

**A framework to support the annotation,
discovery, and evaluation of data in ecology, for
better visibility and reuse of data and an
increased societal value gained from
environmental projects**

DISSERTATION

Dr. rer. nat.

Claas-Thido Pfaff

**A framework to support the annotation,
discovery, and evaluation of data in ecology,
for better visibility and reuse of data and an
increased societal value gained from
environmental projects**

Von der Fakultät für Lebenswissenschaften
der Universität Leipzig
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt
von M. Sc. Claas-Thido Pfaff
geboren am 19.11.1981 in Pforzheim, Deutschland

Dekan: Prof. Dr. Tilo Pompe

Gutachter: Prof. Dr. Christian Wirth und Prof. Dr. Christine Römermann

Leipzig, den 23.08.2019
(Tag der Verteidigung)

Für meine Söhne Bela und Juri.
Mögen sie mit wachen Augen durch die Welt gehen
und sich ihre kindliche Neugier erhalten.

Inhalt

Danksagung.....	xi
Bibliographische Darstellung.....	xiii
Glossary.....	xv
Preface	17
Significance Statement	19
Structure.....	21
Abstract	23
Background.....	23
Material and Methods.....	25
Results and Summary	27
Zusammenfassung	33
Einleitung.....	33
Material und Methoden.....	36
Resultate und Ausblick.....	39
General Introduction.....	45
Specific introductions.....	49
Chapter One	49
Chapter Two.....	51
Chapter Three.....	53
General Material and Methods.....	55
GFBio project.....	55
BEF-China project and BEF-Data	55
The vocabulary creation	56

Chapter One	59
Chapter Two.....	69
Chapter Three	83
Abstract.....	83
Introduction.....	84
Material and methods.....	86
BEF-China.....	86
The Essential Annotation Schema for Ecology (EASE).....	87
The annotation process and complementing data.....	88
Selected aspects for the analysis and their background	90
Results.....	93
Data collection activity and throughput	93
Coverage and dynamics of topics	95
Public perception of the project.....	98
Collaboration structure.....	100
Discussion.....	103
Data collection activity and throughput	103
Coverage and dynamics of topics	104
Public perception and collaboration structure.....	104
Wrap up and outlook.....	106
Acknowledgement	108
Appendix	109
Graphics.....	109
Tables	110

General Discussion	113
Chapter One	113
Chapter Two.....	115
Chapter Three.....	120
Structural Synthesis.....	122
Conclusion	123
General Appendix	129
The EASE XSD.....	129
Tables chapter two.....	132
Author Contributions.....	135
Chaired sessions.....	138
Workshops.....	138
Talks.....	139
Posters	140
Related publications	140
Curriculum Vitae	141
Selbstständigkeitserklärung.....	142
Bibliography	143

Danksagung

An dieser Stelle möchte ich mich herzlich bei allen bedanken, die mich in den letzten Jahren in Bezug auf meine hier vorgelegte Arbeit unterstützt und vorangebracht haben. Allen voran möchte ich mich hier bei meinen beiden Betreuern Christian Wirth und Birgitta König-Ries bedanken, für die zahlreichen Chancen für eine professionelle und persönliche Entwicklung. Ich danke euch für euer entgegengebrachtes Vertrauen, die gute Zusammenarbeit und für all die Herausforderungen denen ich mich stellen und an denen ich über die Zeit wachsen durfte.

Mein Dank gilt auch all meinen Kollegen in der Arbeitsgruppe von Christian Wirth. Ich danke euch für das angenehme Miteinander, das starke füreinander, sowie für das gewissenhafte Lesen meiner Manuskripte und die konstruktive Kritik, die mich immer vorangebracht hat. Danken möchte ich hier auch meinen Coautoren Anne C. Lang, Sophia Ratcliffe, Xinxing Man, Karin Nadrowski und Mario Liebergesell, David Eichenberg und Helge Bruelheide. Ich danke euch herzlich für die intensive Zusammenarbeit und den fachlich kompetenten Informationsaustausch.

Weiterhin möchte ich mich auch bei Karin Nadrowski und David Eichenberg in ihrer Rolle als Betreuer und Förderer meines Studiums bedanken. Danke euch beiden für die vielen Stunden, die wir mit intensiven Diskussionen in fachlichem als freundschaftlichem Austausch verbracht haben. Mein Dank gilt hier des Weiteren auch Britta Kummer für die kontinuierliche Hilfestellung und die fachlich kompetenten Hinweise, welche für das Überleben im Papierdschungel der Bürokratie unabdingbar waren.

Ein ganz besonderer Dank gilt auch meiner Familie, die mich mit viel Herz und Geduld unterstützt und begleitet hat und ohne welche dieser Lebensabschnitt bei Weitem nicht so reibungslos und angenehm verlaufen wäre.

Bibliographische Darstellung

Claas-Thido Pfaff

A framework to support the annotation, discovery, and evaluation of data in ecology, for better visibility and reuse of data and an increased societal value gained from environmental projects

Fakultät für Lebenswissenschaften, Universität Leipzig

Dissertation

159 Seiten, 172 Literaturangaben¹, 28 Abbildungen, 8 Tabellen

Die vorliegende Dissertationsschrift beschäftigt sich im Kern mit der Verwendung von Metadaten in alltäglichen, datenbezogenen Arbeitsabläufen von Ökologen. Die vorgelegte Arbeit befasst sich dabei mit der Erstellung eines Rahmenwerkes zur Unterstützung der Annotation ökologischer Daten, der effizienten Suche nach ökologischen Daten in Datenbanken und der Einbindung von Metadaten während der Datenanalyse. Weiterhin behandelt die Arbeit die Dokumentation von Analysen sowie die Auswertung von Metadaten zur Entwicklung von Werkzeugen für eine Aufbereitung von Informationen über ökologische Projekte. Diese Informationen können zur Evaluation und Maximierung des aus den Projekten gezogenen gesellschaftlichen Mehrwerts eingesetzt werden. Die vorliegende Arbeit ist als kumulative Dissertation in englischer Sprache abgefasst. Sie basiert auf zwei Veröffentlichungen als Erstautor und einem zur Einreichung vorbereiteten Manuskript, welche im Folgenden aufgelistet sind.

1. Pfaff C-T, König-Ries B, Lang AC, Sophia R, Christian W, Xingxing M, Karin N (2015) rBEFdata: documenting data exchange and analysis for a collaborative data management platform. *Ecology and Evolution* 5(14):2890-2897. doi:10.1002/ece3.1547. (23)
2. Pfaff C-T, Eichenberg D, Liebergesell M, König-Ries B, Wirth C (2017) Essential Annotation Schema for Ecology (EASE)—A framework supporting the efficient data annotation and faceted navigation in ecology. *PLOS ONE* 12(10): e0186170. <https://doi.org/10.1371/journal.pone.0186170>. (24)
3. Pfaff C-T, Bruelheide H, Eichenberg D, König-Ries B, Wirth C; On the evaluation of ecological projects using their metadata

¹ Duplicates with the included publications are counted

Glossary

Cyberinfrastructure:

According to Atkins et al., 2003 it is: “infrastructure based upon distributed computers, information, and communication technology.” Thus it serves as the foundation for the transfer, storage and analysis of data.

Metadata:

It is the data about data which holds information about aspects of the data described which can help to clarify the content and the context of its creation. In other words, metadata allows preserving the semantics of the data described which is an essential prerequisite for the reuse of data. It can be descriptive (e.g., representing a resource for preservation and discovery), structural (describing the relation of compound objects) or administrative (e.g., information to help manage a resource).

Data life-cycle:

A concept which describes common steps along scientific interest in a circular fashion where research data is involved. The measures range from the data collection over their manipulation and analysis to their publication and preservation. In the presented work the data life-cycle is used in an extended form including the project planning as a prerequisite for the data life-cycle.

Full-text search:

According to Beall 2008, it is: “... the type of search a computer performs when it matches terms in a search query with terms in individual documents in a database and ranks the results algorithmically.”

Controlled vocabulary:

It provides the foundation for the organisation of knowledge. Controlled vocabularies provide a carefully selected list of authorised terms. They allow tagging units of interest like, e.g., datasets, for their potentially better retrieval during a search.

Folksonomy:

A folksonomy is a natural language vocabulary. It can be derived from the annotation of items (e.g., datasets) by the users of online platforms. The annotation is not restricted which allows a natural growth of the vocabulary along new items and with the needs of the community.

Taxonomy:

A taxonomy is a knowledge representation system based on a classification along the hierarchical relationship which exists among the related terms and their subterms. A taxonomic organisation of the terms derived from the periodic table of elements, for example, would organise “Carbon” underneath the term “Elements”.

Thesaurus:

A thesaurus is a knowledge organisation system which is quite similar to a taxonomy. However, it allows using a broader range of relations between terms which go beyond hierarchies. It typically defines relations like broader, narrower, related, synonym, and “use for” e.g., specified in the DIN 1463-1 and ISO 2788 standards.

Ontology:

An ontology is a vocabulary which comprises the concepts and categories in a domain of knowledge and their relations amongst each other. In other words, it is an explicit specification of a conceptualisation (Mankovskii et al. 2009)

Preface

Metadata have a long tradition in libraries with their roots in the ancient Greek library of Alexandria where the librarians attached tags on books containing the name of the author, the publication date and topic related keywords. This concept also has been used in museum collections for a long time to keep an overview of the items in a catalogue. With the broader proliferation of personal computers in the late 20th century, this concept of organising content was finally transferred into the digital age. Approximately a decade ago metadata started to gain attraction as a component to describe ecological data. It has been recognised as essential for the long-term success of ecology as a discipline. It allows to store, organise and discover data more efficiently and it enables the reuse and integration of data in analyses. Despite the benefits, many of the tools and workflows used by ecologists today still lack support for using metadata or do not yet exploit its full potential. This work focuses on the integration of metadata into the daily grind of ecologists and highlights problems which can be solved using it. Along those lines, this work shows how metadata can improve the data management and related workflows of ecologists and how it finally allows maximising the overall value of their precious work. I hope that this work will help to improve the image of data management and particularly the use of metadata which often is perceived as a burden forced upon researchers, who have to provide it; at least it is like this based on my personal experience. I further hope that this work falls on fertile ground and that it can inspire the reader to promote a wider adoption and a more creative use of metadata in ecology.

Significance Statement

Today ecology is more interdisciplinary than ever before. It is bridging different disciplines while working in close synergy with groups of interested people, e.g., in citizen science projects, or including expert and indigenous knowledge. Ecology gathers data in observations and experiments while instruments are often aiding the data collection. The broad instrumentation enables ecologists to collect an increasing amount of high-resolution data while in parallel it enables covering broad spatial, temporal and organismic scales. This data has the potential for addressing various important questions of public relevance. Further, ecology has become an invaluable source informing political decisions, e.g., with new ideas for solutions on the mediation of impacts related to an increased human resource usage or the pollution of natural systems. The data collected in ecology is highly valuable; carefully treated, used and reused it has the potential to unify theory and to enable policy decisions based on evidence rather than on instinct. In that context, the importance of data management along the full life-cycle of ecological data has become more apparent than ever before. The here presented work touches some of the integral parts of the data life-cycle in ecology. It is discussing the integration of metadata into routines and tools that researchers use and highlights the resulting benefits. Further, the presented work comprises two open source tools. These tools allow the documentation, discovery, the processing and evaluation of data in ecology. They were made publicly available along with the published papers while the open licensing contributes that the ecological research community can easily use them or adapt them to their needs if required.

Structure

The work presented here is a cumulative thesis. It consists of two peer-reviewed scientific articles and a chapter which is ready for publication. The two publications from here on will be referred to as chapters as well, for the sake of simplicity. All three chapters are located along the thematic surface which exists at the intersection between the disciplines of ecology and informatics. The chapters are further embedded in the broader context of a data life-cycle which includes steps that are related to data and project management along the interests of scientists (c.f. Figure 1). The data life-cycle involves steps related to planning a project, goes over the actual data collection before ending with the publication of the results in journals. The chapters are touching a subset of the steps in the data life-cycle while focusing on the use and potential benefits of metadata (c.f. Figure 2).

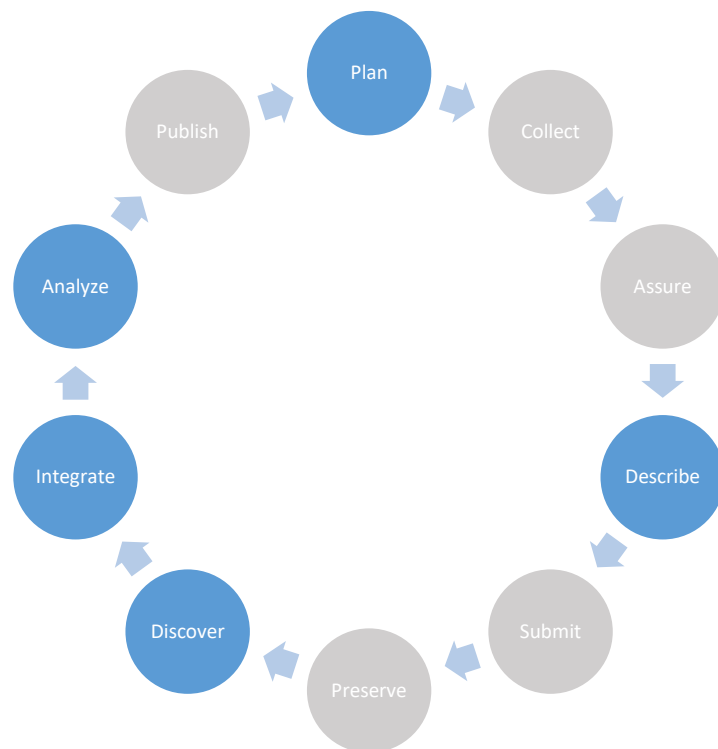


Figure 1 The life-cycle of data encompasses all tasks which are related to the handling of data in ecology. These steps include the planning of projects, the collection and organisation, the quality assurance and the metadata creation, preservation and discovery as well as the integration and analysis of data. Metadata can provide support for various of the aspects along the life-cycle of data. In this work, the focus is on the aspects highlighted in blue. Chapter one touches all the highlighted aspects (except planning), whereas, chapter two deals with the description and the discovery of data.

The third chapter deals with the evaluating of ecological projects based on their metadata to create feedback for the project management which is part of the project planning step in the life-cycle. The graph above represents information created by the Data Observation Network for Earth project (DataONE), adapted by Claas-Thido Pfaff. Noteworthy is the extended fasion of the life-cycle. Here it includes a planning step which is a prerequisite for a data life-cycle to start of.

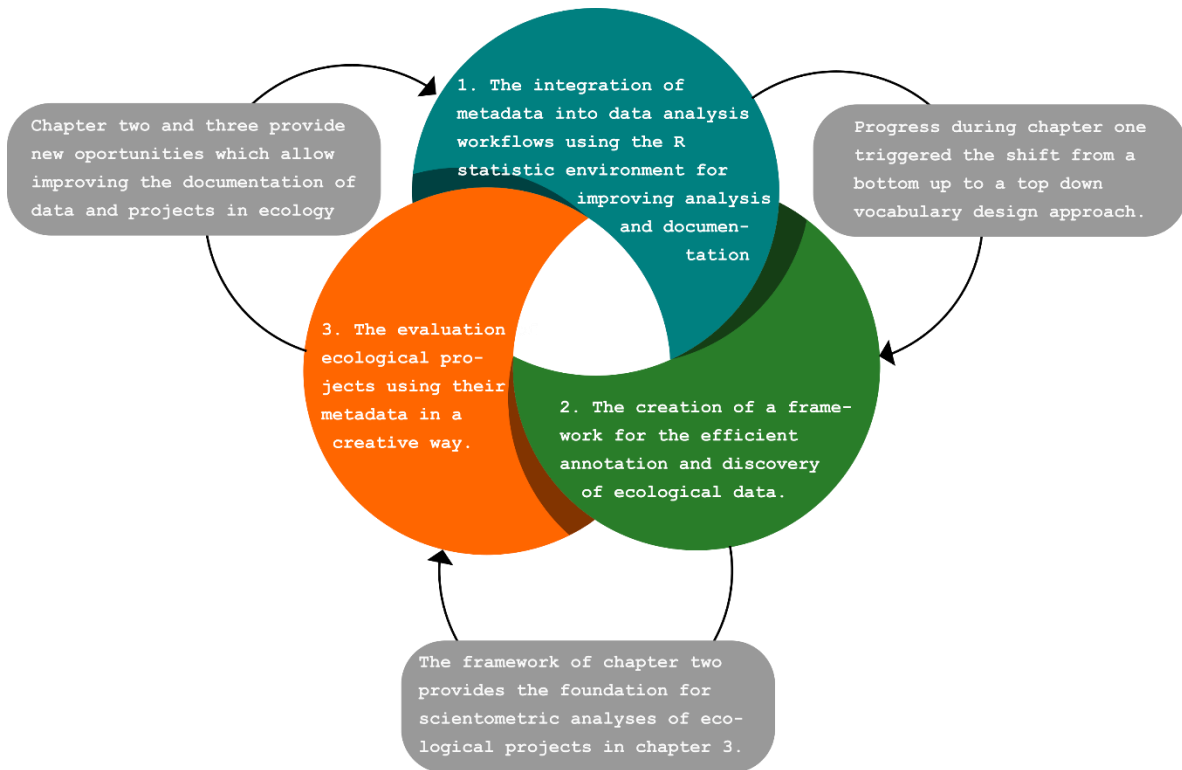


Figure 2 A schematic structure to present an overview which highlights how the three chapters inform and complement each other. Metadata can support research in ecology in various ways. The first chapter shows the integration of metadata into the analysis work-flow of R and how the documentation of the analysis helps to overcome the underrepresentation of information about the data processing in the final publication. The second chapter introduces the creation of a framework and a tool which support the description and the discovery of data in ecology. The third chapter makes use of the metadata framework and tool from chapter two in order to describe a decade of research from the long-term ecological research project (BEF-China). It also develops ideas for tools which allow a better overview and the evaluation of ecological projects.

Abstract

Background

Today ecology is recognised as an interdisciplinary and integrative oriented scientific discipline. It is characterised not only by growing global research networks and large-scale, long-term projects but also by its open and interdisciplinary research approach. The discipline is bundling increasing synergies of expertise across scientific disciplines (e.g., expert knowledge) as well as the broad involvement of interested people (citizen science, Silvertown 2009), indigenous experts (traditional knowledge, Pierotti 2010; Díaz et al. 2015) or hobbyist scientists (Kuhnert, Martin, and Griffiths 2010). The observational and experimental studies which are designed and carried out across ecology today are characterised by hand-collected data which is complemented by a growing amount of data derived from instruments (Michener and Jones 2012). The instrumentation includes gene sequencers and mass spectrometers in laboratories as well as various sensors which can be embedded in the environment or mounted on vehicles like satellites, aeroplanes and drones (Woodward, Lomas, and Kelly 2004; Anderson and Gaston 2013). The involvement of different groups of people in an interdisciplinary research approach, as well as the increasing use of instrumentation and new methodology, lead to the creation of a continuously growing amount of diverse and highly detailed ecological data (Borgman, Wallis, and Enyedy 2007).

