incorporate more data from the openDS schema to benefit its user base, such as MIDS levels and annotations. These opportunities are likely to become more clear as design and development of openDS, MIDS and other aspects of the ECOI schema (such as organisations, collections and transactions) continues to progress.

UX/UI and functionality

A number of UX/UI improvements have been identified in the course of the review and pilot activities. Those critical to the interoperability with DiSSCo, such as exposing DiSSCo PIDs on specimen pages, are relatively low effort. Others are more general suggestions for improvements that, while relevant to the uptake and success of the platform amongst the wider DiSSCo user community, are not so critical to the successful integration of the two platforms. These would be better contributed to the wider GeoCASe roadmap through engagement with the CETAF Earth Sciences Group.

Resourcing and sustainability

The current resourcing and funding situation for GeoCASe is precarious, with only 0.2 FTE of developer resource afforded by TalTech's current funding stream until the end of 2022, and no guaranteed resource for development, support and maintenance after that point. Without further investment, GeoCASe is likely to struggle to:

- progress with the planned technical roadmap for GeoCASe 2.0;
- provide the outreach and support required to increase the number of institutions contributing data to the platform;
- carry out the recommended work to enable integration with the DiSSCo;
- guarantee the continued stability and availability of the platform; and
- in the longer term, manage the build-up of technical debt in the platform.

Recommendations

Recommendations for DiSSCo Construct

1. In collaboration with TalTech, MfN, the GeoCASe Advisory Board and the CETAF Earth Science Group:

    - continue to develop the joint vision as expressed in this document for the development of GeoCASe, in alignment with DiSSCo requirements and focusing on adding value to DiSSCo in terms of visualisations and data validations;

    - explore models and opportunities for funding and resourcing GeoCASe development either through DiSSCo channels or external funding opportunities; and

    - assess the overlap with GBIF in the palaeontology domain, and consider whether there might be viable strategic approaches in either 1. focusing the GeoCASe scope on the mineral, rock and meteorite domains or 2. sourcing palaeontology data from a GBIF integration rather than directly from institutions.

2. Explore a phased approach to interoperability with GeoCASe.

    Phase 1a: Focus initially on publishing data to GeoCASe through the preparation and presentation of Darwin Core Archives that GeoCASe can ingest. This could offer several benefits:

    - allow GeoCASe to focus its limited resources initially on

        a. the minimum enhancements needed to incorporate and link with DiSSCo data, and

        b. further developing and scaling the DwC-A ingestion approach to reduce the technical barriers that BioCASe Provider Software implementation presents for prospective new data providers;

    - enable GeoCASe to onboard more data providers through early adopters of DiSSCo's mobilisation of data from institutional collections management systems.

    The proof of concept and increase in the number of institutions providing data to GeoCASe may also help to strengthen GeoCASe's case for external funding.

    Phase 1b: In parallel with phase 1a, work closely with GeoCASe on an early proof of concept for event-driven interoperability with an external service. Use the process to:

    - develop a blueprint and set of minimum requirements for external services to interoperate with DiSSCo using the event-driven architecture;

- develop a detailed specification, resource requirements and costings for GeoCASe development to automate ingestion from DiSSCo using the event driven approach.

Phase 2: Implement the changes specified in Phase 1b, if appropriate resources can be made available.

Recommendations for GeoCASe roadmap

In addition to the recommendations above, the research and pilot activities have identified some specific recommendations for the GeoCASe roadmap. These are:

1. Add the facility to store Digital Specimen identifiers against GeoCASe specimen records.
2. Incorporate Research Organization Registry[10] (ROR) identifiers as PIDs for organisations, to help to facilitate links with organisations in DiSSCo, GBIF, etc.
3. Improve standardisation of data against controlled vocabularies, including:
   - specimen type, and
   - country codes (conforming to ISO 3166[11]).

Mindat

Overview

Mindat[12] is an online database of mineralogical classifications, localities and occurrences. The platform was founded by Jolyon Ralph in 1993 as a personal database and DOS application in 1993. It was migrated to a Windows 95 app in 1995, and then first launched as the Mindat website in 2000. With a worldwide scope, significant volumes of data and large user base of professional and amateur mineralogists, geologists, and mineral collectors[13], Mindat claims to be the largest mineral database and mineralogical reference website on the internet[14].

More detailed information about Mindat can be found in **Appendix C**.

Requirements

The requirements below use a MUST/COULD/SHOULD prioritisation approach loosely based on the MoSCoW method[15].

