Research Ideas and Outcomes 5: e32987
doi: 10.3897/rio.5.e32987

Workshop Report

Access to Geosciences - Ways and Means to share and publish collection data

Mareike Petersen‡, Jana Hoffmann‡, Falko Glöckler‡

‡ Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

Corresponding author: Mareike Petersen (mareike.petersen@mfn.berlin)

Reviewable v1

Received: 10 Jan 2019 | Published: 11 Jan 2019

Citation: Petersen M, Hoffmann J, Glöckler F (2019) Access to Geosciences – Ways and Means to share and publish collection data. Research Ideas and Outcomes 5: e32987. https://doi.org/10.3897/rio.5.e32987

Abstract

Natural history collections are invaluable tools for various questions regarding biodiversity, environmental, and cultural studies. All object metadata thus need to be findable, reachable and interoperable for the scientific community and beyond. This requires a good structuration of data, appropriate exchange formats, and web sites or portals making all necessary information accessible. Collection managers, curators, and scientist from various institutions and nationalities were surveyed in order to understand the importance of open geoscientific collections for the respective holding institution and their daily work. In addition, particular requirements for the publication of geoscientific collection object metadata were gathered in a two-day workshop with international experts working with paleontological, mineralogical, petrological and meteorite collections. The survey and workshop revealed that common data standards are of crucial importance though insufficiently used by most institutions. The extent and type of information necessary for the publication and discussed during the workshop will be considered for domain specific application schema facilitating the publication and exchange of geoscientific object metadata. There is a high demand for comprehensive data portals covering all geoscientific disciplines. Gathered portal requirements will be taken into account when improving the already running GeoCASe aggregator platform.

Keywords

ABCDEFG, data standard, data portal, GeoCASe, natural history collections, minerals, rocks, fossils, meteorites

Introduction

Our natural history heritage is distributed worldwide. The collection objects are stored at various organizations and information about the specimens come, in most cases, from independently developed databases. It is a great challenge to make these heterogeneous data sources interoperable and to unite them in data aggregators and portals accessible for both the scientific community and the broader public. Unique and domain specific data standards with precisely defined elements are required, in order to allow for an exchange and a standardized publication of collection object related data (hereafter referred to as metadata) with appropriate granularity. Following a common standard schema, data from various institutions can be integrated, displayed and accessed via data portals in a sophisticated manner.

The data standard Access to Biological Collection Data (ABCD, Berendsohn 2007) is a well-known and widely accepted exchange format primarily used for natural history collection and observation data. In 2005, ABCD was ratified as a standard by the nonprofit scientific and educational association Biodiversity Information Standards (TDWG), formerly known as the Taxonomic Databases Working Group (www.tdwg.org). In order to specify data related to geoscientific objects, an ABCD extension was developed covering petrology, mineralogy, and paleontology. During two subsequent workshops, demands of experts representing the relevant disciplines were collected and formed the basis for the extension EFG (Extension for Geoscience, Kiessling et al. 2006). The EFG schema is currently being reviewed by TDWG for ratification. Since its development, the standard ABCD and its extension EFG has already been actively used and has been accepted for data publication in miscellaneous portals (summary in Petersen et al. 2018).

Within the scope of the research and service project “ABCD 3.0 – A community platform for the development and documentation of the ABCD standard for natural history collections”*1 (https://abcd.biowikifarm.net/) ABCDEFG was imported into the TDWG Terms Wiki (https://terms.tdwg.org/wiki/ABCD_EFG), a collaborative developmental platform for the definition, curation, translation, annotation and discussion of basic concepts and terms in biodiversity terminology. As part of the project, workshops with representatives of different scientific communities were carried out to:

1. review parts of the schema,

2. extend the schema where necessary, and

3. draw up application schema for specific use cases.