The reuse of data became more attractive along with the growing amount of collected ecological data. It developed into an essential method in contemporary ecological synthesis projects (Arnqvist and Wooster 1995). The reuse of data comes along with many benefits (Reichman, Jones, and Schildhauer 2011). Mainly it allows extending the scope of new ecological studies on spatial and temporal scales as well as across the boundaries of environmental context. For example, meta-analyses reusing data across various scattered experiments have allowed the

development of the theory of multifunctionality in biodiversity/ecosystem functioning research (Reich et al. 2012). Further, it extended functional biodiversity research from plots to continents (Ratcliffe et al. 2016) and enabled the parameterisation of global climate models (Brovkin et al. 2009). The reuse of data provides the potential for continuously better understanding of our ecosystem. Thus, in turn, it can enable the development of solutions for land management and preservation of nature (Raupach et al. 2005). Finally, derived insights can be utilised to inform political decisions on the maintenance of ecosystem services (Maes et al. 2012). These services include the supply of food and potable water as well as the production of fibre on all of which humanity essentially depends upon (de Groot et al. 2012). Taken together this indicates that ecological data has an inherent value, which is going far beyond the interests of research projects or single individuals involved (e.g., publications).

Hence, ecological data should be carefully curated and preserved for the reuse in future generations of research. It can serve as the foundation for a better reproducible science, the precondition for synthesising knowledge and as a breeding ground for new ideas (M. D. Wilkinson et al. 2016). Despite the vast potential of the collected data, the past has shown that data is likely lost over the course of time if there are no countermeasures applied (P. Bryan Heidorn 2008). The need to process an increasing amount of data, the insight into the long-term value of data, but also the fear to lose valuable ecological data have been strong drivers behind the idea of data curation. The curation of data has been dealt with in developing theory (M. D. Wilkinson et al. 2016), various tools (e.g., Nadrowski et al. 2013; Kattge et al. 2011; Fegraus et al. 2005) and policy (European Science Foundation 2008). An important element of the theory is the life-cycle of data along which tasks of scientific interest are situated (Michener and Jones 2012; Rüegg et al. 2014). It starts with the planning of projects and progresses over the data collection

and metadata creation to the data use and its reuses up to the final preservation and publication of results.

Descriptive metadata can play a vital role in the steps along the life-cycle of data. Most importantly it can enable the reuse of data preserving information about the content and context of the data (Fegraus et al. 2005, e.g., methods used, or meaning of variables). The information may further be used to increase the visibility but also discoverability of data (Giles 2011, e.g., search based on information in the metadata) and as an enabler of integration and analysis of data (Michener and Jones 2012). Despite these benefits which come along metadata and proper documentation, the researchers in the ecological community are not yet fully utilising its potential.

This work is focusing on the creation and use of metadata along several steps included in the life-cycle of data in ecology. These steps are: 1.) The documentation of data including their processing and analyses. The first chapter discusses how to link raw data with derived data products (i.e., images, tables) and the knowledge (e.g., publications); 2.) The discovery and reuse of data: The second chapter focuses on the development of an ecological vocabulary and annotation framework to support the faceted navigation based search in ecology; 3.) The evaluation of ecological projects (which is part of the planning and management of projects in the life cycle). The third chapter discusses potential uses of metadata for the evaluation of ecological projects and on developing tools that provide feedback about resources in a project to the researchers.

Material and Methods

The first chapter focuses on the integration of an ecological metadata standard into the statistical analysis framework R (R Development Core Team 2016). R has been chosen as the environment of best acceptance across ecologists (Touchon and Mccoy 2016). An R package was developed allowing the exchange of data and metadata while bridging the R environment with the web-based data management

platform of the BEF-China project (Bruehlheide et al. 2014). BEF-China is a biodiversity and ecosystem functioning (BEF) experiment located in a subtropical forest ecosystem in the south-east of China (Jiangxi, Zhejiang). Three datasets from the BEF-China experiment and their metadata in Ecological Metadata Language (EML) format, (Fegraus et al. 2005) were used to create an example data analysis workflow. The analysis itself is about the nitrogen acquisition and retention in the study plots of the BEF-China project, as it is detailed in Lang et al. 2014. The example was employed to highlight how the metadata can be used in a real-world data analysis scenario. In particular, it is used for guiding decisions relevant along the synthesis and analysis of the data in R. Further, the chapter points out how a combination of the R environment with an online data management platform could help to enable the proper documentation of the processing of data. The documentation allows preserving information about the provenance of derived data products and results to finally help building a bridge between primary data repositories and knowledge repositories.

The second chapter deals with the annotation and discovery of ecological data. A vocabulary was created first based on a number of sources. These sources included terms from folksonomies of four ecological research projects (Fischer et al. 2010; Baeten et al. 2013; Bruehlheide et al. 2014; Weisser et al. 2017), ontologies (Degtyarenko et al. 2007; Buttigieg et al. 2013), textbooks, scientific publications and expert knowledge derived in workshops (c.f. general Material and Methods section). Based on the vocabulary, a framework was developed to support the annotation and discovery of data in ecology. It has been created along the idea of a multi-hierarchical classification of search objects to support a faceted navigation search approach (English et al. 2002). Several workshops have been carried out to help define the design principles of the vocabulary and to agree on crucial top-level categories or attributes for the annotation as well as on the contents for the annotation vocabulary. Based on collected ideas and agreements for attributes and

vocabulary a metadata schema was developed. It is using the Extensible Markup Language (XML) and the related schema definition language standard (XSD, Fallside and Walmsley 2004). Based on the schema a web-based application has been developed which in concert with the vocabulary allows for a fast annotation of arbitrary data formats in ecology (e.g., images, tables, videos). The vocabulary, the annotation schema, and the web-tool were finally published via GitHub as open source contributions. In that way, the ecological research community can easily access and adapt them to their needs if required.

The third chapter focused on information stored in metadata and how it can be used beyond its original purpose. For this, 250 datasets of the BEF-China experiment have been annotated using the EASE annotation framework and its web-based annotation tool (c.f. chapter two). The information from the EASE annotation was complemented with further metadata extracted from the data management platform of the BEF-China project and the Scimago citation database (e.g., the H-Index for journals published in). Several analyses were carried out describing processes along a decade of research in the BEF-China project like data collection events or the development of research topics. Further, the analyses included investigations into the networks formed between the researchers during data collection events and publication. The chapter discusses how the metadata can be used beyond the original purpose of documentation and how it can be applied along the context of project evaluations (e.g., tracking the collection of specific variables or the development of the topics in a project). Further, the chapter highlights how metadata can serve feedback mechanisms for researchers in ecological projects and how this can help to improve the project management to finally increase the overall value gained from a project.

Results and Summary

Despite the benefits of descriptive metadata many of the tools and data related workflows along the lifecycle of data in ecology still do not use or exploit its full

potential. The presented work shows how metadata can be used as a valuable resource of information and how the information can be turned into a benefit for ecological research and beyond. With a broader adoption and integration of metadata into the tools, researchers use for their daily work we can improve on various of the tasks along the life-cycle of data while in parallel supporting a sustainable scientific culture including, e.g., the curation and reuse of data but also the reproducibility of experiments. While metadata can help to better understand and process data, the R environment as the most widely used analysis framework in ecology is missing support for ecological metadata. Chapter one exemplified the integration of an ecological metadata standard into the R environment benchmarked along an example analysis workflow to discuss the resulting benefits. The first chapter also introduces the need for documentation beyond structured metadata. The provenance as a link between the raw data and finally derived data products and knowledge is essential for tracking down potential errors in analyses and finally for reproducibility of scientific studies. Here the chapter utilises the R environment in concert with the data management platform of the BEF-China project to achieve a more holistic form of documentation which is linking the original research idea, the raw data, manipulation steps and the derived knowledge in the form of publications in a single location.

Bundling raw data with analysis scripts and the data products and knowledge is an essential step in the right direction to finally help in closing the gap between primary data and knowledge repositories. However, storing only scripts as documentation for data processing has the downside that they can be hard to read and interpret. They require proper documentation on their own or a visual representation which helps to communicate better what the script does. First tools have been developed for the visual representation and the tracking of data manipulation along an R script of which none is yet widely used (e.g., RDataTracker and DDG Explorer, Lerner and Boose 2015). The integration of such

tools into the core of R would be useful in order to ensure a broader adoption by the community. Saving the data manipulation scripts along with the original research idea, the raw data and the derived products in the form of, e.g., tables, images and publications in an online repository has much potential to serve as full documentation of an analysis. However, storing documentation in a private repository also increases their likelihood of getting lost over time (P. Bryan Heidorn 2008). Thus preserving a documentation package would require a more holistic approach. It could involve the data producer on the one hand and the publisher on the other where both carefully curate and help to preserve not only data products but also the raw data and the documentation which is linking them.

Over time much effort has been put into the development of metadata standards for ecology (e.g., EML, ABCD, DwC). They allow for the proper documentation of context information of ecological data and collection items. When it comes to data discovery, however, they are typically lacking the required detail and explicitness. While the full-text descriptions are useful for humans, a computer and particular the most widely used full-text search algorithms cannot make much sense of it (Beall 2008). Significant progress has been made along with natural language processing (e.g., Chowdhury 2005) and ontologies (e.g., Walls et al. 2014), where both of which in the future might be able to help better approach or solve the problem of inaccessible information. Good quality ontologies, however, are the most limiting factor here. Their development is not trivial requiring diverse expertise on the one hand but also a broad agreement on concepts as the common language used for communication in the field of the covered research on the other. Today ontologies which are modelling topics relevant to ecology are somewhat underdeveloped (e.g. structurally like a thesaurus), topic-wise patchy or modelled along an unsuitable perspective. Particularly an ontology embracing the interdisciplinary character of ecology is lacking. Ontologies are typically representing a philosophically motivated model of the real world which likely

differs significantly from person to person. Further many parts in such models can be under philosophical dispute and thus represent uncertain or changing topics (particularly in very active scientific fields). Thus, merging existing, relevant ontologies is not a trivial solution at the moment. The merging of ontologies comes along many conflicts of subtle nature. For example, ontologies using the same terms while their modelling is different. These differences are unlikely resolved automatically and are even hard to resolve manually. This comes from the fact that the ideas which are modelled can happen to use the same words but their actual meaning can be completely incompatible. Thus, detangling the differences, while reusing similarities remains a mostly manual effort which requires not only a broad interdisciplinary expertise and time but likely also the involvement of research communities in an effort of clarification and agreement on the used terms. With the EASE framework, we bundled important information, from a perspective of ecological researchers onto their data, into an explicit framework for a fast and precise annotation of data in ecology. With faceted navigation based on annotated datasets, a search can be narrowed down to meet specific requirements. Suitable options for a restriction of the search space can help to discover relevant data for an analysis faster improving the reuse of data.

Beyond the purpose of documentation and discovery, information stored in metadata can be used in many more scenarios. Chapter three developed a use-case which highlights its use in the evaluation and feedback in ecological projects. Several analyses have been carried out to highlight internal processes of data collection and how topics in the project evolve in the course of time. Currently, it shows that the BEF-China project is finishing its data collection in time, that information about organisms is dominating the project and that it is sharing workload across the involved individuals with data collection. In publications, researchers are networking more with each other integrating the data to derive new knowledge. By doing so, it reaches a relatively high amount of good impact

journals. While the meaningfulness of the shown analyses is restricted at the moment (based on one project only), it has the potential to be developed into a more general framework. A workshop involving researchers from Ecology and Scientometrics could help to drive the framework towards an increased value. Particularly, analysing more projects will help to increase the interpretability of the results. This effort might pay out with potential predictability for other projects. This predictability could help to finally guide the funding of new projects (e.g. specifically oriented projects and topics are likely to take more time or workforce). Implementing such an analysis framework into the management tools like BExIS and BEF-Data could finally help to provide better access to the presented ideas for a broader audience. The tools could take shape as a graphical dashboard, e.g., built into user-profile pages to provide feedback for each involved researcher in a project (e.g., find potential collaboration partners) and for principal investigators helping with project management (e.g., are variables measured in time, resolution and the intensity planned initially and are topics covered as outlined).

Zusammenfassung

Einleitung

Die moderne Ökologie hat sich zu einer hochgradig integrativ orientierten wissenschaftlichen Disziplin entwickelt. Sie zeichnet sich nicht nur durch eine wachsende Zahl globaler Forschungsnetzwerke und räumlich groß angelegte und langfristige laufende Projekte aus, sondern auch durch ihren offenen, interdisziplinären Forschungsansatz (e.g. Bruehlheide et al. 2014). Dadurch finden Informationen durch diverse Projekte und Personengruppen unterschiedlichster Disziplinen Eingang in die Ökologie. Die Quellen umfassen Expertenwissen sowie Beiträge von Laien (Citizen Science z.B. Silvertown 2009), indigenen Experten (tradiertes Wissen, Pierotti 2010; Díaz et al. 2015) oder Hobby-Wissenschaftlern (Kuhnert, Martin, and Griffiths 2010). Die in der Ökologie durgeführten Studien spannen einen Bogen von reinen Freilandbeobachtungen bis hin zu kontrollierten Experimenten. Die Daten werden dabei oft noch vollständig manuell erhoben. Diese Erhebungen werden jedoch zunehmend durch Daten ergänzt die mit Hilfe technischer Instrumente erfasst werden (Michener and Jones 2012). Als Instrumente kommen Gensequenzierer und Massenspektrometer in Laboren zum Einsatz sowie ein Spektrum diverser Sensoren für den mobilen Einsatz im Freiland. Die Sensoren werden dabei sowohl zur lokalen Erhebung mit fester Installation in Ökosystemen eingesetzt als auch zur Erhebung von Daten aus der Ferne unter Zuhilfenahme von Drohnen, Flugzeugen oder Satelliten (Woodward, Lomas, and Kelly 2004; Anderson and Gaston 2013). Die Vielzahl der Datenquellen und beitragenden Disziplinen hat schlussendlich zu einem starken Anwachsen einer vielfältigen und detaillierten ökologischen Datenbasis geführt (Borgman, Wallis, and Enyedy 2007).

Die Wiederverwendung von erhobenen ökologischer Daten gewann mit zunehmender Menge immer weiter an Attraktivität. Sie entwickelte sich im Laufe der Zeit zu einer wesentlichen Methoden, die heute in ökologischen Syntheseprojekten

regelmäßig Anwendung findet (Arnqvist and Wooster 1995). Die Wiederverwendung von Daten bringt verschiedene Vorteile mit sich (Reichman, Jones, and Schildhauer 2011), allen voran aber die Möglichkeit den Geltungsbereich ökologischer Studien kontinuierlich über räumliche und zeitliche Skalen hinweg auszuweiten als auch die Grenzen des abgedeckten Umweltkontextes einer Studie zu erweitern. Ein Beispiel dafür sind groß angelegte Metaanalysen, welche die Daten diverser Experimente zusammengeführt haben und dabei die Bildung einer zentralen Theorie zur Multifunktionalität in der Biodiversitäts-/Ökosystemfunktionsforschung erst ermöglicht haben (Reich et al. 2012). Weitere Beispiele hierfür sind zunehmende Erschließung großer räumlicher Skalen in der funktionellen Biodiversitätsforschung bis hin zur kontinentalen Ebene (Ratcliffe et al. 2016) sowie die Parametrisierung von globalen Klimamodellen (Brovkin et al. 2009). Die Wiederverwendung von Daten bietet ein hohes Potenzial, um unser Verständnis über Ökosysteme kontinuierlich zu erweitern. Aus der genaueren Kenntnis der Ökosysteme können wiederum neue Lösungen für eine bessere Landwirtschaft und den Naturschutz entwickelt werden (Raupach et al. 2005). Schließlich können die Erkenntnisse auch einen Beitrag dazu leisten, dass politische Entscheidungen zum Erhalt der Ökosysteme und ihrer Dienstleistungen informiert werden können (Maes et al. 2012). Die Dienstleistungen umfassen dabei die Versorgung mit Nahrungsmitteln und Trinkwasser genauso wie die Herstellung von Rohstoffen (z.B. Fasern als Baumaterial), auf welche die Menschheit im Wesentlichen angewiesen ist (de Groot et al. 2012).

Diese Beispiele zeigen, dass ökologische Daten einen inhärenten Wert besitzen, welcher weit über die Interessen einzelner Personen oder ganzer Forschungsprojekte hinausgeht. Daher sollten ökologische Daten möglichst sorgfältig gepflegt und für die Wiederverwendung durch zukünftige Forschungsgenerationen aufbereitet und verwahrt werden. Sie können dann helfen die Grundlage für eine reproduzierbare Wissenschaft zu bilden und darüber hinaus die Wegbereiter für die

Synthese von Wissen und der Nährboden für neue Ideen sein (M. D. Wilkinson et al. 2016). Leider hat die Vergangenheit gezeigt, dass einmal erhobene Daten im Laufe der Zeit leicht verloren gehen, wenn keine Gegenmaßnahmen ergriffen werden (P. Bryan Heidorn 2008). Glücklicherweise sind sowohl die Erkenntnis über den langfristigen Wert von ökologischen Daten, als auch die Angst um deren Verlust starke Treiber hinter der Entwicklung neuer Ideen um die Aufbereitung und Sicherung von Datenbeständen. Die Wichtigkeit der Pflege von wissenschaftlichen Daten hat sich sowohl in der wissenschaftlichen Theorie (M. D. Wilkinson et al. 2016) als auch in verschiedenen Software-Werkzeugen (Fegraus et al. 2005; Nadrowski et al. 2013; Kattge et al. 2011) und Regelwerken niedergeschlagen (European Science Foundation 2008). Ein zentrales Element der Theorie ist der Lebenszyklus von Daten (Michener and Jones 2012; Rüegg et al. 2014). Angefangen mit der Planung und dem Management von Projekten bewegt sich der Zyklus über die Datenerhebung und Metadatenerstellung sowie die Datennutzung und deren Nachnutzung bis hin zur endgültigen Aufbewahrung und Veröffentlichung der Ergebnisse.