[10] https://ror.org/
[11] https://www.iso.org/iso-3166-country-codes.html
[12] https://www.mindat.org
[13] https://en.wikipedia.org/wiki/Mindat
[14] https://mgs.geo.umass.edu/biblio/mindatorg
[15] https://en.wikipedia.org/wiki/MoSCoW_method

1. DiSSCo MUST include support for Mindat URIs in the openDS specification for mineral and rock specimens and their classifications

2. DiSSCo COULD develop an enrichment service to resolve classification strings against Mindat classifications and add links to Digital Specimens

3. DiSSCo COULD develop a service to verify classifications and localities in specimen records against known occurrences in Mindat

Interactions with DiSSCo

In many ways, the interactions with Mindat would be similar to those for the biological taxonomy lookup service described earlier in this report. Specimen data will be ingested by DiSSCo from the source system, and data administrators would elect to run an automated annotation service that will make a call to Mindat requesting the Mindat URL. This will attach an annotation to the specimen object, and if accepted the Mindat identifier will become part of the actual Digital Specimen. This could potentially be expanded to both geological classifications and localities.

Gap analysis against requirements

Data interoperability

The DiSSCo interactions described above are heavily dependent on access to Mindat data, preferably through an open and performant API, which Mindat doesn't yet offer. Development of an API for full data access is on Mindat's roadmap[16], but the timescales for delivery are unknown. There are also measures in place that prevent any systematic or automated extraction of information from the pages of mindat.org. The Mindat site does suggest that they are open to discussing automated access on a case-by-case basis.

Further information would also be needed on the adherence of Mindat identifiers to PID principles. The established method, used by GeoCASe, appears to be to use the webpage URLs that encode the identifiers (e.g. https://www.mindat.org/min-49971.html).

Architectural scalability

Mindat.org runs on a single server, and at present doesn't allow data downloads as the load on the server would be too much to deal with[17]. This suggests that there might need to be significant work and investment required to scale up the hardware and software architecture to a level where it could handle the kind of loads that might be put on it by automated annotation services from DiSSCo.

Licensing and copyright

[16] https://www.mindat.org/copyrights.php
[17] https://www.mindat.org/copyrights.php

The current licensing and copyright of Mindat content is quite complex, with the database being copyright of mindat.org, and different rights applying to some of the different components, in particular images. Mindat are working towards opening up the core data (excluding images) under a Creative Commons share-alike licence, but this is still more restrictive than DiSSCo's CC-BY default.

Resourcing and sustainability

Overhead costs for the maintenance of Mindat are supported almost entirely by public donations, and indications are that Jolyon Ralph is the sole developer of the platform. These factors are likely to preclude significant development work on the side of Mindat for DiSSCo integration without additional funds and resources being invested in the project.

Recommendations

Recommendations for DiSSCo Construct

1. Incorporate support for Mindat classification and locality identifiers into the openDS specification
2. Explore a potential relationship with Mindat for automated integration between DiSSCo and mindat.org
    a. Collaborate with Mindat to develop a specification and resourcing estimate for integration using DiSSCo's event-driven approach
    b. Investigate potential use cases and appetite for the integration or linkage of DiSSCo Digital Specimen data in the Mindat platform

Cost and resource estimates

DiSSCo

The DiSSCo developer team estimates that approximately 1 year's work would be required from a 1 FTE Back-End Developer for specific pieces of work relating to recommendations in this document. These include:

1. Developing data mappings and infrastructure to generate Darwin Core Archives from Digital Specimens, for ingest into GeoCASe: 2-3 months
2. Developing elements of the DiSSCo event publisher and API to expose events and data to GeoCASe and COL: 6 months
3. Developing an automated annotation service to parse and validate taxonomic data against ChecklistBank and annotate Digital Specimens with the results: 1.5 to 2 months

4. Developing an automated annotation service to extract geological classification data from mindat.org and annotate Digital Specimens with the results: 1.5 to 2 months

These figures should be considered extremely provisional at this stage, as the development of the DiSSCo architecture is at a very early stage.

These estimates also do not include the potential costs associated with DiSSCo developing its own complete set of taxonomic names services rather than working with COL, as discussed in the next section.

Species 2000

The Global Biodiversity Information Facility spends approximately 300K euro a year on taxonomic names services alone. This encompasses development work for building and maintaining the GBIF BackBone Taxonomy and all associated development work to properly represent mediated occurrence data through a taxonomic name-based discovery and access system for users. In addition, the costs also encompass helpdesk services, resolving taxonomic name problems, and mobilisation efforts and tools for taxonomic checklist data for the improvement of the GBIF Backbone Taxonomy (which is built on top of the COL Checklist). These costs are substantial for GBIF, and together with the desire to build a more open and collaborative system that supports and is supported by the taxonomic community, is one of the triggers to move towards the building of a common and shared infrastructure, ChecklistBank, together with Catalogue of Life and other key biodiversity data initiatives.