In addition to mandatory elements and elements of general importance, an application schema comprises concepts relevant for specific purposes; i.e. discipline, collection, or for publication in a particular data portal. An application schema is thus a defined subset of concepts available in the entire standard and may be completed with concepts derived from other standards where necessary.

This paper summarizes and discusses the results of a survey carried out among scientists and collection managers on the general importance (scientific and public scope) and accessibility of their institutional geoscientific collections. In addition, the results of a two-day workshop focusing on collection object metadata publication in geosciences are presented. Scientists, curators, and data base developers shared their experience using data standards in geoscientific collections. Elements being essential for an application schema in the disciplines of paleontology, petrology, mineralogy, and meteorite research were compiled and are presented herein. This paper concludes with a discussion of all results and outlines the next steps toward a better access to collection object metadata in Geosciences.

Survey on Geoscientific Collection Data

Background

Among biological disciplines, e.g., the Global Biodiversity Information Facility (GBIF, https://www.gbif.org) provides the most comprehensive and prominent search portal for species occurrences including collection object and observation data. As GBIF focusses on biodiversity only, the presentation of fossils is inadequate and data describing abiotic specimens like rocks and minerals are completely missing. In order to overcome this, the Geoscientific Collection Access Service (GeoCASe, http://www.geocase.eu) a domain specific portal aggregating data using the ABCDEFG standard, was developed. Here, geoscientific objects can be published and researched in a convenient way.

The maintenance of the portal, adding new features as well as further developments for adapting to the rapid changes in web technologies are ongoing challenges. In order to increase the acceptance of the portal and provide a sustainable service, it is necessary to focus future efforts primarily on demands arising from the scientific community itself.

For collecting user demands, specialists from the geosciences were asked to participate in an online survey about geoscientific collection data, data publication, and cross-institutional data search (see Suppl. material 1 for the survey template and Suppl. material 2 for all given answers). If not specified, we always give the frequency as the absolute amount of answers given to a particular question.

Results

Twenty-five scientists, of which 48% were curators, completed the questionnaire. They are associated with 22 institutions located in seven different countries. The institutions hold a

variety of geoscientific collection objects with minerals and rocks being the most common specimens (Fig. 1).

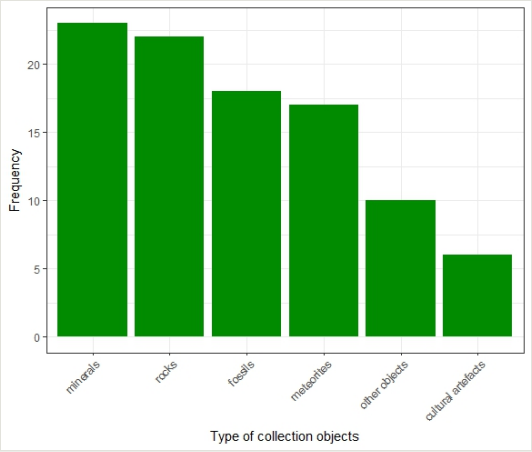

Figure 1. doi

Types of collection objects. Given are the frequencies of particular objects the survey participants' institution or department holds. Other objects include soil, crystal models, historical instruments, boreholes, etc.

Around 50% of the participants are making collection object metadata available. The data are accessible via international data portals such as GeoCASe, GBIF, and Europeana (https://www.europeana.eu/portal); national data portals like the German museum-digital (https://www.museum-digital.de), or their own institutional websites. The usage of data standards varies from 'no standard', to 'agreements within the department or institutions', to using international data standards like Darwin Core or ABCDEFG. However, one third of the participants are not aware of existing conventions for the publication of object metadata (answered 'none' or left question blank).

The perceivedvalue of data portals for cross-institutional search varies among potential audiences. The participants name primarily the scientific community and secondarily the institution as the groups most benefiting from existing data portals. For daily individual work the importance of portals differs among the participants between high and low. Based on the survey, published geoscientific object data is of minor importance for the general public. See Fig. 2 for further details.