Metadaten können eine zentrale Rolle in den einzelnen Schritten des Lebenszyklus von Forschungsdaten einnehmen. Ein wichtiger Aspekt, der durch gute Dokumentation unterstützt wird, ist die Nachnutzung der Daten. Dabei ist es von besonderer Bedeutung, dass sowohl Informationen über den Inhalt als auch den Kontext von Daten erhalten bleiben und genutzt werden können (Fegraus et al. 2005, z. B. die verwendeten Methoden oder die Bedeutung der erhobenen Variablen). Die Informationen können ferner verwendet werden, um die Sichtbarkeit von Daten zu erhöhen, was auch deren Auffindbarkeit in Datenbanken zugutekommt (Giles 2011), z. B. Eine Suche basierend auf Informationen in den Metadaten). Zusammengekommen ermöglichen Metadaten eine bessere Integration und Analyse von Daten (Michener and Jones 2012). Trotz der Vorteile, die mit einer guten Dokumentation

in Form von Metadaten einhergehen, nutzen Forscher der Ökologie bei weitem noch nicht deren volles Potenzial.

Die hier vorgelegte Arbeit konzentriert sich im Wesentlichen auf die Erstellung und Verwendung von Metadaten entlang mehrerer Schritte im Lebenszyklus von Daten in der Ökologie. Die Schritte sind: 1) Die Dokumentation von Daten einschließlich deren Verarbeitung in einer Analyse. Im ersten Kapitel wird dabei erläutert, wie Rohdaten inklusive der von ihnen abgeleiteten Datenprodukte wie Grafiken, Tabellen als auch Wissen in Publikationsform verknüpft und bewahrt werden können. 2) Die Suche und Nachnutzung von Daten: Das zweite Kapitel konzentriert sich dabei hauptsächlich auf die Entwicklung eines Vokabulars und Rahmenwerkes zur Unterstützung der Annotation und der zielgenauen Suche ökologischer Daten. 3) Die Bewertung ökologischer Projekte als Teil der Planung und Verwaltung im Vorlauf zum eigentlichen Lebenszyklus von Daten. Das dritte Kapitel befasst sich mit den Möglichkeiten die Metadaten bieten, um den Erfolg und den Integrationsgrad ökologischer Projekte zu bewerten.

Material und Methoden

Das erste Kapitel konzentriert sich auf die Integration eines ökologischen Metadatenstandards in die statistische Programmiersprache R (R Development Core Team 2016). R wurde für die Implementation gewählt, da es eine weite Verbreitung in der Ökologie gefunden hat (Touchon and Mccoy 2016). Es wurde ein R-Paket entwickelt, das den Austausch von Daten und Metadaten ermöglicht und dabei die R-Umgebung mit der webbasierten Datenverwaltungsplattform des BEF-China-Projekts verbindet (Nadrowski et al. 2013; Bruelheide et al. 2014). BEF-China ist ein Experiment welches sich im Umfeld der Biodiversitäts- und Ökosystemfunktionsforschung (BEF) bewegt. Die Plots des Experimentes liegen in einem subtropischen Waldökosystem im Südosten Chinas zwischen den Provinzen Jiangxi und Zhejiang. Drei repräsentative Datensätze wurden verwendet, die innerhalb des BEF-China-Experiment erhoben wurden und die mit ihnen verbundenen Metadaten im

EML-Format (Ecological Metadata Language). Darauf basierend wurde ein typischer Datenanalyse-Workflow in der Ökologie nachgebildet. In der Analyse geht es um die Stickstoffaufnahme und -allokation in den Untersuchungsflächen des BEF-China-Projekts, so wie es im Detail in Lang et al. 2014 beschrieben ist. Anhand des Beispiels wurde aufgezeigt, wie die Informationen in Metadaten in einem realen Analyseszenario Verwendung finden können. Insbesondere erlauben sie Entscheidungen (z.B. Kompatibilität von Variablen basierend auf der Einsicht in die verwendeten Methoden), die für eine Synthese und reibungslose Analyse der Daten notwendig sind. Des Weiteren zeigt das erste Kapitel, wie die Kombination aus einer Analyseumgebung für Daten und einer Online-Datenverwaltungsplattform einen Beitrag leisten kann hin zu einer übergreifenden Dokumentation wissenschaftlicher Daten. In der Dokumentation können Informationen bewahrt werden, die über die genaue Herkunft abgeleiteter Datenprodukte Auskunft geben, um schließlich eine Brücke zwischen primären Datenrepositorien (z.B. Datenbanken wie die von BEF-China) und Wissensrepositorien (z.B. Zeitschriften) zu schlagen.

Das zweite Kapitel befasst sich mit der Annotation und Suche ökologischer Daten. Es erarbeitet zunächst ein Vokabular, welches auf eine breiten Quellenbasis fußt. Zu diesen Quellen zählen das zur Annotation von Daten verwendete Vokabular von vier ökologischen Forschungsprojekten (Fischer et al. 2010; Baeten et al. 2013; Bruelheide et al. 2014; Weisser et al. 2017), Ontologien (z.B. Degtyarenko et al. 2007; Buttigieg et al. 2013), Lehrbücher, wissenschaftliche Veröffentlichungen und die Ergebnisse diverser kleiner Workshops zur Erfassung von Expertenwissen (siehe Abschnitt Allgemeine Material- und Methodenmethoden). Basierend auf dem erfassten Vokabular wurde ein Rahmenwerk entwickelt, welches die Annotation und Suche von Daten in der Ökologie unterstützen kann. Das Rahmenwerk selbst wurde nach der Idee einer mehrstufigen Klassifizierung von Suchobjekten erstellt, um einen auf Facetten basierten Suchansatz zu realisieren (English et al. 2002). Ge-

gegenstand der genannten Workshops waren neben den Vokabularien auch die Gestaltungsprinzipien des Rahmenwerkes. Dabei wurden sowohl wichtige Attribute für die Annotation (z.B. Chemische Elemente) als auch die jeweils zu verwendenden Vokabular diskutiert und festgelegt (hier z.B. die Elemente nach Periodensystem). Basierend auf den gesammelten Ideen und Vereinbarungen über sinnvolle Attribute und das Vokabular wurde ein Metadatenschema entwickelt. Es verwendet die *Extensible Markup Language* (XML) und den dazugehörigen Standard für die Definition eines Schemas (XSD, Fallside and Walmsley 2004). Basierend auf dem Schema wurde eine Web-Anwendung entwickelt, die in Verbindung mit dem Vokabular eine schnelle Annotation beliebiger Datenformate in der Ökologie ermöglicht (z. B. Bilder, Tabellen, Videos). Das Vokabular, das Annotationsschema und das webbasierte Tool wurden schließlich alle über GitHub als Open Source-Beiträge veröffentlicht. Auf diese Weise ist ein einfacher Zugang zu den Werkzeugen für die ökologische Forschungsgemeinschaft sichergestellt. Darüber hinaus ist es somit auch einfacher, die Werkzeuge bei Bedarf an die Bedürfnisse neuer Projekte anzupassen.

Das dritte Kapitel befasst sich mit der Verwertung von Informationen, die in Metadaten gespeichert sind. Im Besonderen liegt der Fokus hier darauf, in wie fern die Informationen über ihren ursprünglichen Zweck hinaus intelligent genutzt werden können. 250 Datensätze des BEF-China-Experiments wurden verwendet und mit dem EASE-Annotations-Werkzeug annotiert (vgl. Kapitel 2). Die Informationen der EASE-Annotation wurden durch weitere Metadaten ergänzt, welche aus der Datenverwaltungsplattform des BEF-China-Projekts extrahiert wurden (z.B. Teilprojekte und Beschreibungen). Zudem wurden Informationen aus der Scimago-Datenbank (z. B. dem H-Index für Zeitschriften) entnommen, um die Metadaten des Projektes zu ergänzen. Basierend auf den gesammelten Informationen wurden

diverse Analysen durchgeführt, um verfügbare Ressourcen und Prozesse innerhalb des BEF-China Projektes entlang eines Zeitraums von 10 Jahren zu beschreiben.

Die Analysen umfassen unter anderem die Strukturen der Datenerhebung sowie die Abdeckung gemessener Variablen und die Entwicklung von Themenbereichen. Ferner umfasst das Kapitel Analysen der Netzwerke, die sich zwischen den Forschern im Projekt während der Datenerhebungen und der Veröffentlichung von Ergebnissen gebildet haben. In diesem Kapitel wird weiterhin diskutiert, wie die Metadaten im Kontext einer Projekt-Evaluierung angewendet werden können (z. B. Prüfung auf Erfassung bestimmter Variablen oder Entwicklung des Thematischen Fokus eines Projektes). Des Weiteren wird diskutiert, wie Metadaten und die vorgestellten Analysen als Mechanismus der Rückmeldung für Forscher in ökologischen Projekten dienen können und wie dies dazu beitragen kann, das Projektmanagement zu verbessern und letztendlich den Gesamtwert eines Projekts zu steigern.

Resultate und Ausblick

Trotz diverser Vorteile von beschreibenden Metadaten, wird deren volles Potential in der Ökologie noch längst nicht ausgeschöpft. Die hier vorgelegte Arbeit beleuchtet Metadaten als wertvolle Informationsquelle. Sie zeigt, wie die in Ihnen enthaltenen Informationen dabei helfen können, einen Mehrwert für ökologische Forschungsprojekte zu generieren. Durch eine breitere Akzeptanz und die Integration von Metadaten in die entlang des Lebenszyklus von Daten verwendeten Werkzeuge ist es potentiell möglich, nicht nur die Handhabung von Daten zu vereinfachen, sondern auch im gleichen Zuge, eine nachhaltigere wissenschaftliche Kultur zu fördern. Dabei stärkt zum Beispiel die Dokumentation als Teil der Kuratierung von wissenschaftlichen Daten die Nachnutzung und schlussendlich potentiell auch die Reproduzierbarkeit von Experimenten als ein fundamentales Prinzip der Wissenschaften.

Während Metadaten einen wichtigen Beitrag dazu leisten können, dass erhobene Daten besser verstanden und effizienter verarbeitet werden, fehlt es R als am weitesten verbreiteter Analyse-Software in der Ökologie an einer Unterstützung für ökologische Metadaten. In Kapitel eins ist die Integration eines Standards für ökologische Metadaten in die R-Umgebung dargestellt. Die daraus erwachsenden Vorteile werden anhand einer beispielhaften ökologischen Analyse beleuchtet. Das erste Kapitel führt auch die Notwendigkeit einer Dokumentationsform an, die über strukturierte Metadaten hinausgeht. Ein Bindeglied zwischen den Rohdaten und den daraus abgeleiteten Produkten in Form von Tabellen, Grafiken oder erzeugten Texten kann schlussendlich ein solides Fundament zur Dokumentation der Abstammung des aus Daten abgeleiteten Wissens bilden. Dieses Bindeglied ist für das Auffinden potenzieller Fehler in abgeschlossenen Analysen unerlässlich und dient schließlich nicht nur der Reproduzierbarkeit von Analysen, sondern auch der Überprüfung wissenschaftlicher Ergebnisse und des darauf fußenden Wissens. Für die Erstellung einer solchen Verknüpfung wird in Kapitel eins die R-Analyse-Umgebung in Verbindung mit der Datenmanagement Plattform des BEF-China-Projekts verwendet. Um eine ganzheitlichere Form der Dokumentation zu erzielen, werden sowohl die verwendeten Rohdaten als auch die Manipulationsschritte der Daten (R-Code), Datenprodukte und abgeleitetes Wissen in Form von Publikationen an einem einzigen Ort gespeichert. Die gemeinsame Speicherung von Daten und Dokumentation ist ein wesentlicher Schritt, um die Lücke zwischen Forschungsideen, den erzeugten Primärdaten und dem Abgeleiteten Wissen zu schließen.

Das alleinige Speichern von Skripten als Dokumentation für die Datenverarbeitung hat jedoch auch Nachteile. Skripte können unter Umständen schwer verständlich sein (je nach Komplexität der Datenmanipulation oder auch Stil des Programmierers), was im Gegenzug die Interpretierbarkeit und Überprüfung der Resultate be-

einträchtigt. Die Skripte benötigen daher eine eigene Dokumentation oder eine visuelle Aufarbeitung, die hilft, die Datenmanipulation besser zu kommunizieren. Erste Werkzeuge für die visuelle Darstellung und Verfolgung der Datenmanipulationen innerhalb eines R-Skripts wurden bereits entwickelt, von denen jedoch noch keines weit verbreitet ist (z. B. RDataTracker und DDG Explorer, Lerner and Boose 2015). Das Speichern von Analyse-Skripten nebst der ursprünglichen Forschungs-idee (z.B. Anträge), den Rohdaten und den abgeleiteten Produkten bietet gutes Potenzial für die Erhaltung von wichtigen Informationen. Die Speicherung in einem privaten Datenspeicher erhöht jedoch auch die Wahrscheinlichkeit, dass die gesamte Dokumentation im Laufe der Zeit verloren geht (P. Bryan Heidorn 2008). Insgesamt legt dies einen ganzheitlicheren Ansatz nahe, der sich sowohl der Erstellung als auch der Erhaltung eines vollständigen Dokumentationspaketes widmet. Dabei sollten Geldgeber einerseits aber auch Datenproduzenten und die Herausgeber (Verlage) andererseits mit einbezogen werden, um gemeinsam einen Standard zu erarbeiten.

Im Laufe der Zeit ist viel Energie in die Entwicklung von Metadatenstandards für die Ökologie geflossen (z. B. EML, ABCD, DwC). Sie ermöglichen eine strukturierte Dokumentation, um Informationen über den Kontext ökologischer Daten und Sammlungsgegenstände festzuhalten. Für den Anwendungsfall einer Suche von Daten sind sie jedoch nur bedingt geeignet. Dies ist mitunter ihrem Fokus auf text-basierten Beschreibungen geschuldet. Während die Texte für einen Menschen unabdingbar für die Interpretation sind, kann ein Computer mit einer einfachen Volltextsuche die Information leider nicht sehr effizient und oft nicht zielführend verarbeiten (Beall 2008). Die Informationen, welche in einem Text stecken, sind für den Algorithmus nicht einfach zugänglich und es bedarf elaborierterer Methoden für deren Erschließung. Im Besonderen sind hier die Verarbeitung natürlicher Sprache (z. B. Chowdhury 2005) und Ontologien (z. B. Walls et al. 2014) wichtig. Sie bieten das Potential, in der Zukunft einen signifikanten Beitrag dazu

zu leisten, die eben erwähnten Unzugänglichkeiten auszuräumen und Informationen in Texten für eine Suche besser auszuwerten. Ontologien von guter Qualität sind hier jedoch der limitierende Faktor. Ihre technische Entwicklung ist nicht trivial, da sie interdisziplinäre Fachkenntnisse erfordert. Darüber hinaus verlangt sie auch die Mitarbeit der späteren Nutzerschaft. Die Nutzer sollten sich im optimalen Fall über die Konzepte einig werden, die als Grundlage einer gemeinsamen Welt-sicht und Sprache für die Kommunikation der Resultate ihrer Forschung dienen sollen.

Heutzutage haben Ontologien, die relevante Themen der Ökologie berühren, die Tendenz zu struktureller Einfachheit (z.B. wie Thesauri). Sie sind, thematisch unvollständig oder modellieren Wissen, welches für die Ökologie relevant ist, aus einer anderen fachlichen Perspektive (z.B. die eines Chemikers). Insbesondere fehlt der Ökologie eine übergreifende Ontologie (top-level Ontologie), die den gesamten multidisziplinären Charakter der Disziplin umfasst. Daher können existierende Modelle nicht einfach auf ein gemeinsames Grundmodell zurückgreifen. Das erschwert das Zusammenführen einzelner relevanter Ontologien zu einer großen oder einer kleineren spezifischen. Das Zusammenführen von Ontologien birgt viele Konflikte subtiler Natur, die nur schwer programmatisch zu lösen sind (z. B. zwei Ontologien verwenden exakt die gleichen Begriffe, während sich jedoch ihre textbasierte Beschreibung oder ihre Modellierung unterscheiden). Die Analyse der Unterschiede, die zwischen Ontologien existieren, und die Wiederverwendung von Gemeinsamkeiten sind daher von großem manuellem Aufwand geprägt. Während EASE noch nicht als Ontologie modelliert ist, bündelt das Rahmenwerk dennoch Informationen aus der Sichtweise eines ökologischen Forschers auf seine Daten und spezifische Analysen in einem expliziten Modell. Es kann sowohl der Annotation als auch der Suche ökologische Daten dienlich sein. Mit einer facettenbasierten Suche, die auf annotierten Datensätzen mit EASE basiert, können Suchergebnisse

so eingegrenzt werden, dass Anforderungen für bestimmte Analysen erfüllt werden (z.B. räumliche und zeitliche Auflösung von gemessenen Variablen).

Über den Zweck der Dokumentation und Ermittlung hinaus können in Metadaten gespeicherte Informationen in vielen weiteren Szenarien verwendet werden. Kapitel drei entwickelte einen Anwendungsfall, der die Verwendung bei der Bewertung und für die Rückmeldung von wichtigen Informationen in ökologischen Projekten hervorhebt. Hierbei wurden mehrere Analysen durchgeführt, um interne Prozesse wie die Datenerfassung und die Entwicklung von Themen im Projekt im Laufe der Zeit aufzuzeigen. Dabei zeigt sich, dass Informationen über Organismen das Projekt dominieren und die Arbeit während der Erhebung von Daten in kleineren Netzwerken stattfindet. Für die Publikationen vernetzen sich Forscher stärker miteinander, indem sie die Daten integrieren, um neues Wissen abzuleiten. Dadurch kann das Projekt vermutlich viele interessante Publikationen erstellen, die in einflussreichen Zeitschriften publiziert werden. Die Aussagekraft der gezeigten Analysen ist jedoch im Moment noch begrenzt, da sie zum einen nur ein Projekt abdecken, und zum anderen auch nicht alle Anwendungsfälle berücksichtigen. Sie haben jedoch das Potenzial in ein allgemeineres Rahmenwerk überführt zu werden. Ein oder mehrere Workshops mit Forschern aus den Fachbereichen Ökologie und Scientometrics könnte hierbei einen signifikanten Beitrag leisten. Hier kann Expertenwissen eingeholt und diskutiert werden, welche Informationen aus Metadaten noch aufgearbeitet werden können, um Forschern in Projekten einen optimaleren Überblick über das Projekt zu gewähren um damit schlussendlich einen Mehrwert zu generieren (z.B. neue Kollaborationen). Insbesondere kann die Analyse weiterer ökologischer Projekte dazu beitragen, die Interpretierbarkeit der Ergebnisse zu verbessern. Dies Ergebnisse einer größer angelegten Analyse könnten sich schlussendlich potentiell auch auf die Planung neuer Projekte auswirken. Sie könnten zum Beispiel dabei helfen Fragen organisatorischer Natur zu beantworten (z.B. wie viel

Zeit oder Arbeitskraft brauchen Projekte mit einer gewissen thematischen Orientierung typischerweise, um brauchbare Ergebnisse zu erzielen?).