Species 2000 has a long history in providing taxonomic names services and is one of the oldest global biodiversity informatics initiatives still active. Given the current technical plan and scope of DiSSCo it is likely that DiSSCo may in the end have a cost for taxonomic names services that is similar to the costs spent by GBIF if DiSSCo were to develop all the services by itself. For DiSSCo, especially challenging would be the harmonisation of all the taxonomic lists coming from its member institutions. Although the volumes of data between GBIF and DiSSCo may not be comparable at the beginning of the construction of DiSSCo, the scope of the taxonomic issues will be comparable given that a lot of natural history collections in Europe cover large taxonomic sectors of the living tree, and the geographical scope of the collections covers the entire globe.

Species 2000 recommends that DiSSCo takes in its cost book a substantial costing for taxonomic names services into account, in line with what GBIF spends on a yearly basis.

Within ChecklistBank, COL will oversee and manage a custom taxonomic names services product for GBIF to use as a taxonomic backbone for the representation of mediated occurrences in GBIF.org and its associated APIs. This involves at least half of what GBIF already spends on taxonomic naming services on a yearly basis (150K euro a year). The same amount would be used as a standard membership fee for large biodiversity data infrastructures also in need of a taxonomic name services product, such as DiSSCo. In addition, project funds should cover any requirements that are specifically only required for a single initiative. Species 2000/COL will invite DiSSCo to become part of

the governance of ChecklistBank through a formal letter. A service level agreement or similar legal agreement between DiSSCo and COL could be signed to specify the taxonomic name service.

GeoCASe

Initial estimates of resource requirements for GeoCASe development related to the recommendations in this document include:

1. Minimum development to enable automated data ingest from DiSSCo and appropriate storage, linkage, and access: 6 months (Back-End Developer)
2. Full integration with DiSSCo through event-driven architecture: 12 months (Back-End Developer)
3. Additional data, UI and functionality enhancements in alignment with DiSSCo requirements: 12 months (Full-Stack Developer)
4. Ensuring ongoing support and maintenance of the platform, including server administration, and supporting existing and new data providers: 0.5 FTE Data Manager and 10 hours per month Server Administrator (ongoing requirement)

Mindat

A dialogue is still to be opened with Mindat about potential collaboration and integration with DiSSCo. At this point, it would be premature to attempt to make any estimates of the potential costs or resource implications on the part of Mindat.

References

Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Adlard, R., Adriaenssens, E. M., Aedo, C., Aescht, E., Akkari, N., Alexander, S., et al. (2022). Catalogue of Life Checklist (Version 2022-10-20). Catalogue of Life. https://doi.org/10.48580/dfqf

Costello, M. J., DeWalt, R. E., Orrell, T. M. & Banki, O. (2022). Two million species catalogued by 500 experts. Nature 601(7892): 191–191. https://doi.org/10.1038/d41586-022-00010-z.

Fitzgerald, H., Juslén, A., von Mering, S., Petersen, M., Raes, N., Islam, S., Berger, F, von Bonsdorff, T., Figueira, R., Haston, E., Häffner, E., Livermore, L., Runnel, V., De Smedt, S., Vincent, S., Weiland, C. (2021). DiSSCo Prepare Deliverable D1.1 Report on Life sciences use cases and user stories. https://doi.org/10.34960/xhxw-cb79

Glöckler, F., Pim Reis, J., von Mering, S., Petersen, M., Weiland, C., Dillen, M., Leeflang, S., Haston, E., Addink, W. & Fichtmüller, D. (2022). DiSSCo Prepare Deliverable D6.1 Harmonization and migration plan for the integration of CMSs into the coherent DiSSCo Research Infrastructure. DiSSCo Prepare. https://doi.org/10.34960/366d-sf49

Krimmel, E., Karim, T., Little, H., Walker, L. J., Burkhalter, R., Byrd, C., Millhouse, A. & Utrup, J. (2021). The Paleo Data Working Group: A model for developing and sustaining a community of practice. Biodiversity Information Science and Standards 5: e74370. https://doi.org/10.3897/biss.5.74370

Leeflang, S., Weiland, C., Grieb, J., Dillen, M., Islam, S., Fichtmueller, D., Addink, W., & Haston, E. (2022). DiSSCo Prepare Deliverable D6.2 Implementation and construction plan of the DiSSCo core architecture. DiSSCo Prepare. https://doi.org/10.34960/50B9-KJ05