Thereasons for publishing or not publising data on geoscientific collection objects through a data portal were also asked about. According to the responses received, higher accessibility and higher visibility are the main reasons for publishing through a data portal (Fig. 3a). Other reasons like 'special policies', 'nice to have', or 'because others do' are only of minor importance. On the other hand, concerns regarding security, bad data quality and potential misapplication of the data make researchers and curators refrain from publishing particular data sets (Fig. 3b).

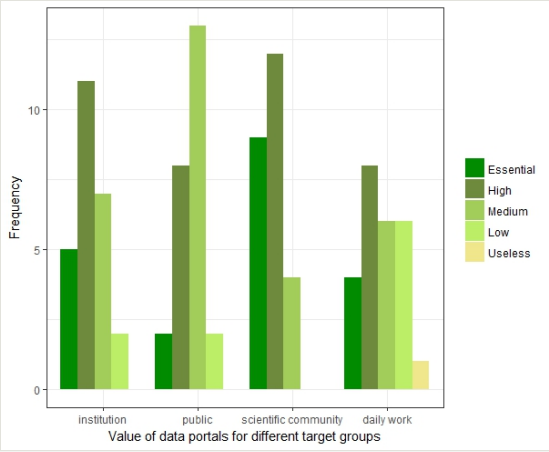

Figure 2. doi

Estimated value of cross-institutional search possibilities through data portals on geoscientific object (meta)data for different target groups. Shown is the frequency of answers for different scopes and their respective importance (see legend).

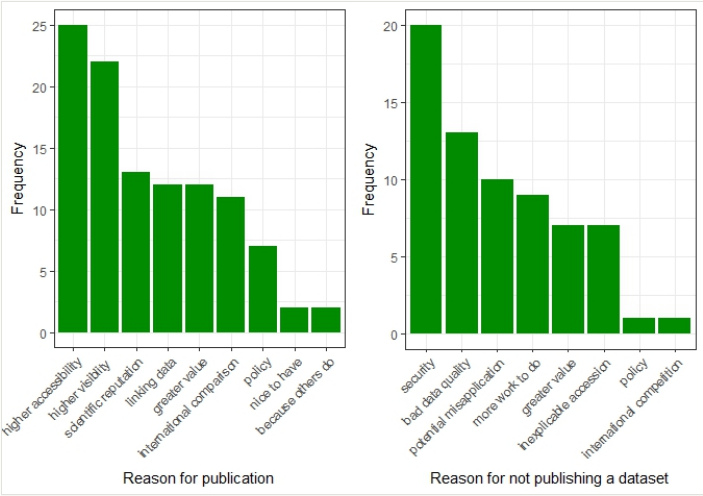

Figure 3. doi

Reasons for publication of (meta)data on geoscientific collection objects (a) and not publishing particular datasets through a data portal (b). Shown is the frequency of answers per reason. Note that the y-axes are scaled differently.

Another question focused on the minimum functionality that a data portal needs to cover and which optional features it could provide in order to increase its value for daily scientific work. Here, multiple differences between each geoscientific sub discipline emerge and several participants emphasize that the minimum information needed depends on the type of collection and scientific task performed.

For paleontological objects, the following properties were mentioned: scientific name / taxonomy, common name, type status, details of the specimen (which part of the skeleton), locality (i.e. country), age, geochronology, lithostratigraphy, quantity, etc.

For rocks and minerals, participants listed common name, synonyms or historical names, chemical composition, locality (including details such as mine, administrative unit, or country), geological horizon/age/locality, color, year entered to the collection (historical purposes), where the rock was used (architectural requests), accession (collector, year of collection), including thesauri such as of mindat (https://www.mindat.org), etc.

For meteorites, the name is unique and all information on a specific meteorite is linked to its name. However, the unique reference number, parent-daughter / sibling relationships, mass and storage in the collection are considered as required fields for a data portal.