Die Implementierung eines solchen Rahmenwerkes zur Analyse von Projekten in Daten-Management Plattformen wie BExIS und BEF-Data könnte schließlich ihren Beitrag leisten, die vorgestellten Ideen und die daraus erwachsenden Vorteile einem breiten Publikum zugänglich zu machen. Die Werkzeuge könnten in Form eines grafischen Dashboards Gestalt annehmen, z. B. in Benutzerprofilseiten integriert, um jedem beteiligten Forscher in einem Projekt spezifisch zugeschnittenes, nützliches Feedback zu geben (z. B. Vorschlag potenzieller Kooperationspartner). Für projektverantwortliche Wissenschaftlicher kommen dann zum Beispiel noch Werkzeuge hinzu, die dabei helfen, einen Umfassenden Überblick zu schaffen um das Projektmanagement zu unterstützen. Hier könnte aufbereitet werden ob Variablen wie vereinbart in der richtigen zeitlichen und räumlichen Auflösung erfasst und ob sich das Projekt thematisch in der geplanten Richtung entwickeln.

General Introduction

Ecology aims at understanding the interactions between the life forms on Earth and the interactions they form with the environments they are occurring in (Friederichs 1958). While the first works with an ecological character reach all the way back to the antique (e.g., Aristotle 384 - 322 BC and Theophrastos 371 - 287 BC, Egerton, 2001), ecology started to take shape much later during the 18th and 19th century. The term "Ecology", e.g., has been coined back in the year 1866 by the German zoologist Ernst Haeckel (1834 - 1919) in his work about a general morphology of organisms (Haeckel 1866). It combines two Greek words being "oikos" (οἶκος) and "logia" (λογία) which translate into "environment" and the "study of ...". Ecology in its beginning was significantly influenced by the scientists and naturalists who travelled and explored the world in the 18th century describing nature based on their observations. These descriptions comprise, e.g., contributions from Alexander von Humboldt (1769 - 1859, with his integrative work on botanical geography) Alfred Russel Wallace (1823 - 1913) and Charles Darwin (1809 – 1882, with the theory of evolution, Darwin, 1859).

From the beginning on ecology has always been interdisciplinary due to its broad scope of interest which encompasses the observation and the study of complex natural systems and their interactions with each other (e.g., Wright and Bartlein 1993). The broad focus finally promoted the formation of a diverse range of sub-disciplines which complement and inform each other standing in synergy with disciplines like, e.g., chemistry, geology or meteorology (Egerton 2012). Ecology has started as an observational discipline and from there went through multiple stages of evolution moving towards a quantitative science. Consecutively, the discipline became characterised by emerging aspects like the model development and a resulting generalisation of theory, and later on by computationally intensive simulations (Petrovskii and Petrovskaya 2012). In parallel to this, ecology has developed from rather simple to more complex project structures involving many

individual researchers. Working in larger collaborations allowed the distribution of labour across individual scientists but also better utilisation of the expertise of specialised sub-disciplines represented by them (Hobbie et al. 2009). In other words, ecology has become a highly collaborative and network-based discipline over time (Borer et al. 2017).

Since a few decades, ecology is transforming into a more data-intensive and globally oriented discipline (Michener and Jones 2012). This trend was heralded by the deployment of the first satellites in the '60s and '70s of the 20th century (e.g. Vanguard 1 for geodetic measurements, Kwa 2005) and carried on with deploying satellite networks like the Earth Observation System (EOS) which enable an observation of the earth from a new perspective and in fine-grained detail (e.g. Landsat). Finally, the trend towards a data-driven science became manifest in the multitude of methods and instrumentation which are broadly used today. The instrumentation comprises, for example, high throughput gene sequencing (Venter et al. 2001), hyperspectral cameras, and a wide range of different sensors which are directly embedded into the environment (Collins et al. 2006) or used for remote sensing with satellites, airplanes and drones (Anderson and Gaston 2013). Today, ecology is a mix of its past influences while the instrumentation of the discipline promotes the collection of data at an increasing pace and in a finer resolution than ever before (Porter et al. 2009; Jones et al. 2006). At the same time, it spans a broad scale which is ranging from molecules (Blomquist and Bagnères 2010) up to the whole biosphere (Hughes 2000).

With the increasing ability of individuals and small groups to collect massive amounts of data, the need for sophisticated data management and trustworthy cyber-infrastructure for ecology became apparent (Atkins et al. 2003). The increasing awareness of the long-term value (Fegraus et al. 2005) and the potential loss of valuable, and sometimes irreplaceable ecological data (P. Bryan Heidorn 2008) further promoted a stimulating environment for the development of tools

and standards to support researchers along the life-cycle of data (Fegraus et al. 2005; Wieczorek et al. 2012; Berkley et al. 2001; Nadrowski et al. 2013). This includes tools for the planning of projects, the collection of research data, the data analysis, the curation and reliable storage of data as well as the publication of data and the derived research results (Michener and Jones 2012; Rüegg et al. 2014). In parallel to this, policies were developed by publishers and funding agencies to complement the standards and tools with the expectation that the publicly financed research data has to be curated carefully before it is finally published in openly accessible repositories (European Science Foundation 2008). This is important, as ecological data has an inherent value of societal relevance which can be unleashed only if the data is carefully documented and broadly accessible for the reuse in new research ideas (M. D. Wilkinson et al. 2016; Roche et al. 2015). In this context, extensive research networks have emerged and cyber-infrastructure projects were set up being responsible for data collection, curation, preservation, and dissemination to finally ensure a broad visibility and a better access to the data in a long-term perspective (Adams 2012, Tenopir et al. 2011, Diepenbroek et al. 2014).

Today ecology is recognised as unifying scientific discipline. It bundles competences and data from science but also includes society and culture (e.g., citizen science, expert knowledge). Whole institutions were dedicated to enable the use and reuse of data but particularly to facilitate the synthesis of data across the boundaries of scientific disciplines and scales to finally develop new knowledge for an increased societal benefit. These institutions comprise, e.g., the National Center for Ecological Analysis and Synthesis (NCEAS, Hackett et al. 2008) or the German Centre for Integrative Biodiversity Research (iDiv) Halle – Jena – Leipzig. They facilitate the overall progress in ecology, they improve our understanding of ecosystems and help to better track their state, both of which are essential prerequisites to approach challenges with a broad societal interest (Peters 2010). These challenges include, e.g., finding solutions to questions around climate

change, habitat loss and the declining diversity on Earth (Pereira et al. 2010). In that context, ecology also has grown into an essential source of information serving as input for decisions in policy, e.g., on how to best maintain services and values (e.g. food, water, fiber production) which are provided to us by our nature (United Nations 1992; Millenium Ecosystem Assessment 2010; Díaz et al. 2015; de Groot et al. 2012; Cardinale et al. 2012).

Specific introductions

Chapter One

Over the last decades, many data in ecology were deposited in disconnected data silos which were solely accessible by individual researchers or small groups. This has been found to be a problem as it significantly increases the probability of the data to get lost over time (P. Bryan Heidorn 2008). With the increasing awareness on the long-term value of ecological data, much effort was put into the development of documentation standards and into tools which are dedicated to the curation, preservation and discovery of data in ecology (Fegraus et al. 2005; Nadrowski et al. 2013). In parallel, policies have been developed and installed not only by governments but also by funding agencies and publishers (Penev et al. 2011). These policies were aiming for the regulation of data documentation, publishing and sharing in order to prevent their loss and maximise the reuse of valuable environmental data (European Science Foundation 2008; Vines et al. 2014).

However, the past has shown that researchers who adhere to the policies often publish data-products only. These products represent aggregates or subsets, which are derived from the collected raw research data. The original data tends to remain in private repositories (Savage and Vickers 2009; Fecher, Friesike, and Hebing 2015). However, not publishing the original research data along with its documentation has several downsides. It prevents the detection of errors in published articles and analyses, it is a barrier to their full reuse (subsets only allow limited analyses) and in turn, increases the chance of costly duplications in data collection efforts (Roche et al. 2015). Overall, the lack of sharing data and documentation represents a significant hurdle towards transparent and reproducible scientific findings. In other words, this impedes the central principles for sustainable progress in science (Tenopir et al. 2011; M. D. Wilkinson et al. 2016).

Chapter one deals with the issue of a growing gap in documentation between the data in private repositories, and the publication of derived knowledge and data

products in journals (Attwood et al. 2011). It discusses how to improve the documentation right at where the data are analysed. R was chosen here as a reference as it is the most widely used suite for statistical data analysis in ecology (R Development Core Team 2016). It has gained popularity over time due to its open-source licensing, its platform independence and its easy extensibility where a modular package system enables the latter. The open character of the R data analysis framework facilitates the implementation of new ideas and also enables quality checks by a large community of researchers (Touchon and Mccoy 2016). The R environment is typically used in an offline fashion in order to analyse data on a single personal computer or cluster while preparing it for publication. However, due to the flexibility of R, tight integration with online data repositories and services is possible as well. The integration of online resources, in fact, has become increasingly popular over the recent years, e.g., in the rOpenScience project as an important source of packages which enable access to online public accessible data sources (Boettiger et al. 2015).

Chapter one introduces the first open source contribution of this work; It is part of the rOpenScience project. The package is functioning as an interface between the R environment and the data management platform of the BEF-China project (Nadrowski et al. 2013, c.f. Methods). The package enables the bidirectional exchange of research data and its associated metadata and is the first R package to import the metadata from the Ecological Metadata Language standard (c.f. Methods). The standard is describing important aspects of data useful during the analysis workflow. Exemplary, the chapter is showing how the primary research data and its metadata can be pulled from an online data management platform into the R environment. It highlights how metadata can contribute to the understanding of the data and how this enables a more efficient processing and analysis. Further, the chapter shows how the results and the processing steps, which are related to the data used are uploaded back to the online platform. The upload establishes a

documentation circle linking the research idea and the original data (stored on the BEF-Data platform, c.f. to the Methods section) with the data products and the knowledge derived in the analysis. In other words, it helps to narrow the gap in documentation between the primary research data and derived data products and knowledge.

Chapter Two

Along with the increasing awareness of the long-term value of ecological data, it was proposed to adapt metadata in order to support the discovery and reuse of data in ecology (Fegraus et al. 2005; Wiczorek et al. 2012). Metadata does not only preserve human-readable documentation but can go far beyond this depending on its structure and the level of formalisation (Madin et al. 2007; Michener and Jones 2012). Metadata can be applied to many scenarios of usage. For example, it is used in the documentation of data which in turn can be used to better discover the data in databases by utilising a full-text search (Brin and Page 1998). This form of search is building an index based on the documentation in metadata associated with each item in the search pool (e.g., datasets). Keywords entered in a search box are then compared against the index in order to find matching results. This type of search, however, comes with several drawbacks. The problems are typically arising from the fact that a full-text search lacks a basic understanding of the semantic meaning of a search query (e.g., synonyms, homonyms). Thus a full-text search often yields unsatisfactory results (Beall 2008).

Several solutions have been developed over time to help compensate for the shortcomings of full-text search (e.g., English et al. 2002; Sy et al. 2012). These include the use of modelled knowledge (e.g., thesauri or ontologies) to complement a search query (e.g., adding broader, narrower or similar terms) or the classification of search items by the annotations with keywords. The latter finally allowed building mechanisms to enable an explicit selection of search items by their relevance (English et al. 2002; Yee et al. 2003). A crucial prerequisite for the

classification of search items is the annotation with a vocabulary. The purest form of vocabulary is a flat list of natural language terms or expressions (Trant 2009). These type of vocabularies are frequently built and used by social sharing, online communities along the classification and organisation of content (e.g., images, blog posts, papers, datasets). These vocabularies, however, have the problem that their most significant advantage unfolds along a tradeoff.

On the one hand, their flexibility and free nature allow the vocabulary to grow with the needs of its community, e.g., adding arbitrary keywords to sort their content (Trant 2009). However, on the other hand, this freedom often leads to redundancy or highly user-specific terms, which are hard to understand and reuse again by other users. Thus the vocabulary needs a curation mechanism to keep the classification clean and useful (Lamere 2008; Weller and Peters 2008). Another big downside of folksonomies is their lack of structure (e.g., no taxonomic hierarchies of the terms). Taken together, this limits their utility in information retrieval as their content is hard to access in another way than either with a word cloud to select from or a full-text search (Hotho et al. 2006).

Faceted navigation became popular over time as it is offering an intuitive and structured mechanism to select from search results (Hearst 2008). It lives from structured metadata and a multi-hierarchical classification of the search items (English et al. 2002). The keywords for the annotation come from a standardised and well-structured vocabulary like, e.g., a thesaurus or an ontology which are built by subject experts to best describe the searchable content (Salton 1980; Oren, Delbru, and Decker 2006). The system can complement a full-text search to help overcome part of its limitations. Thus it contributes to the efficiency of information retrieval (English et al. 2002). Facet navigation can provide a rich set of organised options to a user to select from during a search (Jones et al. 2006). The selection builds a filter pipeline limiting the search results to meet specific requirements in the end. In ecology, these requirements can include, e.g., the interest for data in

which variables have been explicitly measured or experimentally manipulated, data where the temporal and the spatial resolution of the measurements fall into a specific range or data which is coming from a particular biome or region on the earth (Pfaff et al. 2017). For example, searching for “Carbon” using a full-text search will bring up all search items associated with the chemical element as well as with “Carbon”, a village in Alberta, Canada. A filter here finally allows to better disentangle the ambiguity of terms. It can allow selecting according to attributes (i.e., for example, “location” and “chemical element” here) and substitute the need to manually browse, evaluate and decide on the relevance of separate results. This mechanism, in turn, is a prerequisite and first step towards a more efficient discovery of compatible data and finally the integration of the highly diverse data of ecology (Yee et al. 2003).

While implementing the mere mechanism of faceted navigation is straightforward the primary challenge remains in defining suitable attributes and vocabulary to appropriately capture the content and context of search objects while taking into account the needs and interests of the respective community of users (Hearst 2008; Strohmaier, Körner, and Kern 2012). In this context, in chapter one an open source framework for the annotation and faceted discovery of data in ecology has been created. It aims to support researchers in ecology to describe their data in a structured way while on the other hand supporting their interests related to information retrieval. The chapter is discussing the design principles as well as the needs of ecology as a discipline along information retrieval. In parallel the chapter introduces the second open-source contribution. It is a web-based tool with a graphical user interface aiding the structured annotation and discovery along the ideas in the presented framework (<https://github.com/cpfaff/ease>).

Chapter Three

Over the past few decades, ecological projects have grown in size and complexity in order to cover larger temporal and spatial scales while addressing a wider set of

topics (Peters et al. 2008). The projects are often involving the sharing of workload across many individuals from different fields of expertise and nationality (Borer et al. 2017). On top of this, ecology today also tends to set up large research sites which are used for an increasing amount of time (Bruelheide et al. 2014). The growth of project structure and size comes with an increased number of resources that need to be managed; which is typically the job of principal investigators and funding agencies. They have to provide guidance or conduct evaluations in order to measure the progress and success of a project. If the available resources are not recognised appropriately by the project members, they are remaining underutilised; this potentially has the consequence that it limits the overall value which can be gained from a large-scale ecological project.

The third chapter introduces a new use-case for the EASE framework (c.f. chapter two) in particular but opens new perspectives on the use of all the metadata from ecological projects in general. The chapter is discussing the growing complexity of ecological projects and associated problems. It develops ideas around the exploration of ecological projects using their metadata in a creative way, and the use of the metadata in order to support their evaluations. Further, it discusses how the metadata can increase the self-awareness of ecological projects and how it can be turned into instruments which allow informed decisions of principal investigators and funding agencies to take action which finally can improve the overall value, and ensure the success of a project.

General Material and Methods

In the following sections, details of the methodological aspects are provided which are relevant across all three chapters. A more detailed insight into the process of creating the vocabulary for the EASE annotation and discovery framework is covered here as well (c.f. chapter two). This description is of particular interest as two attempts have been made to create the vocabulary. The sections below are explaining some of the problems and the experiences which have been made during the process and how they finally shaped the development of the annotation framework.

GFBio project

The presented work took place in the context of the German Federation for Biological Data project (GFBio, Diepenbroek et al. 2014). This project has been funded by the German Research Foundation (DFG) starting in 2013. Currently, the project involves 19 partner institutions ranging across universities, natural history collections and libraries to bioinformatics and data archives for environmental data. It was set up with the goal to interconnect and build new data management solutions on existing cyber-infrastructure within Germany in order to provide researchers of biological and environmental sciences with services that are related to their data even beyond the lifetime of separate projects. The services cover the full life-cycle of data from the planning of new projects, the data acquisition, the description and the documentation of the data via metadata as well as the long-term preservation for potential data reuse. Finally, the project aims to be a central point of reference for all scientists dealing with environmental and biological research in Germany who receive their funds from DFG.

BEF-China project and BEF-Data

BEF-China is an international research project funded by the DFG (FOR 891). It has been formed to detangle influences of different aspects of plant diversity on functions and services of ecosystems such as primary production, erosion control,

and element cycling in the context of subtropical forest ecosystems (Bruehlheide et al. 2014). The project was established across two sites in the provinces Zhejiang and Jiangxi in southeastern China. It involves 147 researchers from China, Germany, and Switzerland. BEF-China consists of 16 groups, where two are responsible for the coordination of “Central Projects”, and the other 14 sub-projects are researching along a wide range of biological and ecological objectives.

The BEF-China project also developed its own data management platform which is called BEF-data. It allows the project to manage, document, share and curate all of its datasets (Nadrowski et al. 2013). BEF-Data provides a mechanism to initiate and guide new collaborations by allowing the project partners to request data from each other. The request for data in a so-called paper proposal needs to include all relevant information like a description of the new research idea for the data (Nadrowski et al. 2013). Further, these paper proposals serve as a single point of reference. They aggregate and collect information about the research idea, the involved authors, the datasets which are included and finally they are linking the products in the form of publications in journals (c.f. <https://bit.ly/2xdradr> and <https://bit.ly/2K1aFXj>). The metadata of all the datasets in the project is publicly available via the BEF-data portal as well part of the data which has been published already (<https://bit.ly/2QxP77c>).