Little, H., Byrd, C., Karim, T., Krimmel, E. & Norton, B. (2022). Extinct Taxa in an Extant World: Working towards better fossil taxonomic representation. Biodiversity Information Science and Standards 6: e94417. https://doi.org/10.3897/biss.6.94417

von Mering, S., Petersen, M., Fitzgerald, H., Juslén, A., Raes, N., Islam, S., Berger, F, von Bonsdorff, T., Figueira, R., Haston, E., Häffner, E., Livermore, L., Runnel, V., De Smedt, S., Vincent, S., Weiland, C. (2021). DiSSCo Prepare Deliverable D1.2 Report on Earth sciences use cases and user stories. DiSSCo Prepare. https://doi.org/10.34960/n3dk-ds60

Appendix A: Overview of Catalogue of Life and ChecklistBank

Catalogue of Life

Catalogue of Life (COL) is an international collaboration bringing together the effort and contributions of taxonomists and informaticians from around the world. COL aims to address the needs of researchers, policymakers, environmental managers and the wider public for a consistent and up-to-date listing of all the world's known species and their higher taxa. The COL Checklist is a consensus classification (Bánki et al. 2022), based on the underlying taxonomic source databases, managed by a community of more than 500 experts (Costello et al. 2022). The higher taxa are partially based on a management hierarchy. COL, through ChecklistBank, also supports those who need to manage their own taxonomic data and species lists.

In December 2020, the new COL infrastructure in collaboration with GBIF was launched. This infrastructure consists of three parts. First is a public portal[18] that facilitates access to the monthly updated COL Checklists, its underlying taxonomic databases, and general information on COL. The second component is ChecklistBank[19], which is a data repository that facilitates access to original data sources underlying the COL Checklist, all COL Checklist releases, all GBIF taxonomic checklists, all Plazi mediated data from published digital literature (over 45K datasets) and workbench or assembly tooling for the COL Checklist. ChecklistBank tools will be publicly available for future users to build taxonomic backbones with resources publicly held within it. Thirdly, the infrastructure includes a set of APIs[20] to render all COL Checklist data to ChecklistBank, the COL portal and users, provide persistent name and digital object identifiers (DOIs), and support various data standards (e.g. DwC-A, ColDP).

With the migration to the new COL infrastructure in December 2020, COL has also switched to a new algorithm to generate stable identifiers for name usages. The new implementation aims to keep the identifiers stable when the authorship of a name has only been slightly modified, although it does force a change in identifiers when an authorship is added to a record that previously lacked one. Changes in status (accepted name or synonym) and parent/classification changes do not trigger any ID changes. So, when name usages change status from an accepted name to a synonym or vice versa, there is no change in the ID. By combining a name usage identifier and the data set key the user has a stable reference to an immutable name usage in a particular release of the COL Checklist, no matter how the treatment of this name changes over time.

ChecklistBank

[18] https://catalogueoflife.org/
[19] https://www.checklistbank.org/
[20] https://api.checklistbank.org/

ChecklistBank is a high-functionality public repository and portal established to simplify FAIR data sharing for taxonomic and nomenclatural lists. It allows contributors to publish lists using a variety of typical data formats. Each list is then interpreted into a standard data model and accessible through a standard API and reusable web browser components. In future, datasets in ChecklistBank will be cited using a ChecklistBank Digital Object Identifier (DOIs). At present, DOIs are only available for project releases and its underlying data sources such as the COL Checklist. Data publishers benefit both by making their datasets accessible for reuse and attribution and also through ChecklistBank tools for data review and detection of possible issues. Some of the datasets in ChecklistBank serve as authoritative sources for sections of the COL Checklist, and new releases of the COL Checklist are also published as ChecklistBank datasets.

All datasets can be downloaded in multiple formats and accessed via a consistent API. Aggregating taxonomic and nomenclatural lists through a common portal makes it possible for users to locate sources offering differing perspectives on nomenclature and taxonomy.

ChecklistBank is provided as a fundamental tool to ensure that basic data on species names and classifications can be shared and reused in support of the biological sciences and wider societal uses.

Scope and functionality

Catalogue of Life produces the COL Checklist with the aim to deliver a comprehensive and up to date listing of all the world's known species. The COL Checklist is underpinned by 165 global species databases. A substantial number of these databases are mediated through partnering taxonomic initiatives, such as ITIS, WoRMS, TaxonWorks, etc.

ChecklistBank serves as a publishing platform for taxonomic and nomenclatural checklists. ChecklistBank can contain datasets that are not necessarily underpinning the COL Checklist. At present it contains more than 45K datasets. Apart from global species checklists, these contain published articles mediated through Plazi[21], national species lists, policy relevant list like invasive species lists or list from international policies (e.g. the European Environmental Agency), species lists originating from Natural History collections, species lists from citizen science observation networks, but also from genetic data sources such as NCBI taxonomy and UNITE.