In general, the inventory number, the housing institution, and a contact person responsible were identified being essential metadata. If available, a photograph of the object would also increase the value. The most desirable minimum information is the name and collection locality as this will enable most searches carried out by external researchers and the public. The functionality of a respective portal should be similar to that of GBIF. Further demands are a basic search function for all fields (e.g. thematic and regional searches), fuzzy searches, queries for classification, descriptive statistics, and an integrated loan service*2.

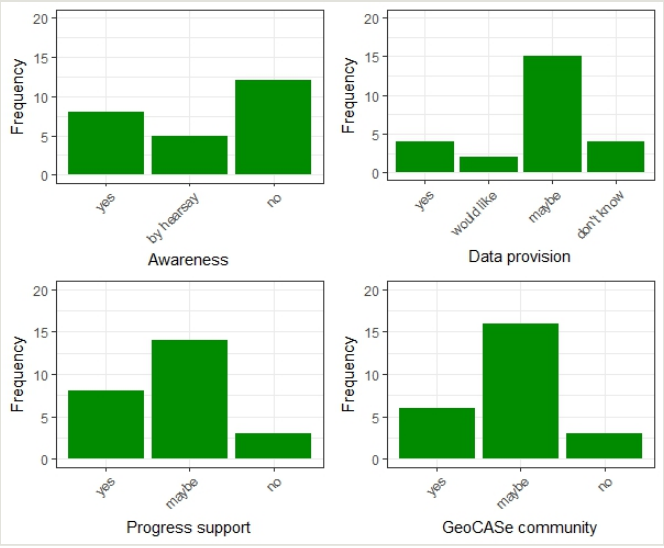

Figure 4. doi

Awareness and usage of Geoscientific Collection Access Service (GeoCASe). Shown is the frequency of answers regarding the awareness of GeoCASe (a), the data provision through the GeoCASe portal (b), the readiness to support the progress of improvements (c), and the interest of being part of the GeoCASe community (d).

The last part of the questionnaire focused on the awareness and usage of the Geoscientific Collection Access Service (GeoCASe). GeoCASe is known by around fifty percent of all participants (Fig. 4a), whereas only four institutions were already providing data to the portal. Most participants would like to use the service but require more information on GeoCASe(Fig. 4b). The majority of participants is willing to support the improvement of the portal (Fig. 4c) and would like be part of the GeoCASe community (Fig. 4d).

Workshop: "Access to Geosciences: sharing and publishing data related to paleontological, mineralogical, and petrological objects using a common data standard"

Summary

The Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science hosted an international workshop focusing on data standards in geosciences on May 29th and 30th 2017. During the workshop geoscientists, curators of geoscience collections, and data base developers shared their experiences using data exchange formats in geosciences. The workshop participants were selected based on their scientific affiliation to paleontology (first day) or petrology, mineralogy and meteorites (second day). Specialists with overlapping expertise were welcome to participate on both days.

Both days were structured similarly. Starting with a short personal introduction of all participants, presentations focusing on recent developments and the digital status of important geoscientific collections from institutes worldwide followed (see https://abcd.biowikifarm.net/wiki/Events:WorkshopEFG2017 and Suppl. material 3). The subsequent session delved into the data standard ABCDEFG with a presentation describing the workflow of publishing an institutional collection in the GeoCASe portal using EFG in particular, and highlighted current developments of this common biodiversity data standard. In the last part of the workshop, the additional needs and requests relating to a data exchange format based on use cases was collected from all participants in an open discussion. Furthermore, fundamental elements being vital for the publication of collection data related to different geoscientific disciplines in general (paleontology, mineralogy, petrology, and meteorites) were identified.

Results and Discussion

Following the thematic talks on the first day, input of all workshop participants regarding the development of a data portal and common data standard in geosciences was collected in an open discussion session.