The vocabulary creation

The development of the vocabulary for data annotation and discovery in ecology was started in parallel with chapter one. The primary goal was a vocabulary, which is close to the real needs of the ecological research community. Thus, keywords of ecological research projects have been collected, which they have created as annotation for their datasets. The collection comprised the BEF-China project (Bruehlheide et al. 2014), the Jena Experiment (Weisser et al. 2017), FUN-Div Europe (Baeten et al. 2013) and the Biodiversity Exploratories (Fischer et al. 2010). The final list of keywords was assembled as a simple flat list comprising a collection of

approximately 1200 atomic keywords (e.g., carbon) and short expressions (e.g., wood-inhabiting fungi). Consecutively, the vocabulary was cleaned and organised. That process included the removal of redundancy and the sorting of keywords into logical groups. These groups comprised, e.g., biological processes, manipulated and measured variables, organisms and biomes. Finally, the terms were organised more robustly along the ISO 2788 standard for thesauri. This standard defines a set of relationships between terms in a vocabulary which can be, e.g., broader, narrower, related or synonym.

Several workshops were organised inviting collaborators in dedication to help further organise the terms in the vocabulary. The workshop participants were asked to sort and organise parts or even the whole set of terms into logical groups and hierarchies. Additionally, they were asked to add new keywords if they felt that essential terms were missing (e.g., introduce new categories or links in the hierarchy). The participants in the workshops organised the given terms in many different ways. The outcome varied from structures that followed simple hierarchical forms up to developing own theories for a better organisation along adding many categories and concepts. All of them finally had in common that they were based on and backed by the personal experience and scientific background of the workshop participants. It turned out that the terms do not fall into a self-evident logical structure. Thus, a classification, similar across the working groups could not be achieved based on the terms themselves.

The structure of the terms has been mainly determined based on individual points of view and the interpretation of each term as the outcome of discussions in the groups. The organisation and structures often needed further explanation to make the location of specific terms in the vocabulary understandable for another person not involved in the very same workshop. The reason for the variability in the organisation of the terms is likely to be found in lexical ambiguity and the lack of context information (e.g., definitions of the terms). If no documentation is available

for a term, then it is hard to know what it meant in the first place (here in projects for annotations). Many places for a term are reasonable in the organisational structure, which all can be backed by according argumentations. This is a universal problem with language and thus likely holds for all types of vocabulary being developed ranging from rather simple thesauri up to complex ontologies.

Initially, it was decided to construct the vocabulary along with a bottom-up approach. This decision was attractive in the perspective of achieving the goal to stay as close as possible to the keywords and real-world language samples from the databases. Thus the vocabulary development started from the most specific terms developing towards more general ones (e.g., Carbon -> Chemical Element -> Thing). Due to the many possible options of structuring the terms, it became finally apparent that it might be better to have a new take and reverse the direction of development. In a top-down approach, the vocabulary was then developed from general terms growing into more detail (e.g., Thing -> Process -> Oxidation -> Nitrification). This approach has the advantage that it allows developing the structures in the vocabulary more strategically, e.g., along the lines of what the vocabulary needs in order to best serve ecological research projects or in particular specific use cases like data retrieval. This approach finally allowed to create a solid fundament on which the rest of the vocabulary could be based on.

The first set of top-level terms and structures were inspired by experiences made with the organisation of extracted keywords from the databases and various discussions in the workshops with colleagues. An initial workshop for the top-down vocabulary was held which involved collaborators from GFBio in order to find an agreement for the most critical top-level categories suitable for a description of data in ecology. Eight categories of topics have been selected finally which then served as “initial nucleus” for the further development of the vocabulary towards a framework which was finally named Essential Annotation Schema for Ecology (EASE, c.f. Chapter two).

Chapter One

Title:

rBEFdata: documenting data exchange and analysis for a collaborative data management platform

Journal:

Ecology and Evolution

Access:

<https://doi.org/10.1002/ece3.1547>

<https://bit.ly/2Dzim7W>

Authors:

Claas-Thido Pfaff¹ (corresponding: claas-thido.pfaff@uni-leipzig.de), Birgitta König-Ries³, Anne C. Lang¹, Sophia Ratcliffe¹, Christian Wirth^{1,2}, Xingxing Man¹, Karin Nadrowski¹

Affiliations:

Universität Leipzig: Institut für Spezielle Botanik und Funktionelle Biodiversität¹, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig², Friedrich-Schiller-Universität Jena: Department of Mathematics and Computer Science³

rBEFdata: documenting data exchange and analysis for a collaborative data management platform

Claas-Thido Pfaff¹, Birgitta König-Ries², Anne C. Lang¹, Sophia Ratcliffe¹, Christian Wirth¹, Xingxing Man¹ & Karin Nadrowski¹

¹Department of Special Botany and Functional Biodiversity, University of Leipzig, Johannisallee 21, 04103 Leipzig, Germany

²Department of Mathematics and Computer Science, Friedrich Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

Keywords

Biodiversity ecosystem functioning, data postprocessing, data sharing, ecological metadata language, metadata, open science, open source, paper proposals, R, reproducible science, semantic integration, workflow.

Correspondence

Claas-Thido Pfaff, Department of Special Botany and Functional Biodiversity, University of Leipzig, Johannisallee 21, 04103 Leipzig, Germany.

Tel: +49-341-97-38587;

Fax: +49-341-97-38549;

E-mail: claas-thido.pfaff@uni-leipzig.de

Funding Information

This project has been funded via the Chinese-European research unit Biodiversity and Ecosystem Functioning China (BEF China, FOR 891) and the German Federation for the Curation of Biological Data (GFBio) project which both are funded by the German Research Foundation (DFG).

Received: 23 January 2015; Revised: 28 April 2015; Accepted: 1 May 2015

Ecology and Evolution 2015, 5(14): 2890–2897

doi: 10.1002/ece3.1547

Introduction

Large amounts of ecological data are gathered each year by researchers worldwide, striving to enhance the knowledge on our ecosystems. Many data management platforms and tools have been developed with the growing awareness on the value of data, the need of data curation, and the potential of data reuse (BEFdata, Nadrowski et al. 2013; Bexis, Lotz et al. 2012; Metacat, Berkley et al. 2001; DataOne, KNB). Data platforms and networks facilitate the access to heterogeneous data typically generated by

Abstract

We are witnessing a growing gap separating primary research data from derived data products presented as knowledge in publications. Although journals today more often require the underlying data products used to derive the results as a prerequisite for a publication, the important link to the primary data is lost. However, documenting the postprocessing steps of data linking, the primary data with derived data products has the potential to increase the accuracy and the reproducibility of scientific findings significantly. Here, we introduce the rBEFdata R package as companion to the collaborative data management platform BEFdata. The R package provides programmatic access to features of the platform. It allows to search for data and integrates the search with external thesauri to improve the data discovery. It allows to download and import data and metadata into R for analysis. A batched download is available as well which works along a paper proposal mechanism implemented by BEFdata. This feature of BEFdata allows to group primary data and metadata and streamlines discussions and collaborations revolving around a certain research idea. The upload functionality of the R package in combination with the paper proposal mechanism of the portal allows to attach derived data products and scripts directly from R, thus addressing major aspects of documenting data postprocessing. We present the core features of the rBEFdata R package along an ecological analysis example and further discuss the potential of postprocessing documentation for data, linking primary data with derived data products and knowledge.

ecological research projects. This is reflected in the variety of study systems, methods, data types, environmental contexts, and the temporal and spatial scales. A specific reuse of data is the fusion of many datasets in meta-analyses. This is of particular interest in ecology as it potentially allows quantitative summaries of research domains to generate higher-order conclusions about general trends and patterns (Arnqvist & Wooster 1995, Koricheva et al. 2013). Conclusions derived from analyzing data are then archived as papers in journals. Although an increasing number of journals today require the data used to derive

2890

© 2015 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

the results as prerequisite for publication (e.g., f1000), the steps on how these data have been assembled from primary data and how the data have been processed during the analysis are often hidden. Losing the link between primary data, derived data products, and knowledge results in a “gulf” between primary data repositories and knowledge repositories (Shotton 2009; Attwood *et al.* 2010).

Linking primary data with a research idea in a proposal is a sociocultural mechanism applied by an increasing number of research consortia (e.g., Stokstad 2011). While this potentially helps to prevent duplication and to maximize synergies in projects, it lacks a persistent and structured documentation. The BEFdata (Biodiversity and Ecosystem Functioning Data) platform provides an implementation of this mechanism as a step-by-step wizard guiding through the creation of a proposal (Nadrowski *et al.* 2013). This mechanism does not only streamline the discussion and collaborations in research projects but also serves for documentation and thus as a single point of reference for information revolving around a research idea. It has the potential to link primary data, the research idea in the form of a rationale, and the derived data products in one place. Documenting the postprocessing steps of primary data towards a final data product has the potential to reduce redundant efforts in projects and increase the accuracy and the reproducibility of scientific findings. However, we need an easy way to add information to a paper proposal right at where the data are used, for example, in statistical scripting environments. We here present the core features of the r-Biodiversity and Ecosystem Functioning Data R package (rBEFdata) along an ecological analysis example highlighting the use of the proposal mechanism for documentation purposes. We further discuss the potential of data postprocessing documentation linking primary data with derived data products and knowledge to help building a bridge over the “gulf” between primary data and knowledge repositories.

The BEFdata Portal

BEFdata is a data management web application written in Ruby on Rails (Nadrowski *et al.* 2013). It is open source software released under an MIT license. The development of the platform is driven by the joint Chinese–German–Swiss research project “BEF-China” (FOR 891). The source code as well as information on how to set up the application is provided via the Github repository and the associated wiki pages (<https://github.com/befdata/befdata>). The BEFdata portal is not comparable to large data archives (e.g., Gbif, DataOne) but is instead meant to be used by research collaborations during a period in which the data are not yet ready for being deposited in large international archives. The portal gives researchers in a project the

opportunity to centrally store, clean, harmonize, and share their data in a private and secure environment. It uses the Ecological Metadata Language standard (EML, Fegraus *et al.* 2005) to format metadata, and it facilitates collaboration using a paper proposal mechanism (Nadrowski *et al.* 2013). Data hosted on instances of BEFdata are private by default, but the metadata is readable by everybody. A researcher can fill a shopping cart with data needed to answer certain research questions. After the selection of the data, a proposal has to be written to inform the data owners. This represents a mechanism that goes far beyond a simple data request by providing a transparent way of communicating the scope and rationale of a planned analysis to all researchers, listing the datasets required for the analysis as well as the potential co-authors. The analysis can start after all participants involved in the proposal agreed on a modus to work on toward a publication. For a more detailed description about the portal functionality, we refer to Nadrowski *et al.* 2013.

Data for the Examples

The BEFdata portal is currently used by two research consortia, the BEF-China (Bruehlheide *et al.* 2014) and the FunDivEUROPE project (<http://www.fundiveurope.eu>). For our examples, we use data provided by the BEF-China project consortium. It is an international research collaboration with the aim to disentangle the role of tree and shrub diversity for production, erosion control, element cycling, and species conservation in Chinese subtropical forest ecosystems (Bruehlheide *et al.* 2014). The research data of the BEF-China experiment are hosted on a BEFdata server available via <http://china.befdata.biow.uni-leipzig.de>. The data for our example analysis have been compiled in a paper proposal revolving around questions about nitrogen acquisition and retention in the study plots of the BEF-China experiment (Lang *et al.* 2014). The data have been collected by different subprojects of the BEF-China experiment and are publicly available via the paper proposal depicted in Figure 1. The vocabulary used in the data discovery example is a work in progress thesaurus created for the domain of Biodiversity and Ecosystem Functioning. It is developed in the context of the project called the German Federation for Biological data (GFBio) and can be accessed via <http://tematres.befdata.biow.uni-leipzig.de/vocab/>.

The rBEFdata Package

The package is developed for the R statistical programming environment (R Development Core Team 2015), and is part of the rOpenSci (<http://ropensci.org/>) package portfolio. It is released under an MIT license and a stable

Mixed afforestation of young subtropical trees promote nitrogen acquisition and retention

Created at: 2013-09-16

Envisaged Journal: Journal of Applied Ecology

Envisaged date: 2013-08-16

Rationale

Knowledge of biodiversity of very limited, particularly as approaches using tree sapling acquisition and cycling in early successional stages of secondary forests and forest plantations. Insights in the potential of nutrient retention of young tree plantations are of particular interest in China, where large areas have been reforested in order to counteract soil erosion and to increase the soils' water and nutrient retention capacity. In this study we planted saplings of four abundant early successional (evergreen and deciduous) tree species in monocultures, two- and four-species combination to test the effect of species richness on nitrogen acquisition and retention by using a ^{15}N tracer experiment.

A crucial question in BEF research is the appropriate time scale of experiments which allows species richness effects to emerge. This question gains importance when long-lived and slowly growing

systems is still Experimental identifier nutrient

Proposal Metadata

Author data owners

Author

Anne Lang

Main Proponents

Werner Härdtle, Prof.

Michael Scherer-Lorenzen, Prof.

Figure 1. The page of the proposal we used in our example in this paper. Proposals serve as starting point for analyzing data hosted on BEFdata platforms. They can serve as central point of coordination and for documentation. They aggregate information such as the title, the author, and the date of creation as well as the envisaged journal and a detailed rationale. After an analysis, derived data products and scripts can be attached to complement this information. The proposal page that is shown here is available under the url <http://china.befdata.biow.uni-leipzig.de/paperproposals/90> and the analysis has been published by Lang et al. 2014.

version is available via CRAN. We here introduce the rBEFdata R package and its core features available with version 0.3.5. rBEFdata provides programmatic access to features of the BEFdata data management platform. It allows to search for data and integrates with external thesauri hosted via a TemaTres vocabulary server (<http://www.vocabularyserver.com>) to improve the data discovery. rBEFdata allows to download and import data and metadata into R hosted on a BEFdata instance. A batched download is available as well which works along the paper proposal mechanism of BEFdata. This fetches all data associated with a specific research idea in one single step for analysis. Upload functionality of the R package allows to push new data to the BEFdata portal and to attach derived data products right from within R to their original paper proposal.

rBEFdata has a set of options that, for example, determine which instance of BEFdata is used. Issuing the command `bef.options()` provides an overview about the options available and their current values (in the following, all R commands embedded into the text appear in italics). Most of the option fields come with predefined

values (related to BEF-China). This can be simply changed by the assignment of a new value. For example, the URL to an own BEFdata server can be set by `bef.options("url" = "http://my.own.befdata.instance.com")`. This allows one to use the R package with one or several different instances of BEFdata. Other options are dedicated to the URL of a TemaTres vocabulary server and the name of a download folder. This folder is created to store downloads, for example, attachments of paper proposals like R scripts or images. For a detailed overview about all available commands, we refer to the package manual, which is available here <http://cran.r-project.org/web/packages/rbefdata/rbefdata.pdf>.

Data Discovery and Thesaurus Integration

Datasets can be tagged with keywords on BEFdata portals. These keywords can then be used for data discovery from R issuing the command `bef.portal.get.dataset.for_keyword("carbon")`. Here in this example, we search for all data tagged with "carbon". Multiple keywords can be used in

combination with concatenation, for example `c("carbon", "nitrogen")`. The command returns a data frame with the titles and the IDs of the data found. The IDs can then be used to access the data with `bef.portal.get.dataset_by(id = xx)`. In a second step, we improve the search along an external thesaurus using the TemaTres server integration. If we search the BEF-China instance of BEFdata for "plant organ" and "weight" with `bef.portal.get.dataset_for_keyword(c("plant organ", "weight"))`, which returns 24 datasets, this misses data that are tagged with narrower terms for plant organs. We access the external thesaurus to improve the search terms used towards more detailed terms. We receive a list of narrower terms issuing the command `bef.tematres(task = "fetchDown", term = "plant organ")`. The list returned from the vocabulary server then includes "leaf", "root", "twig", "seed" and "stem". Repeating the search from the above with the extended list not only yields more-than-twice the number of datasets, but also results in the data tagged with the narrower terms (57). Exchanging the external thesaurus via the options command of rBEFdata allows to use an own or reuse existing vocabularies of preference like e.g. the LTER thesaurus.

Data Access

The data of other researchers are only accessible if they made it public or when the access has been granted through a paper proposal. From rBEFdata, the access is controlled via credentials in which a registered user can find on its profile page on a BEFdata instance. The credentials can be set via the command `bef.options("user_credentials" = "xyz")`. After the approval of a proposal on BEFdata, it is possible to batch download all requested datasets. The batch download requires the ID of the proposal, which can be found in its URL. The proposal we use in our example contains three datasets and it has the ID 90 (Fig. 1). We import the datasets using the command `bef.portal.get.datasets.for_proposal(id = 90)` (Fig. 2). We need two of the datasets for our example

analysis and we assign these datasets to different variables. We call them "n_retention" and the other "design". A look into the column headers of these both datasets reveals a common column header named "plot_id". If we want to use that column to merge the two datasets into a synthesis product, we need to ensure that they contain the same categories and have the same meaning. As rBEFdata provides access to the EML formatted metadata, we have access not only to the title, the abstract but also to the column descriptions of each dataset we just downloaded. We can access the metadata using the R builtin `attributes()` command (e.g., `attributes(n_retention)`). We can check whether the two datasets can be merged by the "plot_id" column inspecting the column descriptions from the metadata. For example, for the dataset we just assigned to the variable "design," we can inspect the column description issuing the command `attributes(design)$columns[1,]$description`. The description from the first dataset can then be compared with the description of the "plot_id" column in the second dataset (Fig. 3).

Analysis and Postprocessing Documentation

After the import of the data, the data preparation, and the merging of the data into a synthesis data product, we can start with our analysis (Lang et al. 2014; fig. 3). In our example analysis, we create a plot figure representing our results (Figs. 3, S1). We attach the R script containing the analysis and the plot figure we created to the proposal for documentation (Fig. 4). We can attach to proposals using the command `bef.portal.attach_to_proposal(id = xx, attachment = "path/to/file", description = "desc")`. We need to provide the command with the ID of the proposal we want to attach to, a path to a file we want to attach, and a description. Documenting the postprocessing steps in the form of a script and derived data products like a plot figure together with the paper proposal preserves valuable information (Fig. 5).

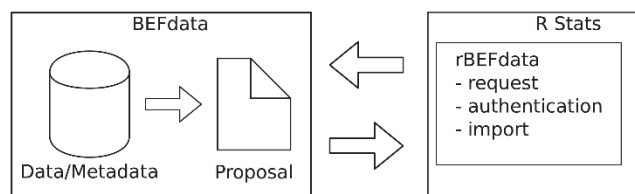


Figure 2. In our example workflow with rBEFdata, we first download and import all data that have been attached to the proposal depicted in Fig. 1. You can check out the detailed R code via a Gist we made available on GitHub (<https://gist.github.com/cpfaff/2a927c772342fe398466>). Furthermore, the script containing all code of our examples in one file is available via figshare <http://dx.doi.org/10.6084/m9.figshare.1365364> and Gist from here <https://gist.github.com/cpfaff/c7dcfa1c971ee61150b2>.