Technical architecture

The COL public portal as well as the ChecklistBank API and UIs are hosted by the Global Biodiversity Information Facility (GBIF) in Copenhagen, Denmark. ChecklistBank is an open-source project with multiple repositories hosted in GitHub[22]. The back end[23] is implemented in Java as a Dropwizard application that drives the COL ChecklistBank API. The front-end[24] is a React user interface

[21] https://plazi.org/
[22] https://github.com/CatalogueOfLife
[23] https://github.com/CatalogueOfLife/backend
[24] https://github.com/CatalogueOfLife/checklistbank

application that uses the ChecklistBank API and supports public exploration of all data in ChecklistBank. It also includes (for appropriately authorised users) the tools for assembling taxonomic checklists from multiple sources.

GBIF and COL are developing a semi-automated part of the COL Checklist. The addition of a semi-automated part will make the COL Checklist more comprehensive by including extended information and enriched data for example coming from Plazi mediated species information and (M)OTU's coming from ENA and UNITE. It will improve taxonomic coverage and usefulness of the COL Checklist also in delivering taxonomic services for GBIF-mediated occurrences (**Figure 9**). The semi-automated part of the assembly of a checklist may in the future also become available as a generic function in the 'ChecklistBank project functionality'. For example, to build a specific backbone product geared towards the needs of DiSSCo, and to provide discovery and access services for scientific specimen objects based on scientific names.

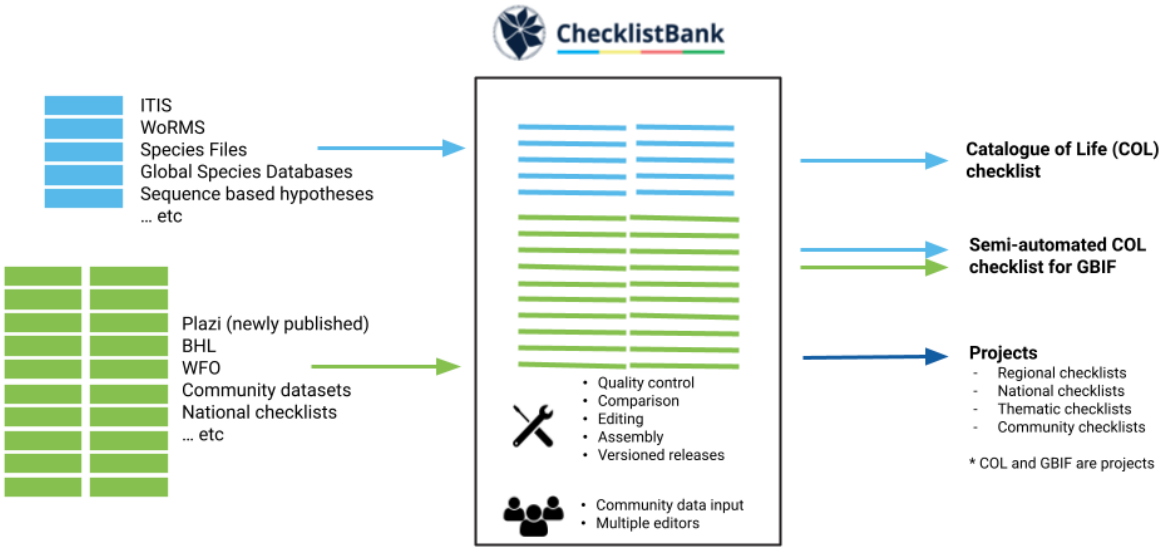

Figure 9: Schematic figure of ChecklistBank datasets and the products coming out of it.

At present datasets underpinning the COL Checklist (blue) and the GBIF Backbone Taxonomy (green) come into ChecklistBank. The COL Checklist is constructed from datasets in ChecklistBank. COL and GBIF now construct a semi-automated part of the COL Checklist that would serve as a candidate for the replacement of the GBIF Backbone Taxonomy. Data products, checklists of different scope like for instance specifically for DiSSCo, can also be generated through the 'ChecklistBank project functionality'.

Governance, funding and resourcing

Species 2000, originating in 1996, is one of the oldest biodiversity informatics initiatives in the world. Its aim has always been to serve as a federation of taxonomic databases. In the coming period,

Species 2000 will focus more on becoming a functional organisation for the representation of the taxonomic community in respect to species list building. Recently, a new organisation called the Catalogue of Life was set up in the Netherlands as a not-for-profit organisation. The Catalogue of Life organisation will at first focus on the oversight, management, and further development of the ChecklistBank infrastructure. Various biodiversity data initiatives that are dependent on the COL Checklist for taxonomic name services, will form part of the governance of ChecklistBank. This will also result in a new financing model for the Catalogue of Life. Up to now, Species 2000 / COL has been financed through a distributed consortium of institutes across the world. In the new situation, major users that are dependent on the COL Checklist services will help carry the financial sustainability.