Collection database, data portal and research data

It became clear that the perception of researchers, curators as well as collection managers regarding a database or data portal differ. Whereas curators and collection managers are

seeking for an appropriate infrastructure for collection object related information (inventory database), researchers mostly use databases with collection object related information derived from already published literature for their daily work (reference database). Thus, to better facilitate collection object-based research, critical collection information must match the most important scientific criteria such as age model, stage geography, stratigraphy, and taxonomy.

The linkage of research data with the respective collection database is mostly favoured but always a challenge. However, there are different approaches to realize this added value, for example by a reference connected to the specimen (e.g. type specimen database Stuttgart; http://www.dbsmns.naturkundemuseum-bw.de) or searchable references through a DOI (Digital Object Identifier) associated to collection objects in data portals like GeoCASe.

A common database, covering the requirements of researchers and collection managers is possible and should be our common goal. Such a database would merge and store information derived from collection objects (research data) together with elementary metadata closely related to the object itself (collection data).

Locality and geo-data

During the workshop we spent a considerable amount of time discussing details of the requirements and challenges regarding the sampling locality of an object. In this context, participants remarked that future developments or adjustments of data standards should always follow the guidelines of the European spatial data infrastructure INSPIRE (http://inspire.ec.europa.eu) for associated geo spatial data. Historical collection objects are often accompanied with missing (e.g. only region or country is given), defective (e.g. wrong location name or GPS coordinates), or inconsistent (e.g. projection, scales) collection site information. In general, time and manpower for correct georeferencing and data cleaning are lacking. New vouchers collected more recently, however, mostly have better information on their collection site. Although geospatial information are essential (meta) data, it should also be possible to hide details on the collection site in order to protect sensitive collection sites.

Information relevant for an application schema

In the last part of the first workshop day, collection object metadata of interest for paleontological researchers were collected. (Table 1) lists all of the collected information. It is furthermore stated whether they were classified as essential or optional for the publication of collection object related data. These elements are suggested to be part of the application schema for paleontological collection objects.

Building upon the results obtained from day one, participants in the second workshop day were divided into three thematic groups by expertise: rocks, minerals, and meteorites. Within these groups researchers and collection managers were asked to identify essential elements for the publication of metadata for each respective collection. Here, not only

researchers but also other potential user groups (e.g. artists, creatives, industry, educational sector) of the data and their requirements were considered.

Table 1.

Important information relevant for paleontological application schemata. Given are concepts essential or optional for the publication of collection object associated data as well as additional comments or action items (AI) regarding the respective concept.

Essential	Optional	Comment /AI
Unit		
taxon name		special case: micropaleontology; several taxa in one sample
collection and/or field ID		
holding institution		
	analytic results	
Gathering		
location (preferred GPS coordinates in WGS 84)		check INSPIRE guidelines for additional requriements
	site description / particular areas (landscape, coal mine)	the use of the ABCD elements AreaClass ABCD Schema Task Group 2005band AreaName (ABCD Schema Task Group 2005) is preferred compared to the EFG element NamedGeologicalFeature (ABCDEFG Development Team 2005
	collecting person	
	expedition / collection	
	historic researcher	
Stratigraphy		
most recent stratigraphy		
	historical stratigraphy	check for common standard, mandatory fields, and controlled vocabulary
	geochronology / chronostratigraphy	age models could be of interest
	lithostratigraphy	check for „national“ scale; continental mostly done
Unit associated material		
	Image (imaging method; technical metadata, camera, SEM, 3D standard, Micro CT…)	
	reference (DOI if available)	

The following terms, particularly important for each respective domain, were listed during the following discussion.

Rocks: time plus alteration event (incl. metamorphs plus time steps), time plus geography, inclusions, usage of rocks (place / time), preparation (e.g. for artists), color (e.g. color chart plus reference), rock texture (existing terms), rock mineralization, rock structure (including relationships of rocks), deformation, classification, classification history, link to specific dates or geological events (e.g. volcano eruptions), holding collection.