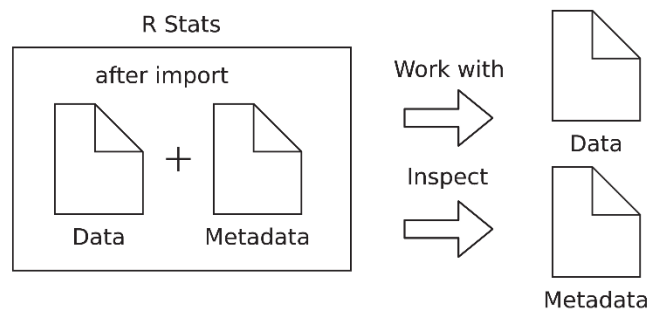


Figure 3. After the import of the data from the BEFdata platform, we can start working with it right away. Furthermore, we are able to inspect the metadata as it is attached as attributes to the R data frame. In our example, we first inspect the metadata of the datasets. This can help in understanding data and guide decisions on merging data into synthesis data products. The R script covering the steps of inspecting the metadata is available via Gists (<https://gist.github.com/cpfaff/f0ac88924b5cbde48949>). The scripts covering the analysis parts and the visualization can be found here <https://gist.github.com/cpfaff/3dc94da4f6191fedaad1> and here <https://gist.github.com/cpfaff/63ecba903b4b4b8a4783> (Hothorn *et al.* 2008, Weisberg 2011, Pinheiro *et al.* 2014). Furthermore, the script containing all code of our examples in one file is available via figshare <http://dx.doi.org/10.6084/m9.figshare.1365364> and Gist from here <https://gist.github.com/cpfaff/c7dca1c971ee61150b2>.

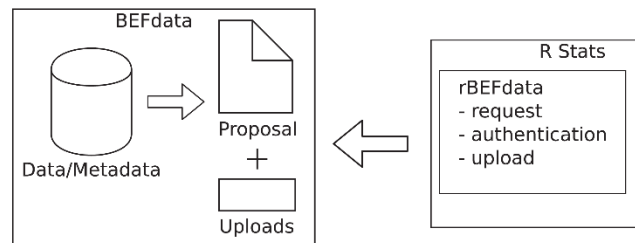


Figure 4. After the analysis has been performed, derived datasets, images, as well as the R script itself can be uploaded and directly attached to the original paper proposal. In our example, we upload a plot in PNG format, depicted in Figure 5, and the R script as attachment to the proposal which is shown in Figure S1. Uploading the R script used for analysis allows to keep a whole provenance record linking the primary data with derived products. The R code covering the steps of data upload we provide a Gist on GitHub <https://gist.github.com/cpfaff/37c131c7a6b903db2f00>. Furthermore, the script containing all code of our examples in one file is available via figshare <http://dx.doi.org/10.6084/m9.figshare.1365364> and Gist from here <https://gist.github.com/cpfaff/c7dca1c971ee61150b2>.

Discussion

As part of the rOpenSci community (<http://ropensci.org/>), we aim to make scientific data available from within R that is typically spread across many databases and networks around the world (e.g., DataOne, KNB, GBIF). With rBEFdata, we provide one piece to the puzzle offering convenient access to data and metadata hosted on instances of the BEFdata data management platform (Nadrowski *et al.* 2013). The core functionality of the rBEFdata package is quite similar to what other packages offer which allow to access scientific data from within R (e.g., rDataOne, rGBIF). By the time of writing, to the

best of our knowledge, the rBEFdata package was the first R package persisting EML structured metadata over the import of data into an analysis workflow with R. Furthermore, the support and the access to metadata from R is not yet widely spread (e.g., see rDataOne). However, we currently work on a more generic framework for R that supports read and write of metadata as well as import and export of data based on the EML metadata standard (<https://github.com/ropensci/EML>). This framework can then be used to provide other R packages with the functionality and the benefits that come along with having access to metadata. Metadata has a wide range of applications during scientific analysis workflows. We highlighted how it

Selected Datasets

Datasets (3)

Competition of saplings for N -Pilot- 15N recovery in leaves and fine roots (main)

Competition of saplings for N -Pilot- system 15N retention (main)

Plottreatment and -location within the blocks of the Pilot-Experiment (main)

Calculated Authors

Anne Lang, Helge Bruelheide, Prof., Werner Härdtle, Prof., Keping Ma, Prof., Michael Scherer-Lorenzen, Prof., Stefan Trötschel, Bo Yang, Goddert von Ohelmb, Martin Baruffol, Sabine Both, Matteo Brezzi, Martin Böhnke, Karin Nadrowski, Juliana Nates Jimenez, Ricarda Pohl, Yann Salmon

Attached from R

Files (2)

rbeffdata.Rmd (26.4 KB)
The R script that has been used to derive the results in the published paper

results_plot_proposal_90.png (10.2 KB)
Nitrogen (N) retention affected by species richness. N retention summed as the recovery of soil, roots and leaves (a), relative leaf recovery (b), relative root recovery (c) and relative soil recovery (d). Significant differences as revealed by post hoc Tukey's test ($P < 0.05$) are indicated by different letters.

Data request state

Preparation > Project Board > Data Requests > **Finished**

Data request approved. Download rights will expire on Wed 16 Sep 2015.

Helge Bruelheide, Prof.

Keping Ma, Prof.

Michael Scherer-Lorenzen, Prof.

Project

SP02e Individual plant growth and branch demography as a function of species richness and composition

Email lists

Author and proponents

Author, proponents and owners of main datasets

Author, proponents and all dataset owners

Author, proponents, dataset owners and data helpers

Figure 5. The proposal page shows the primary data put together to answer a certain research question. After completing the analysis, derived data products can be attached in with no limitation to a specific format. In our example, we attach a plot figure and an R script that has been used to derive the results. As we show, it is possible to add a description to the attachment which allows to provide more information. The description for the plot figure we attached here provides similar information like a figure caption on how to interpret the results.

can be used to guide the decisions on manually merging primary data into synthesis products. Metadata also represents a first step toward automatic data integration. If the metadata, for example, includes unit information, this can be used to level out granularity differences making the creation of synthesis data products easier. However, a full automatic assemblage and integration of data would need further support of higher-order semantics in the form of ontologies (Michener and Jones 2012).

Although thesauri may be of limited use as direct help in full automatic data integration, they provide important guidance toward that step. They can be employed to structure harmonization efforts within and across projects, providing mechanisms to agree on concepts and definitions which helps with speaking a common language. A controlled vocabulary applied over data can simplify manual data integration and potentially help in linking data to ontologies (Michener and Jones 2012). Furthermore, thesauri have the potential to be used for data discovery as we have shown in our example. With rBEFdata, we start to provide this important kind of functionality based on TemaTres, as the thesaurus of the Long-Term Ecological Research network (LTER) is based

on this technology (Porter *et al.* 2011). Reusing existing vocabulary is attractive as this can save more effort. However, reuse is not always an option if there is no vocabulary available that suits the project. As TemaTres is open source provided under a GPL license, this allows any project to set up an own instance for the development of a project-specific vocabulary (<http://www.vocabularyserver.com/>). The flexibility to switch between vocabularies in rBEFdata allows projects to create and use a vocabulary on their own or just decide to reuse existing vocabularies of their choice. Searching BEFdata from R, however, currently has some limitations: As data on BEFdata instances are private by default, there is no way to use the respective data right away except it has been made public by the data owner. According to the workflow defined in BEFdata, a proposal has to be written to gain access to the data, but there is currently no functionality implemented that would allow to generate a paper proposal from R. The creation of proposals from R, however, is planned for future releases of rBEFdata which then allows to fully exploit the potential of the search mechanism.

Scientific workflow tools like Kepler or Pegasus have a long tradition in documenting the postprocessing of data

(Altintas et al. 2004; Oinn et al. 2004; Deelman et al. 2009). Using the attachment functionality of rBEFdata with the paper proposal mechanism of BEFdata provides a simple way to use documentation mechanism for the postprocessing of data from within R. An extensive use of the attachment functionality significantly increases the value of paper proposals as they can become a single point of reference holding all information necessary to understand and reproduce a scientific analysis. Our documentation mechanism, however, currently is very application-specific to BEFdata. Thus, we are working on combining our documentation mechanism with features based on ideas of reproducible reporting in R (Yihui Xie 2014). BEFdata will then provide export functionality of paper proposals holding all their information which could be published along with primary data and the paper. Although journals today more often require the data products used to derive the results as a prerequisite for publication, the link to the primary data is lost. Without this information, however there is no way to control for underlying errors that might occur on data preparation steps, which typically represent about 70% of the whole scientific analysis workflow (Garijo et al. 2014). Providing the primary data and the postprocessing information together with the publication would allow to control for underlying problems which is important as improperly designed, or incorrectly analyzed experimental data can lead to incorrect conclusions (Tilman 1989). Publishing primary data and postprocessing information along with a paper has the potential to improve the reliability and reproducibility of scientific findings.

Conclusion

In creating the rBEFdata R package as a companion to the open source data management portal BEFdata (Nadrowski et al. 2013), we provide a convenient tool to communicate with BEFdata servers directly from R. The open source licensing of BEFdata and its companion R package rBEFdata allow research projects to run their own sophisticated data management, curation, analysis, and documentation setup. Combining the R package with the data management platform can significantly improve the data analysis workflow, the productivity, and collaboration in any ecological project while promoting best practices in the data management and reproducibility through good documentation.

Acknowledgments

Thanks goes to all colleagues from our working group of Special Botany and Functional Biodiversity at the University of Leipzig for their continuous feedback on the

manuscript. We also like to thank the members of the BEF-China project for providing access to their data for the examples. We also like to acknowledge the funding of the DFG through the projects GFBio and BEF-China (FOR 891).

Conflict of Interest

None declared.

References

- Altintas, I., C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock 2004. Kepler: an extensible system for design and execution of scientific workflows. *16th International Conference on Scientific and Statistical Database Management, 2004. Proceedings*, pp. 423–424.
- Attwood, T. K., D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne. 2010. Utopia documents: linking scholarly literature with research data. *Bioinformatics* 26: i568–i574.
- Arnqvist, G., and D. Wooster. 1995. Meta-Analysis: synthesizing research findings in ecology and evolution. *Trends in Ecology & Evolution* 10:236–40.
- Berkley, C., M. Jones, J. Bojilova, and D. Higgins 2001. Metacat: a schema-independent XML database system. *Thirteenth International Conference on Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings*, pp. 171–179.
- Bruehlheide, H., K. Nadrowski, T. Assmann, J. Bauhus, S. Both, F. Buscot, et al. 2014. Designing forest biodiversity experiments: general considerations illustrated by a new large experiment in subtropical China. *Methods Ecol. Evol.* 5:74–89.
- Deelman, E., D. Gannon, M. Shields, and I. Taylor. 2009. Workflows and e-Science: an overview of workflow system features and capabilities. *Fut. Gen. Comput. Syst.* 25:528–540.
- Fegraus, E. H., S. Andelman, M. B. Jones, and M. Schildhauer. 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* 86:158–168.
- Fox, J., and S. Weisberg 2011. *An R Companion to Applied Regression*, Second. Sage, Thousand Oaks, CA. Available at <http://socserv.socsci.mcmaster.ca/~jfox/Books/Companion>
- Garijo, D., P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. 2014. Common motifs in scientific workflows: an empirical analysis. *Fut. Gen. Comput. Syst.* 36:338–351.
- Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. *Biom. J.* 50:346–363.
- Koricheva, J., J. Gurevitch, and K. Mengersen. 2013. *Handbook of meta-analysis in ecology and evolution*. Princeton University Press, <http://press.princeton.edu/titles/10045.html>
- Lang, A. C., G. von Oheimb, M. Scherer-Lorenzen, B. Yang, S. Trogisch, H. Bruehlheide, et al. 2014. Mixed afforestation of

- young subtropical trees promotes nitrogen acquisition and retention. *J. Appl. Ecol.* 51:224–233.
- Lotz, T., J. Nieschulze, J. Bendix, M. Dobbermann, and B. König-Ries. 2012. Diverse or uniform? — Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecol. Inform.* 8:10–19.
- Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol. Evol.* 27:85–93.
- Nadrowski, K., S. Ratcliffe, G. Bönisch, H. Bruehlheide, J. Kattge, X. Liu, et al. 2013. Harmonizing, annotating and sharing data in biodiversity–ecosystem functioning research. *Methods Ecol. Evol.* 4:201–205.
- Oinn, T., M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, et al. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20:3045–3054.
- Pinheiro, J., D. Bates, S. DebRoy, and D. Sarkar & Team, R.C. 2014. nlme: Linear and Nonlinear Mixed Effects Models. Available at <http://CRAN.R-project.org/package=nlme>
- Porter, J., M. O. Brien, D. Costa, D. Henshaw, C. Gries, E. Mendez, et al. 2011. A controlled vocabulary for LTER data keywords.
- R Core Team 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Shotton, D. 2009. Semantic publishing: the coming revolution in scientific journal publishing. *Learn. Publ.* 22:85–94.
- Stokstad, E. 2011. Open-source ecology takes root across the world. *Science* 334:308–309.
- Tilman, D. 1989. In *Likens*. Pp. 136–157 in E. Gene, ed. Long-term studies in ecology. Springer, New York, NY. <http://link.springer.com/10.1007/978-1-4615-7358-6>.
- Xie, Y. 2014. knitr: A general-purpose package for dynamic report generation in R. R package version 1.6.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. The plot we created to visualize the results in our analysis example. The code covering the steps to create the pot figure is available from here <https://gist.github.com/cpfaff/63ecba903b4b4b8a4783>. For a detailed explanation of the results see (Lang et al. 2014).

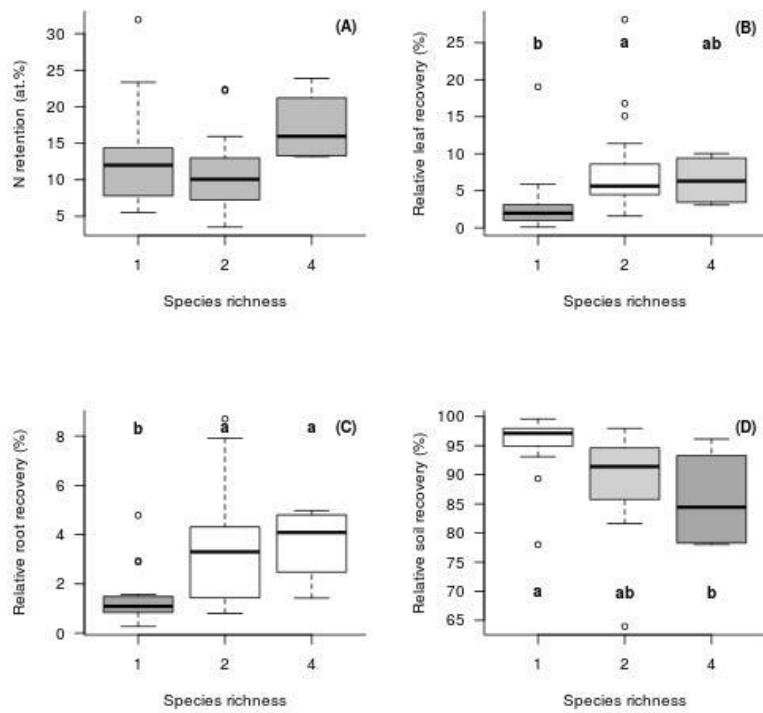


Figure 3 The boxplot visualises the results from our analysis example used in the first chapter. The code to produce the figure is published here <https://gist.github.com/cpfaff/63ecba903b4b4b8a4783>. For a detailed explanation of the ecological analysis and the results see (Lang et al. 2014).

Chapter Two

Title:

Essential Annotation schema for Ecology (EASE) – A framework supporting the efficient data annotation and faceted navigation in ecology.

Journal:

PLOS One

Access:

<https://doi.org/10.1371/journal.pone.0186170>

<https://bit.ly/2R7jxP0>

Authors:

Claas-Thido Pfaff¹ (corresponding: claas-thido.pfaff@uni-leipzig.de), David-Eichenberg², Mario Liebergesell², Birgitta König-Ries³, Christian Wirth^{1,2}

Affiliations:

Universität Leipzig: Institut für Spezielle Botanik und Funktionelle Biodiversität¹, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig², Friedrich-Schiller-Universität Jena: Department of Mathematics and Computer Science³, Martin Luther University Halle-Wittenberg: Institute of Biology/Geobotany and Botanical Garden⁴

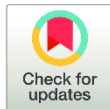
RESEARCH ARTICLE

Essential Annotation Schema for Ecology (EASE)—A framework supporting the efficient data annotation and faceted navigation in ecology

Claas-Thido Pfaff^{1*}, David Eichenberg¹, Mario Liebergesell², Birgitta König-Ries³, Christian Wirth^{1,2}

1 Department of Special Botany and Functional Biodiversity, University of Leipzig, Germany, **2** German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany, **3** Department of Mathematics and Computer Science, Friedrich Schiller University of Jena, Germany

* claas-thido.pfaff@uni-leipzig.de



 OPEN ACCESS

Citation: Pfaff C-T, Eichenberg D, Liebergesell M, König-Ries B, Wirth C (2017) Essential Annotation Schema for Ecology (EASE)—A framework supporting the efficient data annotation and faceted navigation in ecology. *PLoS ONE* 12(10): e0186170. <https://doi.org/10.1371/journal.pone.0186170>

Editor: Junwen Wang, Mayo Clinic Arizona, UNITED STATES

Received: February 15, 2017

Accepted: September 26, 2017

Published: October 12, 2017

Copyright: © 2017 Pfaff et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data is available via the public github repository at <https://github.com/cpfaff/ease> and <https://github.com/cpfaff/EaseAnnotationTool>

Funding: Deutsche Forschungsgemeinschaft, GZ: WI 2045/1 1-2, (DFG) <http://gepris.dfg.de/gepris/projekt/229241684> The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Ecology has become a data intensive science over the last decades which often relies on the reuse of data in cross-experimental analyses. However, finding data which qualifies for the reuse in a specific context can be challenging. It requires good quality metadata and annotations as well as efficient search strategies. To date, full text search (often on the metadata only) is the most widely used search strategy although it is known to be inaccurate. Faceted navigation is providing a filter mechanism which is based on fine granular metadata, categorizing search objects along numeric and categorical parameters relevant for their discovery. Selecting from these parameters during a full text search creates a system of filters which allows to refine and improve the results towards more relevance. We developed a framework for the efficient annotation and faceted navigation in ecology. It consists of an XML schema for storing the annotation of search objects and is accompanied by a vocabulary focused on ecology to support the annotation process. The framework consolidates ideas which originate from widely accepted metadata standards, textbooks, scientific literature, and vocabularies as well as from expert knowledge contributed by researchers from ecology and adjacent disciplines.