DiSSCo is one of the biodiversity data initiatives that have been part of the steering committee of the development of the new COL infrastructure. It is envisioned that DiSSCo, like other sister organisations, will become part of the COL organisation to help maintain ChecklistBank as a global joint resource.

GBIF has put integration with COL for improvement of taxonomic services into their Global Work Programme and Strategic Plan since 2016. This work programme item is yearly resourced. The partnership between GBIF and COL around ChecklistBank is also part of the new GBIF Strategic Plan 2023 - 2027.

Data access, licensing and copyright

ChecklistBank and the COL Checklist support open data licenses, mostly Creative Commons CC-BY 4.0 and CC-0 licenses.

Current roadmap

The immediate roadmap for development for COL and GBIF will be to build and operationalise the semi-automated part of the COL Checklist. This semi-automated part will enrich the taxonomic community underpinned part of the COL Checklist, e.g. by adding new and already published taxonomic names, references, (M)OTUs to the COL Checklist. In addition, with generic list building tools the semi-automated part of the COL Checklist may also be a product that is of interest to DiSSCo as a means to address specific taxonomic names services for the discovery and access of scientific specimen collections.

Appendix B: Overview of GeoCASe

Scope and functionality

Search

The GeoCASe UI provides quick search, filter, and faceted search capabilities across the aggregated specimen data, with tabular, image gallery and map result views. The individual specimen pages include dataset and provider metadata in addition to the specimen-level information.

Data access

GeoCASe 2.0 includes a basic REST API, which returns specimen data using default Apache Solr search and responses. Data may also be downloaded directly via the main search UI in CSV and Excel format, but this is limited to a maximum of 1000 records at a time by the pagination of the result.

Data linkage

GeoCASe 2.0 includes links at the specimen and dataset level to several external platforms, including mindat.org[25], the Paleobiology Database (PBDB)[26], fossiilid.info[27] and the CETAF Registry of Collections[28].

Technical architecture

Platform technology

GeoCASe uses the Berlin Harvesting and Indexing Toolkit (BHIT)[29] based on the GBIF HIT. It is used to harvest data from different data providers that use the BioCASe Provider Software (BPS) to expose their collections data in the ABCD(EFG) data standard. The data is stored in a MariaDB[30] database and is indexed with the help of Apache Solr[31]. The GeoCASe User Interface and API connects to the indexed data to perform a scalable search experience. Both, the UI and API, has been developed as a NodeJS application. Images related to the data records remain on the data provider's server infrastructure, but thumbnails are created and cached by a microservice based on Imaginary[32].

The GeoCASe 2.0 infrastructure is hosted by the Tallinn University of Technology[33] (TalTech).

[25] https://www.mindat.org/
[26] https://paleobiodb.org/
[27] https://fossiilid.info/
[28] https://cetaf.org/registry-of-collections/
[29] https://wiki.bgbm.org/bhit
[30] https://mariadb.org
[31] https://solr.apache.org/
[32] https://github.com/h2non/imaginary
[33] https://taltech.ee/en/

Data ingestion

To contribute data to GeoCASe, providers have historically been required to install a local instance of the BioCASe Provider Software (BPS)[34], an open-source software package provided by the Botanic Garden and Botanical Museum Berlin (BGBM). BPS is used to extract the specimen data from one or more local databases or collections management systems (CMSs), map it to the ABCD+EFG data standard, and expose it to GeoCASe via a public endpoint. GeoCASe uses the Berlin Harvesting and Indexing Toolkit (B-HIT)[35] data collection tool to crawl BioCASe providers and harvest the data. This harvesting is currently triggered manually by GeoCASe, as fully automating the process would present issues with the current architecture.

A more recent development has been to add the ability to ingest data from Darwin Core Archives, which is intended to lower the technical barrier that the BPS approach has historically represented for many potential data providers.

Governance, funding and resourcing

GeoCASe is currently managed by the GeoCASe Advisory Board, who are responsible for the development, maintenance, promotion, and quality control of the portal. Further coordination, steering and stakeholder engagement is contributed by the CETAF (Consortium of European Taxonomic Facilities) Earth Sciences Group (ESG).

The initial version of GeoCASe was developed by the Museum für Naturkunde (MfN), Berlin in 2007 with funding from the SYNTHESYS[36] project. An upgraded version, GeoCASe 2.0, was developed by the MfN and Tallinn University of Technology, Estonia (TalTech), launching in 2021, and resulting in the transfer of primary responsibility for GeoCASe's development and maintenance from MfN to TalTech in 2021. TalTech's work has been funded through a national research infrastructure roadmap linked to the Estonian DiSSCo node, but that funding is due to finish at the end of 2022. At present, TalTech has an estimated 0.2 FTE of developer resources available to GeoCASe, which is also not guaranteed to persist beyond the end of 2022.