Minerals: inclusions, horizons, mineral structure, analytic results (e.g. chemical composition), link to chemical authority databases (MINDAT, https://mindatd.org/), secondary minerals, solid solutions, relationships between minerals (age relation).

Meteorites: reference to the database of the Meteoritical Society (https://www.lpi.usra.edu/meteor/metbull.php), geological event (fall), observations (observed fall, pseudo-observed fall, not observed), inventory number, size, weight, preparation, local storage, link to owner institution / collection, attached material (video, sound).

Conclusion and Outlook

The workshop revealed not only differences between geoscientific disciplines, but also between collections and institutions with respect to publication of collection object associated information. Although data standards were known by most participants, the majority of them neither publish data using common exchange formats nor do they publish in domain specific portals. Some institutes, however, make images of their collection objects and / or a subset of associated metadata available on their website.

Researchers from all disciplines shared the opinion that rocks and fossils are strongly related with each other and collections objects composed of both should always be described together. Therefore, a data standard should provide the option to refer to related elements e.g. the host rock of a fossil or the mineral structures associated with a fossil. This is of course also true for objects composed of several rock types or minerals.

In contrast to other geoscience collection objects and natural history objects in general, meteorites have a unique name and all information (data, locality, observation etc.) is linked to this particular name. All data is freely available in the reference database of the Meteoritical Society (https://www.lpi.usra.edu/meteor/metbull.php). An institutional inventory number, weight, size, local collection storage, and associated multimedia objects, are however important properties for institutional and research purposes as well as the broader public and thus should be available in any exchange format or data standard used.

In addition, the importance of publically available information also differs among user groups. Scientists ask for different facts compared to the broader public or the creative industry. The color of minerals for example, is only of minor importance for geoscientists but of high importance for an artist planning an exhibition. Given the heterogeneous group of participants of the workshop, the requirements of various potential user groups of

geoscience object related data could be collected. Scientists, collection managers, and data base developers exchanged their experience. In addition to the demand for publically available data, requests of past users of their institutional collections were also taken into account. The elements relevant for the publication of paleontological object metadata (Table 1) and the collection of important metadata of rocks, minerals, and meteorites (see above) will be incorporated in the application schemata for the different disciplines.

During the workshop, the participants were asked to identify object metadata which need to be publically available. A couple of the terms identified for one subject are of general importance for all natural history collection objects and will of course be part of all application schemata.

Discussion & Conclusion

The workshop and survey revealed that data standards are known by most of the interviewed geoscientists and curators, though some know only little about data standards or misunderstand their principles. Despite the fact that most institutions do have internal conventions, they are not publishing their data using common standards. The exceptionally diverse composition of the workshop, with international experts covering most geoscientific disciplines, was a unique possibility to collect and explicitly discuss the requirements for a common data standard. Together with some of the survey results regarding the minimum expected functionality of a geoscientific data portal, collection object metadata elements, which need to be publically available and thus are indispensable in a domain-specific data exchange format, were identified. (Fig. 5) summarizes the terms for each geoscientific discipline separately and illustrates overlapping information in the center.

There is a need to make collection object associated data publicly available for scientific purposes as well as for the broader public. Although this can partly be realized using institutional websites, a data portal is preferred. A portal reproduces the content of each collection and collection items of different collections can be explored together, offering a higher amount but also higher variety of objects from one domain. In addition, features such as galleries with associated multimedia objects and an integrated loan service could be implemented. Although GeoCASe covers most demands of the workshop and survey participants it appears not to be sufficiently known in the community. Assuming successful funding, we will improve and extend GeoCASe towards the required features and services, but also focus on drawing attention tothe importance and assets this portal provides for the scientific community andthe general public.