Introduction

Technological progress is driving the efficient acquisition, the dissemination and the reuse of data in ecology. Today data is created at an increasing pace and large research networks are used to provide access to ecological data for a broad audience [1,2]. With an improved access to a wide range of ecological data many potential benefits arise. It can help to reduce the amount of redundant data acquisition efforts or facilitate the formation of new collaborations. The reuse of data in fact has become one of the most important strategies in contemporary ecological synthesis projects (e.g. NCEAS: [3,4]). It is not only a basis for reproducible science

Competing interests: The authors have declared that no competing interests exist.

but also a precondition for synthesizing knowledge. Data reuse allows to extend the scope of studies in order to cover wider temporal and spatial scales which are relevant to human societies and which help to generalize theory across environmental contexts. For example, meta-analyses reusing data from scattered experiments have allowed to develop the theory of multifunctionality in biodiversity/ecosystem functioning research [5], and extended functional biodiversity research from plots to continents [6] and the parameterization of global climate models [7]. Although the reuse of data is important for research in ecology it can be challenging to find suitable data which qualifies for the reuse in a specific context and which comes with context data necessary for an integration into meta-analyses.

The essential prerequisites for an optimal reuse of ecological data are detailed metadata, annotations and efficient search strategies [8,9]. Metadata standards which are used in the context of ecology cover a broad range of information (e.g. [8]). They deal with topics like the temporal and spatial extent of data, or the organisms and methods covered by a study [8,10]. This information is typically provided in large detail using full text descriptions which then can serve as a basis for a full text search. A full text search basically matches strings which are given in a search box with strings that are contained in data and metadata (e.g. abstracts, method descriptions, variable names). A full text search, however, comes with several idiosyncrasies which are reducing its effectiveness. For example, it is typically not aware of synonyms or homonyms nor does it account for broader, narrower or closely related terms relevant for a specific search term. On top of that a full text search lacks the understanding of the semantic meaning of a search query and thus often fails to provide satisfactory results [11,12]. As an example: Searching for the keyword “Carbon” using a full text search across an ecological database will potentially yield a host of results. This might include results from global change studies using elevated *carbon* dioxide concentrations as experimental treatment, soil survey reporting *carbon* concentrations in the subsoil, paleoclimate studies employing *carbon* isotope discrimination in tree rings, or field observations near *Carbon Village* in Alberta, Canada.

Faceted navigation is a mechanism which is frequently used in combination with full text search as it allows a refinement of the search to improve the results. As prerequisite a faceted navigation requires the search objects (e.g. datasets, pictures or products) to be classified in categories. This classification can be done along an arbitrary amount of categories. However, the categories are often reflecting the inherent characteristics of the search objects. In e-commerce that means for example the price, the type or the brand of a product whereas in ecology, for example, the name of the study regions, authors or information related to time and date are potential categories. The categorization is typically stored as an annotation which in turn is stored as sidecar file with the search object [13]. A selection from the categories during a search can then incrementally build up a filter which is restricting the results to match the selected criteria. With respect to the example mentioned above: If the full text search on “Carbon” is complemented by a faceted navigation based on a classification which is using “experimental treatment” and “variable name” as categories then the information can be used to separate the results selecting one or the other explicitly depending on the requirements (e.g. picking “experimental treatment = elevated *carbon* dioxide concentration”).

While the general mechanism of a faceted navigation is simple, the main challenge remains in defining useful parameters and vocabulary which are relevant for the classification of the search object to allow for an efficient discovery [14]. In the context of the German Federation for Biological Data (GFBio, <http://www.gfbio.org>) and in close collaboration with the German Center for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig we set out to define requirements for an efficient annotation of ecological data optimized for a faceted navigation discovery. We screened various sources of information like metadata standards, textbooks,

scientific literature and vocabularies to search for useful patterns and concepts suitable for an annotation of ecological data.

Here we present a framework that we call the Essential Annotation Schema for Ecology (EASE) consisting of two parts. The first part is an annotation schema which is based on XML Schema Definition (XSD). It allows to store the information about the classification of search objects along several categories serving as a basis for a faceted annotation and navigation application. The XML schema is accompanied by a vocabulary with a focus on ecology which provides support for the annotation through the provision of ecologically relevant conceptual keywords. The framework is a synthesis which consolidates ideas that originate from expert knowledge, widely accepted metadata standards, and ecological theories and concepts (e.g. used to structure content in textbooks), scientific literature and standardized vocabularies. In the following we present the framework and the underlying design principles and provide an outlook towards a tool based on the framework supporting time efficient annotations and the faceted navigation for an improved discovery and reuse of ecological data.

Project context

GFBio has the goal to bundle available cyber infrastructure in Germany in order to support researchers in biology and ecology along the whole life cycle of data. GFBio thus aims at supporting the planning of new projects, the acquisition and analysis of data, the publication process, the curation of data and metadata as well as the long term storage of data. Finally, the GFBio web portal will serve as a central point of reference in Germany for the access to biological data including advanced search and features to foster the reuse of biological data and the collaboration between researchers. In order to support the development of the EASE framework several (10 in total) workshops have been set up in close collaboration with the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. Domain experts from ecology and adjacent disciplines have been invited to contribute their ideas formulating general design principles for the framework and to discuss and drive the development of the vocabulary.

Design principles

As a first step, design principles have been defined to set up the general guidelines for the development of the EASE framework.

Parsimony

In order to support a time efficient annotation, the framework should be kept as simple as possible in regards of structure and the content. This optimization, however, should be done carefully by still maintaining a differentiated and consistent description of ecological data. An example: Time represents an important aspect in ecology which is typically covered by calendar dates and times. Larger time frames are covered by numerical references (e.g. 18 Mio years ago) or by named geological time periods. The International Chronostratigraphic Chart (ICC) is an effort which aims to define the geological time frames of earth history. It defines eons (5 in total: Phanerozoic, Precambrian, Proterozoic, Archean, and Hadean), eras (10 in total: e.g. Cenozoic, Mesozoic, Paleozoic), periods (22 in total: e.g. Quaternary, Neogene, Paleogene), epochs (34 in total: e.g. Holocene, Pleistocene, Pliocene) and ages (98 in total: e.g. Calabrian, Gelasian, Piacenzian). The time frames are getting more granular from eons to ages and the fine granular time frames are nested in the larger ones. For simplicity of the framework and the annotation process it could be argued to ignore e.g. "ages" or at least make them optional.

While this would sacrifice some granularity, it would simplify the annotation and still provide a consistent classification depicting the larger temporal context.

Comprehensiveness

Despite the fact that the framework is striving for parsimony it also has the goal to achieve comprehensiveness. EASE aims at defining essential orthogonal dimensions according to which ecological content can be precisely described. Comprehensiveness is not accomplished by using many different, but rather a few and strictly complementary dimensions. This is reflected by using broad domain relevant topics which are covered in the annotation schema (e.g. time: start time and end time, space: name of locations, method: general approach of the study) but also by the quality how the topics are covered in detail. As an example: Understanding processes and mechanisms is an important aspect to many ecological studies. Thus, the annotation schema contains a part dealing with ecological processes. The processes are covered in a certain breadth asking not only for the name of the process itself but also for related aspects like the objects which are involved (e.g. Organisms, Chemical, Matter, and Energy) and for a generic characterization of the process (e.g. Uptake, Release, and Exchange). The vocabulary is providing a list of widely used and well defined ecological processes which supports the annotation process providing suggested content for the process name field in the schema. As the number of processes used in ecology is potentially endless a list has been designed covering widely used and well defined generic processes e.g. demography (i.e. death, birth, growth), disturbances (e.g. windstorm, fire) or interactions (e.g. parasitism, mutualism).

The framework

Vocabulary

Several workshops were carried out comprising in total 35 researchers from ecology and adjacent disciplines. Top level categories for the framework have been collected and eight categories were finally selected. These top level categories represent orthogonal dimensions of information in the search space relevant in ecology (e.g. time, space, methods). In the workshops the selected top level categories have been substantiated in a top-down approach defining a vocabulary with increasing detail. Additional material such as textbooks [15–17] and standardized vocabularies (e.g. World Reference Base for Soil Resources: <http://www.fao.org/soils-portal/soil-survey/soil-classification/world-reference-base/en/>, International Chronostratigraphic Chart: <http://www.stratigraphy.org/index.php/ics-chart-timescale>) have been reviewed in order to find useful conceptual keywords and patterns for the annotation framework. The vocabulary of the framework is detailed below along the selected top level categories. The complete framework is available on GitHub (<https://git.io/v1Vty>) and the sections below are containing references to the according parts of the vocabulary hosted online.

- Time

This is the facet of EASE which captures temporal aspects relevant for ecology. It includes the start and the end of a data acquisition, geological time frames as well as the temporal resolution and extent of the study. The dates and times in EASE are conform to ISO8601 and names of time zones follow the IANA time zone database (<http://www.iana.org/time-zones>). The geological time frames refer to those given in the International Chronostratigraphic Chart (ICC) which defines and names time ranges in order to express the time scale of earth history (<http://www.stratigraphy.org/index.php/ics-chart-timescale>). For the temporal extent and the temporal granularity, the vocabulary contains categories along common units of time e.g. “Second”, “Minute”, “Hour”, and “Day” (c.f. vocabulary <https://git.io/v1Vtd>). In a faceted

discovery that ultimately allows to select for data which is matching a desired temporal resolution. For example, studies interested in a fine seasonal resolution typically search for data carried out over at least a whole year with measurements taken on a daily or hourly basis (e.g. atmospheric temperature measurements).

- Space

The space facet of the EASE framework deals with information related to localities and regions. It captures the names of locations, the location type as well as the hierarchical relation of a location to countries and continents. For the location type as well as for the countries and the continents the EASE vocabulary provides predefined lists. They are containing e.g. “City”, “Stream”, and “Lake” (c.f. vocabulary <https://git.io/v1sA1>) for location types or names of countries and continents like “Andorra”, “Afghanistan”, “Africa”, “Asia” and “Europe” (c.f. vocabulary <https://git.io/v1sAS>) which has been incorporated from the GeoNames ontology (<http://www.geonames.org/>). In addition to such explicit definitions of locations, the EASE framework allows to specify a bounding box as well as the exact study site coordinates. The bounding box provides a coarse localization using decimal degree values. The coordinates are captured using the Universal Transverse Mercator (UTM) and the World Geodetic System 1984 (WGS84) datum. Similar as in the time facet the space facet provides a resolution and an extent. To this end the vocabulary provides predefined categorical values being “Point” (<1 m²), “Plot” (1 m²–0.01 km²), “Region” (0.01 km²–10000 km²), “Continent” (10000 km²–100000000 km²) and “Global” (larger) (c.f. vocabulary <https://git.io/v1Vtj>). This allows to filter for data which comes with the desired spatial resolution and extent. For example, data that has been gathered at the landscape scale (exceeding 10 km²) but within which several localized study plots were established where measurements have been taken.

- Sphere

The sphere part comprises aspects of the pedosphere, the hydrosphere, the atmosphere and the lithosphere. It complements the spatial information of the EASE framework covered in the location facet by identifying compartments and vertical layers within ecosystems or larger spatial reference units. For example, it allows to specify a distinct layer within the atmosphere (e.g. Troposphere, c.f. vocabulary <https://git.io/v1OUU>) or a layer within a body of water (e.g. Abyssopelagic, c.f. vocabulary <https://git.io/v1OUI>) to state where the data has been gathered. Apart from this, the sphere facet also captures the levels of biological organization. For that purpose the vocabulary provides predefined categories ranging from the “Atom” over “Cell” and “Organ” up to the “Biosphere” (c.f. vocabulary <https://git.io/v1Of7>). This finer level of granularity in faceting allows in the end for the selection of data which focuses on a specific organizational level or which comes from a specific compartment in the biosphere like a certain layer in the atmosphere or the soil. Fig 1 shows an example how the annotation could look like with a potential user interface. Based on the definitions given in the vocabulary, annotation (and search) can be achieved by ticking the matching category provided by the tool.

- Biome

The biome facet of EASE captures relevant aspects which describe biomes. This comprises the latitudinal (e.g. Boreal, Temperate, Tropic, c.f. vocabulary: <https://git.io/v1OU4>) and altitudinal zonation (e.g. Nivale, Montane, c.f. vocabulary: <https://git.io/v1OUE>), the moisture regime (e.g. Humid, Arid, c.f. vocabulary: <https://git.io/v1OUi>), the continentality (e.g. Continental, Maritime, c.f. vocabulary: <https://git.io/v1OUX>) as well as the physiognomy of the biome (e.g. Savannah, Shrubland, c.f. vocabulary: <https://git.io/v1OUD>) [18]. Below these higher levels of information, the EASE framework also extends into more specifics which are

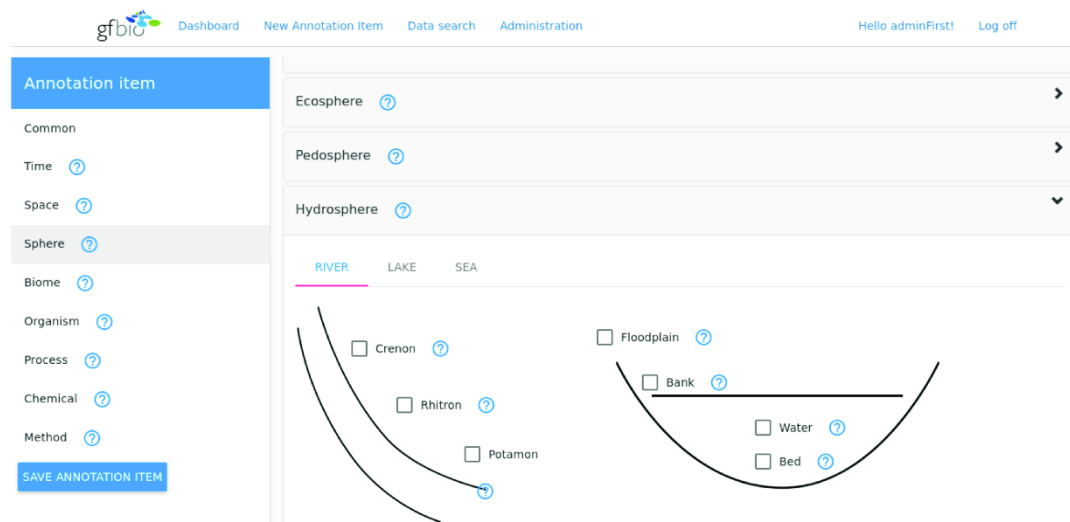


Fig 1. A mockup of a potential annotation tool which will be developed in the future based on the principles of the EASE framework. The figure here specifically depicts the sphere part, detailing the sub-facet hydrosphere. To allow for a finer granularity within the hydrosphere, the tool will allow to specify specific zones in and related to rivers, lakes or the sea. Within these sub-facets, one can easily state where measurements or samples have been taken. To guide the user and allow for a qualitative annotation, definitions of the respective concepts should be displayed e.g. by resting with the cursor over the question marks. In case the user does not find suitable concepts in a specific category he will be provided with an option to extend the annotation vocabulary on the fly (i.e. not shown here).

<https://doi.org/10.1371/journal.pone.0186170.g001>

dealing with oro- and pedobiomes, as well as elevation and edaphic features. The vocabulary provides conceptual keywords for selection which are containing e.g. “Amphibiome”, “Halo-biome” or “Helobiome”(c.f. vocabulary: <https://git.io/v1Ofj>). The biome part also deals with the classification of biomes comprising their general condition with “Natural” or “Urban” and their dominant form of usage with e.g. “Agriculture”, “Forestry” or “Fishery” (c.f. vocabulary: <https://git.io/v1OvN>). It is important to note that many of these features are difficult to infer from the location alone because the fine-scale heterogeneity of hydrography, soil types, physiognomy and land-use is not appropriately resolved in digital maps.

- Organism

The organism facet of EASE deals with the scientific names and taxonomy of organisms. The schema captures scientific names separately for botanical, zoological, fungal organisms and for viruses). For the taxonomy of organisms, the schema of EASE is containing elements named along the main ranks of the Linnean topology which are “Domain”, “Kingdom” (e.g. Plantae, Animalia), “Division” (botany) or “Phylum” (zoology), “Class”, “Order”, “Family” and “Genus”.

- Process

The process facet deals with relevant aspects of ecological processes. To this end the vocabulary supports the annotation by providing a generic list of ecological processes which comprises e.g. the “Adaption”, “Speciation” and “Migration” (c.f. vocabulary: <https://git.io/v1OfZ>). Additionally the process part deals with interactions, where the user is presented with the

option to specify the interacting partners based on kingdoms (e.g. "Plantae", "Animalia"), the direction of the interaction ("Mutual", "Affects", "Is Affected By") and the quality of the interaction (e.g. "Amensalism", "Antagonism" c.f. vocabulary: <https://git.io/v1OfE>). Not only does this allow to select a particular process in the end but also to carry out a search for interaction process related datasets in a very generic way. For example, one can select all data that deals with the interaction between fungi and plants where the direction from the first to the second interaction partner is specified as "Affects" with the quality being "Antagonistic". That in the end would select data dealing with fungi as plant parasites but not as symbionts (see Fig 2).

- Chemical

The chemical facet deals with all aspects of chemistry being part of ecological data. This comprises chemical elements and compounds which have been measured as well their function in the biological context. The vocabulary here supports the annotation by providing a list of elements based on the periodic table as well as a list of chemical compounds and classes of compounds e.g. "Lipids", "Carbohydrates", "Amino Acids" (c.f. vocabulary: <https://git.io/v1OfT>) which has been compiled from various sources [16,17,19]. Moreover, the biological functions of chemicals which are relevant in ecological studies are covered by conceptual keywords like e.g. "Antibody", "Attractant" or "Repellent" (c.f. vocabulary: <https://git.io/v1OfY>) which has been inspired by parts from the Chemical Entities of Biological Interest ontology (CHEBI) (<http://www.ebi.ac.uk/ols/ontologies/chebi>).

- Method

The methodological facet of the EASE framework captures the general approach and the context of the study. The vocabulary provides a list of generic approach types being either "Virtual" (e.g. simulation), "Manipulative" (i.e. with experimental factors mostly controlled) or "Observational" (i.e. where plot selection creates factor gradients) (<https://git.io/v1OfK>). The context of the study approach is captured by categories like "Microcosm" (e.g. lab experiment), "Mesocosm" (e.g. ecotron, greenhouse experiment) to "Macrocosm" (e.g. field studies) (<https://git.io/v1Ofi>). On top of that the method part of EASE captures the variables that either have been manipulated in a study. The vocabulary provides a list of aspects which are manipulated frequently to form gradients containing conceptual keywords like e.g. the "Producer diversity", the "Consumer density" or the "Nutrient availability" (<https://git.io/v1OfD>).