Licensing and copyright

GeoCASe recommends that data providers assign open licences to their data, but this is ultimately at the discretion of the data providers, and responsibility is with data users to comply with the licences and assume all rights are reserved if none is supplied. There is a field in the visible dataset that holds image licences, and from a visual scan there appears to be a mixture of Creative Commons Attribution (CC-BY), Non-Commercial (CC-BY-NC), and Share-Alike (CC-BY-SA) licences, and some instances of the Creative Commons open waiver (CC0). There doesn't however seem to be a field that reflects the licence for the data visible at the record or the dataset level.

[34] https://www.biocase.org/products/provider_software/
[35] https://wiki.bgbm.org/bhit/index.php/Main_Page
[36] https://www.synthesys.info/

The GeoCASe code[37,38,39] is openly available in GitHub under the GPL 3.0 Licence.

[37] https://github.com/geocollections/geocase-ui
[38] https://github.com/geocollections/geocase-infrastructure
[39] https://github.com/geocollections/geocase-thumbnail

Appendix C: Overview of Mindat

Scope and functionality

Mindat[40] is an online database of mineralogical classifications, localities, and occurrences. The platform was founded by Jolyon Ralph in 1993 as a personal database and DOS application in 1993. It was migrated to a Windows 95 app in 1995, and then first launched as the Mindat website in 2000. With a worldwide scope, significant volumes of data (**Table 3**) and large user base of professional and amateur mineralogists, geologists, and mineral collectors[41], Mindat claims to be the largest mineral database and mineralogical reference website on the internet[42].

Table 3: Mindat resource statistics[43] as at 2022-10-19

Mineral species	5,835
Rock names	3,054
Other names	45,332
Localities	385,694
Occurrences	1,436,078
Photos	1,229,152
Articles	3,116
Glossary items	25,986
Registered users	66,392

Figure 10 below provides an interpretation of some of the data scope and relationships within mindat.org, based on a manual review and analysis of the website pages and interlinks.

[40] https://www.mindat.org
[41] https://en.wikipedia.org/wiki/Mindat
[42] https://mgs.geo.umass.edu/biblio/mindatorg
[43] https://www.mindat.org