As GeoCASe is an aggregating platform for data of different institutions, the quality of the search results greatly depends on the data quality of each provider. Assuming a consistent quality from each individual provider, the main quality issue is the compilation of different data sources. Thus, it was identified that in addition to a harmonized data structure provided by the ABCDEFG standard, a harmonized vocabulary for the content (i.e. gazetteers for stratigraphy and geography) is a desirable goal. This can be achieved with both, the existing technology in order to check for data consistency e.g. by using the

BioCASe Monitor Service (Glöckler et al. 2013), and by adding mandatory vocabularies and recommended references to existing authority files to the next version of the ABCDEFG standard.

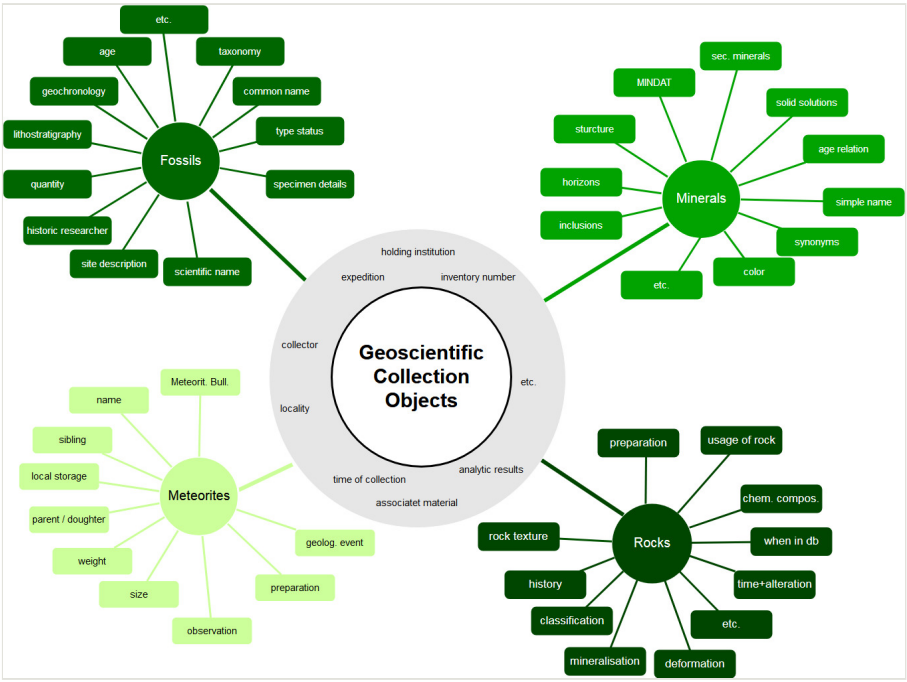

Figure 5. doi

Elements identified by workshop and survey participants as being important for the publication of geoscientific collection objects. The four different object classes (fossils, rocks, meteorites, and minerals) are highlighted in different shades of green, terms important for the respective class are assigned around the object name and likewise colored, terms in the center (grey) are relevant for all four object classes. Note: The figure only summarizes the mentioned terms, there might be more properties that are important for the publication of collection object associated data for the different object classes and / or more terms mentioned for one class only but important for several.

The findings of the workshop, the survey results and our experience in using ABCDEFG (Holetschek 2015, Holetschek 2016, Petersen et al. 2018) provide the basis for a domain-specific application schema. As a next step we will condense all information and prepare an application schema for each geoscientific collection type and / or use case identified. This compilation of terms will serve as a general guideline for the publication of geoscientific collection object associated data. In addition to a descriptive and illustrative version, we will compile functional XML schemata. These schemata can be used to exchange and publish data (e.g. GBIF or GeoCASe) with the free and open-source BioCASe Provider Software (BPS; http://www.biocase.org/products/provider_software/).

Acknowledgements

The workshop was supported by ABCD 3.0, a DFG project funded under the LIS infrastructure platform, and by Geo.X, the Research network for Geosciences in Berlin and Potsdam. We thank L. Hecht, D. Fichtmüller, B. Baltruschat, and L. Lertsutham for their help preparing the content and conducting the workshop and all participants for their valuable contribution. An anonymous reviewer significantly improved the manuscript by editing the language and proofreading.