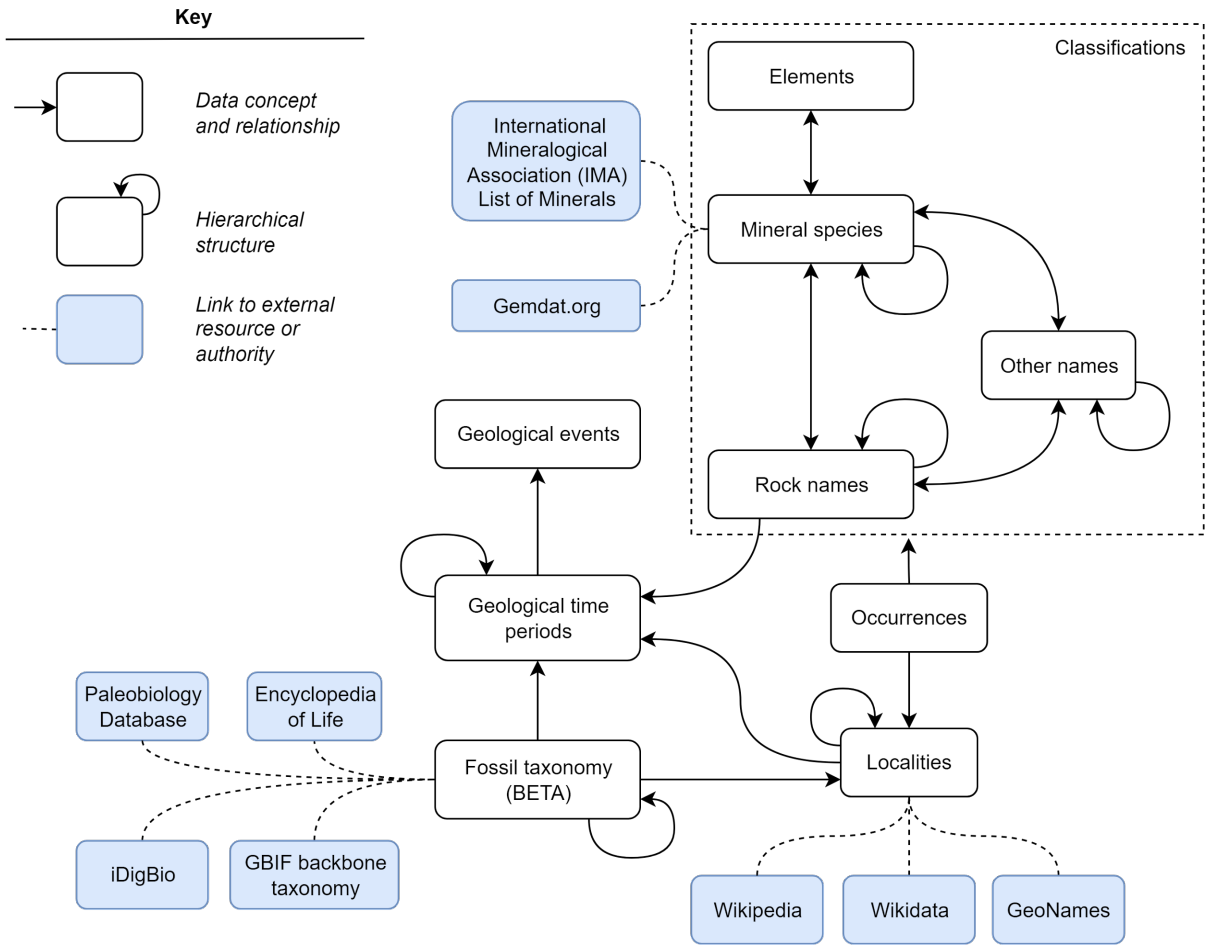

Figure 10: A high-level interpretation of the main data concepts in Mindat, with links to external services and authorities. This depiction was constructed from a manual review of web pages and links in mindat.org and may not be fully accurate and comprehensive.

Localities

Mindat includes an extensive dataset of mineral localities, arranged hierarchically from region/country level down through political subdivisions to deposits and mines. Locality records include links to relevant records in Wikipedia, Wikidata and GeoNames.

Classifications

Mineralogical classifications in Mindat include the list of official International Mineralogical Association (IMA) Approved Mineral Species[44], and a large dataset of alternative names. The latter includes a hierarchy of rock names[45], mineral varieties, mixtures, and synonyms, as well as the

[44] https://www.mindat.org/minerals.php
[45] https://www.mindat.org/min-50468.html

constituent chemical elements. Classification records are extensively linked to represent a variety of interrelationships between concepts.

Occurrences

Occurrences in Mindat represent an established link between a classification and a locality, used to aggregate the lists of recorded commodities, minerals and rock types for a given locality and vice versa. In addition to the link, occurrences also include basic metadata on the validity of the assertion and a reference.

Geochronology and geological events

A chronostratigraphic hierarchy, representing the Geologic Time Scale (GTS) and aligned predominantly with the International Commission on Stratigraphy (ICS) Stratigraphic Chart, is included in the data structure. This also has an associated dataset of geological events related to the geological time periods. The stratigraphic data is linked with the rock name classifications, and with relevant lists and counts of taxa derived from data imported from the Paleobiology Database as part of the taxonomic functionality in development (see below).

Taxonomy (beta / in development)

The taxonomy pages in mindat.org are publicly available but are currently flagged as being experimental and under development. The data are sourced from or linked to external platforms including the Paleobiology Database (PBDB), the GBIF taxonomic backbone, iDigBio, the Encyclopedia of Life and Wikipedia.

Technical architecture

Mindat.org runs on a single server with mirrored storage, and backup server in a different location with nightly synchronisation. Investigation of the mindat.org site suggests that the technology is based around a LAMP (Linux/Apache/MySQL/PHP) architecture.

Governance, funding and resourcing

Mindat is run by the Hudson Institute of Mineralogy[46], a not-for-profit research, cultural and educational entity chartered in 2003 and approved in 2004. Overhead costs for the maintenance of Mindat are supported almost entirely by public donations.

The site is supported by a management team of around 50 experts, who review submitted content and oversee editorial policies. Beyond this, there are around 450 expert contributors around the world who enjoy a higher trust level in submitting and editing content, and around 4000 regular contributors.

[46] http://www.hudsonmineralogy.org/

Data access, licensing and copyright

Mindat doesn't currently offer data download functionality through the standard UI[47]. The current server architecture isn't sufficiently powerful to support bulk downloads, and there are also considerations around copyright and appropriate use of the data. They are, however, working towards the development of an API for full data access and a more open licensing approach (see below).

The current copyright status of Mindat content is quite complex, as the database as a whole is copyright of Mindat under database copyright law but contains data elements considered as scientific facts (not copyrightable) and also content, images etc that remain copyright of the individual contributors. However, they are working towards opening up their core data (excluding content subject to third party copyright) under a Creative Commons Share-Alike (CC-BY-SA) licence.

The Mindat program code is copyright of Jolyon Ralph, apart from some portions (e.g. the message boards) that use third party open source software.

[47] https://www.mindat.org/copyrights.php