Funding program

Deutsche Forschungsgemeinschaft (DFG). Programme: Scientific Library Services and Information Systems (LIS)

References

- ABCDEFG Development Team (2005) NamedGeologicalFeature. https://terms.tdwg.org/wiki/abcd-efg:NamedGeologicalFeature. Accessed on: 2018-6-26.
- ABCD Schema Task Group (2005a) AreaName. https://terms.tdwg.org/wiki/abcd2:Gathering-NamedArea-AreaName. Accessed on: 2018-6-26.
- ABCD Schema Task Group (2005b) AreaClass. https://terms.tdwg.org/wiki/abcd2:Gathering-NamedArea-AreaClass. Accessed on: 2018-6-26.
- Berendsohn WG (2007) Access to biological collection data. ABCD Schema 2.06 – Ratified TDWG Standard. URL: http://www.bgbm.org/TDWG/CODATA/Schema/default.htm
- Glöckler F, Hoffmann J, Theeten F (2013) The BioCASe Monitor Service - A tool for monitoring progress and quality of data provision through distributed data networks. Biodiversity Data Journal 1: e968. https://doi.org/10.3897/bdj.1.e968
- Holetschek J (2015) BioCASE Concept survey, Biological CollectionAccess Service for Europe. http://www.biocase. org/whats_biocase/concept_survey.cgi. Accessed on: 2018-7-09.
- Holetschek J (2016) Commonly used ABCD 2.06 concepts, Documentation on Wiki of the BioCASe Provider Software. http://wiki.bgbm.org/bps/index.php/CommonABCD2Concepts. Accessed on: 2018-7-09.
- Kiessling W, Copp C, Risonné A, Döring M, Mewis H (2006) The EFG extension to the ABCD schema. In: Belbin L, Rissoné A, Weitzmann A (Eds) Proceedings of TDWG: Abstracts of the 2006 Annual Conference of Biodiversity Information Standards (TDWG). St. Louis, USA, 15–22 October 2006.
- Petersen M, Glöckler F, Kiessling W, Döring M, Fichtmüller D, Laphakorn L, Baltruschat B, Hoffmann J (2018) History and development of ABCDEFG: a data standard for geosciences. Fossil Record 21 (1): 47-53. https://doi.org/10.5194/fr-21-47-2018

Supplementary materials

Suppl. material 1: Survey on Geoscientific Collection Data doi

Authors: Glöckler et. al.
Data type: survey
Brief description: Sheet on the survey on geoscientific collection data
Filename: Survey_sheet_neu.pdf - Download file (231.36 kb)

Suppl. material 2: Results of the survey on geoscientific collection data doi

Authors: Glöckler et. al.
Data type: table
Brief description: Results of the survey on geoscientific collection data. Names, institutions and contact details are anonymized
Filename: survey_on_geoscientific_collection_data_anonym.xlsx - Download file (19.71 kb)

Suppl. material 3: Workshop Programm doi

Authors: Petersen et. al.
Data type: table
Brief description: Programm of the two-day workshop Access to Geosciences: sharing and publishing data related to paleontological, mineralogical, and petrological objects using a common data standard. Museum für Naturkunde, 29-30 May 2017
Filename: Suppl_Workshop Program_without_names.pdf - Download file (75.09 kb)

Endnotes

*1 ABCD 3.0: Funded by the German Research Foundation (Deutsche Forschungsgemeinschaft), Scientific Library Services and Information Systems; partners: Museum für Naturkunde Berlin (MfN) and Botanical Garden and Botanical Museum Berlin Dahlem (BGBM).

*2 Object loan service: This service could cover various functionalities, including information about the availability of an object for loan, lending restrictions, communication between collection managers and requesting scientists, reminder function for deadlines etc.