Schema

In parallel to the development of the vocabulary detailed above the EASE XML Schema has been created to serve as foundation for an annotation and faceted navigation application. It is built using the XML Schema Definition (XSD) standard. In order to discover structures suitable for reuse in the annotation schema we screened three XML based metadata standards which are frequently used in the context of ecology (see also S1–S5 Tables). These were:

1. Darwin Core (version: 2015-06-05) which is a standardized metadata schema maintained by the members of the Biodiversity Information Standards (TDWG). It started as a loose collection of terms with a clear semantic meaning. The focus of DwC is to capture and to exchange detailed information about organisms. Darwin Core is separated into nine main topics, six of which are dealing with information like the acquisition event of data, locations, and the geological context, the occurrence of organisms and their taxonomy and the authority of identification. The other parts of the schema deal with general context information which comprises the names and addresses of institutions as well as the nature of the data record [10].

Fig 2. A mockup of a potential annotation tool which will be developed in the future based on the principles of the EASE framework. The figure here specifically depicts the interaction part of processes. It allows to specify the interaction name, the partners, the direction and the quality of the interaction. For the free input fields like the name of the interaction here in this part of the annotation tool auto completion functionality will be provided. This allows to pick from suggestions during the annotation which come from the EASE vocabulary. If a user however is not able to find the right conceptual keyword the vocabulary could be extended creating a new term as required and adding it to the list of annotation terms to be reused by others.

<https://doi.org/10.1371/journal.pone.0186170.g002>

2. The Ecological Metadata Language standard (EML, version 2.0.0) is developed and maintained by the Knowledge Network for Biocomplexity (KNB). It is an initiative with the goal to provide a sophisticated metadata standard for ecology. It has a modular and flexible design which allows using specific parts while neglecting others depending on the use case. It has four top level modules which represent resources that can be described. This comprises dataset, literature, software and protocol. The schema defines a host of modules which allow to capture detailed information about the resources (e.g. Access Rights, Physical Aspects: e.g. File format; Related Parties: e.g. associated people and organizations; Time and Organism related aspects: e.g. Time frame, Taxonomy) [8].
3. The Access to Biological Collection Data (version 2.06) is a metadata standard for the access and the exchange of data about specimens in collections and observations. It is used by the Global Biodiversity Information Facility (GBIF) and the Biological Collection Access Service for Europe network (BioCAsE: [20]). The schema is strongly hierarchically organized capturing e.g. aspects about biotopes, specimen, data acquisition events and contacts (e.g. authors, institutes) as well as a detailed history about the location of physical collection objects (<https://github.com/tdwg/abcd>).

All of the schemas equally well cover aspects of time and space as well as methods and organisms which are essential for a description of data in ecology (see also S1–S5 Tables). The EASE schema provides a well-organized structure for an efficient annotation in ecology which is revolving around the eight facets of the vocabulary detailed above. Apart from that the schema it also defines elements which store general information like responsible parties (e.g.

contact and author names and addresses), a reference to the hosting data center, the title and the abstract of the search object and information about how to access the data (e.g. URL, file path, database id). The schema has been designed with an application in mind which is supporting the future maintenance and growth of the vocabulary. Thus the schema allows to store new conceptual keywords not only including their scientific definition but with their associated Unique Resource Identifier (URI) which also provides a link to external vocabularies like ontologies or thesauri [9].

Discussion

Metadata which is associated with ecological data today is often utilized to support full text search [9]. Although full text search has seen some improvements over time it comes with several inherent issues which often lead to unsatisfactory search results [11]. Faceted navigation is a strategy which gained much popularity over the last decade and by today is successfully applied in a multitude of applications ranging from e-commerce to science [13]. While the basic principle of facets is simple the main challenge remains in the design of the classification attributes [13]. They require a careful design adapted to the specific use case and in order to reflect not only the bare characteristics of a resource but also the requirements of the searching user. The existing metadata schemata that we reviewed for the design of the schema were already covering many aspects we needed in fine detail which have been reused in the structure of the EASE schema (e.g. time and date from EML [8] and organism related aspects of ABCD) but many other detailed aspects have been developed during the workshops based on discussions revolving around particular user needs (e.g. simple temporal and spatial extent and resolution of data or detailed interactions). Next to appropriate attributes which capture information about the search object a vocabulary which is supporting the annotation is equally important.

There are basically two opposing strategies for the provision of a vocabulary. The first follows a top-down approach, where the developer of the annotation schema creates a fixed hierarchy and finite list of terms. The advantage of this approach is that the resulting vocabulary does clearly focus on the essential dimensions and terms. However, top-down designed vocabulary is likely to be incomplete compared to real user requirements. The second strategy is a bottom-up approach like it is known from social tagging [21]. There users are allowed to freely tag their resources (e.g. pictures, datasets). The resulting pool of keywords forms an unstructured vocabulary which is called a folksonomy [22]. This strategy can be very powerful. It is easy to use even without any prior knowledge about a specific vocabulary or annotations and the vocabulary can flexibly grow to reflect the interests and the needs of a user community. However, maturing folksonomy are likely to inflate quickly accumulating redundancy e.g. in form of synonyms, spelling mistakes and different language terms referring to the same semantic concept and they are also likely to contain highly personalized tags which are hard to understand and reuse for others [23].

With the EASE framework we set out to strike a balance between the methods mentioned above. In the creation of the ecological annotation vocabulary we started with a top down approach which is based on a multitude of standards, textbooks and expert knowledge. In the schema we do stick to the top down approach forcing the user to pick from a limited set of vocabulary options for many of the annotation attributes (content restricted attributes). This is especially true where frequent changes of vocabulary are unlikely (e.g. time zones, countries, continents) or where the vocabulary reflects a finite and use case specific gradient (e.g. temporal and spatial resolution). However, there are other parts in the schema which are more open and basically follow a combined approach. There, some vocabulary is provided as an option to

pick from but they are not exclusively restricted to these terms which allows the vocabulary to grow (e.g. names of processes, the chemical compounds and the names of variables used as gradients in a study). However, the growth of the vocabulary in these elements should not be uncontrolled. An application on top of the schema should subject new vocabulary to a curation process which (i) 'harvests' the emerging new concepts and (ii) and allows a curator to incorporate them in their original or a modified form into the backbone of the EASE vocabulary in order to prevent the problems we see arise with folksonomies.

In the near future we aim to develop an application based on the EASE framework and the mockups we have shown (GitHub: <https://git.io/v5wWe>). It will provide features which allow for the efficient and fast annotation of data in ecology. It will come with an auto completion so it is possible to pick from meaningful suggestions during the annotation. If a user should not be able to find an appropriate term for the annotation, the tool will help to create vocabulary on the fly and then subject the new created concepts to a curation process. The application will provide support for the annotation of data in a single and batch mode and allow to create annotation templates which then can be applied to any amount of data to speed up the annotation process which is e.g. useful with data coming from the same project (some aspects are not changing). The tool will also integrate with a set of carefully selected external services to provide further vocabulary resources e.g. to fuel the suggestion mechanism beyond the EASE basic vocabulary (e.g. the GFBio terminology service <https://www.gfbio.org/data/annotateandconnect>). Here it is important to note again that the EASE annotation schema allows storing the URIs of terms used in an annotation. This enables a path to all the content and the knowledge which is modelled in external vocabularies and it allows to link resources described via EASE with many other resource even if they have not been described with EASE. For example, when we pick an environment from the ENVO ontology (e.g. soil) during the annotation in EASE and store the URI this allows us to query and compare all resources which use terms from ENVO for the annotation no matter of the annotation format (e.g. search for datasets which contain soil related parameters).

With the EASE framework we provide a basis for a detailed and highly organized annotation of ecological data which allows to situate data in the ecological search space. The framework can serve as a starting point for new projects and can help them to maintain a harmonized vocabulary facilitating data discovery with a faceted navigation. At the moment the EASE vocabulary is a simple controlled vocabulary. However, the combination of the schema, the vocabulary and the future application together provide a potential platform which allows communities of ecologists to produce and agree on a useful folksonomy which later on can be harvested as raw material for the creation of more elaborate ontologies [24]. Our framework is highly compliant with the topics that are covered by widely used metadata standards in ecology. Thus it is straight forward and easy to ingest information about resources already described via metadata in form of EML, ABCD, or DwC (S1–S5 Tables). The extendibility of the framework can potentially provide new insights increasing the knowledge in metadata sciences and allow a fine granular control over the yield of results combined with a full text search for a better discovery of data in ecological databases.

Supporting information

S1 Table. It shows the conceptual topics of time in EASE in relation to how the topics are covered in EML, ABCD and DwC metadata standards (X = not explicitly available as element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented e.g. using XSLT transformations. (DOCX)

S2 Table. It shows the conceptual topics for space in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented e.g. using XSLT transformations. (DOCX)

S3 Table. It shows the conceptual topics for biomes in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented e.g. using XSLT transformations. (DOCX)

S4 Table. It shows the conceptual topics for organisms in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented e.g. using XSLT transformations. (DOCX)

S5 Table. It shows the conceptual topics for methods in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented e.g. using XSLT transformations. (DOCX)

Acknowledgments

We like to thank all the participants of the workshops for their contributions: Alex Klein, Alexander Weinhold, Alexandra Weigelt, Alfred Radl, Anne Lang, Anton Güntsch, Bettina Ohse, Björn Reu, Claudia Guimaraes-Steinicke, Daniel Marra, David Fichtmüller, Harald Vacik, Hongmei Chen, Ingo Schöning, Ivaylo Kostadinov, Jens Kattge Karin Nadrowski, Katharina Grosser, Katherina Pietsch, Kristin Baber, Maren Gleisberg, Marion Schruppf, Martin Freiberg, Naouel Karam, Nico Eisenhauer, Nicole Van Dam, Pelin Yilmaz, Pier Luigi Buttigieg, Robert Huber, Robert Tolksdorf, Rolf Engelmann, Ronny Richter, Sophia Ratcliffe, Steffen Bode. We also like to thank all the people from the institute *Spezielle Botanik und Funktionelle Biodiversität* at the University of Leipzig for their continuous help in form of fruitful discussions about the vocabulary and proof reading of the manuscript.

Author Contributions

Conceptualization: Claas-Thido Pfaff, David Eichenberg, Mario Liebergesell, Christian Wirth.

Funding acquisition: Christian Wirth.

Investigation: Claas-Thido Pfaff.

Methodology: Christian Wirth.

Supervision: David Eichenberg, Birgitta König-Ries, Christian Wirth.

Visualization: Claas-Thido Pfaff.

Writing – original draft: Claas-Thido Pfaff, Christian Wirth.

Writing – review & editing: Claas-Thido Pfaff, David Eichenberg, Mario Liebergesell, Birgitta König-Ries, Christian Wirth.

References

1. Hey T. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Res; 2009.
2. Kattge J, Díaz S, Lavorel S, Prentice IC, Leadley P, Bönsch G, et al. TRY—a global database of plant traits. *Glob Change Biol*. 2011; 17: 2905–2935. <https://doi.org/10.1111/j.1365-2486.2011.02451.x>
3. Sala OE, Chapin FS, Iii, Armesto JJ, Berlow E, Bloomfield J, et al. Global Biodiversity Scenarios for the Year 2100. *Science*. 2000; 287: 1770–1774. <https://doi.org/10.1126/science.287.5459.1770> PMID: 10710299
4. Goodale CL, Apps MJ, Birdsey RA, Field CB, Heath LS, Houghton RA, et al. Forest Carbon Sinks in the Northern Hemisphere. *Ecol Appl*. 2002; 12: 891–899. <https://doi.org/10.2307/3060997>
5. Reich PB, Tilman D, Isbell F, Mueller K, Hobbie SE, Flynn DFB, et al. Impacts of Biodiversity Loss Escalate Through Time as Redundancy Fades. *Science*. 2012; 336: 589–592. <https://doi.org/10.1126/science.1217909> PMID: 22556253
6. Ratcliffe S, Liebergesell M, Ruiz-Benito P, Madrigal González J, Muñoz Castañeda JM, Kändler G, et al. Modes of functional biodiversity control on tree productivity across the European continent. *Glob Ecol Biogeogr*. 2015; 25: 251–262. <https://doi.org/10.1111/geb.12406>
7. Brovkin V, Raddatz T, Reick CH, Claussen M, Gayler V. Global biogeophysical interactions between forest and climate. *Geophys Res Lett*. 2009; 36: L07405. <https://doi.org/10.1029/2009GL037543>
8. Feigraus EH, Andelman S, Jones MB, Schilchauer M. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bull Ecol Soc Am*. 2005; 86: 158–168. [https://doi.org/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)
9. Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol Evol*. 2012; 27: 85–93. <https://doi.org/10.1016/j.tree.2011.11.016> PMID: 22240191
10. Wiczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE*. 2012; 7: e29715. <https://doi.org/10.1371/journal.pone.0029715> PMID: 22238640
11. Beall J. The Weaknesses of Full-Text Searching. *J Acad Librariansh*. 2008; 34: 438–444. <https://doi.org/10.1016/j.acalib.2008.06.007>
12. Blair DC, Maron ME. An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. *Commun ACM*. 1985; 28: 289–299. <https://doi.org/10.1145/3166.3197>
13. English J, Hearst M, Sinha R, Swearingen K, Yee K-P. Flexible Search and Navigation Using Faceted Metadata. 2002. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.8556>
14. Hearst MA. UIs for faceted navigation: Recent advances and remaining open problems. *HCIIR 2008: Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval*. 2008. pp. 13–17. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.3770&rep=rep1&type=pdf#page=13>
15. Schaefer M. *Wörterbuch der Ökologie [Internet]*. Heidelberg: Spektrum Akademischer Verlag; 2012. Available: <http://link.springer.com/10.1007/978-3-8274-2562-1>
16. Müller-Esterl W. *Biochemie: Eine Einführung für Mediziner und Naturwissenschaftler*. 1. Aufl. 2004. Korr. Nachdruck 2009. München; Heidelberg: Spektrum Akademischer Verlag; 2004.
17. Vollhardt KPC, Schore NE, Peter K. *Organische Chemie*. 4. vollst. überarb. u. aktualis. Auflage. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2005.
18. Woodward FI, Lomas MR, Kelly CK. Global climate and the distribution of plant biomes. *Philos Trans R Soc B Biol Sci*. 2004; 359: 1465–1476. <https://doi.org/10.1098/rstb.2004.1525> PMID: 15519965
19. Riedel E, Janiak C. *Anorganische Chemie*. 6th ed. Berlin; New York: de Gruyter; 2007.
20. Holeschek J, Dröge G, Güntsch A, Berendsohn WG. The ABCD of primary biodiversity data access. *Plant Biosyst—Int J Deal Asp Plant Biol*. 2012; 146: 771–779. <https://doi.org/10.1080/11263504.2012.740085>
21. Lamere P. Social Tagging and Music Information Retrieval. *J New Music Res*. 2008; 37: 101–114. <https://doi.org/10.1080/09298210802479284>

22. Strohmaier M, Körner C, Kern R. Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. *Web Semant Online*. 2012; 17: 1–11. <https://doi.org/10.1016/j.websem.2012.09.003> PMID: 23471473
23. Trant J. Studying Social Tagging and Folksonomy: A Review and Framework. *J Digit Inf*. 2009; 10. Available: <https://journals.tdl.org/jodi/index.php/jodi/article/view/269>
24. Dotsika F. Uniting formal and informal descriptive power: Reconciling ontologies with folksonomies. *Int J Inf Manag*. 2009; 29: 407–415. <https://doi.org/10.1016/j.ijinfomgt.2009.02.002>

Chapter Three

Title:

On the evaluation of ecological projects using their metadata

Authors:

Claas-Thido Pfaff¹ (corresponding: claas-thido.pfaff@uni-leipzig.de), Helge
Bruelheide^{2,4}, David-Eichenberg², Birgitta König-Ries³, Christian Wirth^{1,2}

Affiliations:

Universität Leipzig: Institut für Spezielle Botanik und Funktionelle Biodiversität¹,
German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig²,
Friedrich-Schiller-Universität Jena: Distributed Information Systems³, Martin
Luther University Halle-Wittenberg: Institute of Biology/Geobotany and Botanical
Garden⁴

Abstract

Today ecological projects have grown into highly complex endeavours along with the global demand for a deeper understanding of the Earth's ecosystems and the related services. Ecological projects often rely on large collaborations to bridge the expertise across disciplines and set up spatially extent research platforms used as the basis for data collection over extended periods of time. While the growing complexity of project structures allows for better insights into the systems studied, they also involve increasing challenges for principal investigators and funding agencies as they have to provide guidance or evaluate the progress and success of a project. Thus, we here we want to suggest to make use of metadata of ecological projects to allow gaining a better insight into the project. We exemplify the use of metadata describing selected aspects across a decade of research carried out in the BEF-China project. For the description of the data, we mainly used the Essential Annotation Schema for Ecology (EASE) and the companion data annotation tool in order to create the metadata. We show and discuss how metadata of ecological

projects can provide useful insights into the project. These insights comprise the collaboration structures or how topics emerge and evolve along time, and we discuss how the metadata can help to inform principal investigators and funding agencies during an evaluation or how it can serve as feedback for members in a project to better exhaust existing resources.

Introduction

In response to the increasing public demand for solutions to acute problems of global relevance (e.g., rapid climate change, species loss, ecosystem degradation, Cardinale et al. 2012) ecological projects have grown in size and complexity. In order to approach the scales of conservation and land management, ecological projects today frequently form large collaborations and create spatially extent study platforms used as long-term observatories (Hobbie et al. 2009; Weigelt et al. 2010; Fischer et al. 2010; Bruelheide et al. 2014). These large setups ensure a collection and analysis of compatible data while comprehensively characterising the study system and enables the continuous integration of the existing data with new emerging research ideas. While extensive and more elaborate projects have the potential to vastly improve our understanding of ecological systems they also come with their very own challenges (Borgman, Wallis, and Enyedy 2007). For example, it is getting harder to keep an overview of the resources in a project comprising collected samples and data analyses, topics that are covered, collaborations formed or the projects which have been planned, etc. However, if its members do not recognise the resources of a project, they cannot fully exploit them (e.g., reuse existing data, or increase synergy with another researcher). This lack in turn potentially limits the overall value that can be gained from a project. A detailed overview about the resources that are available in a project is crucial for its overall success. It is of interest not only for each researcher but also for principal investigators and funding agencies which both are in charge of providing guidance and evaluating the progress and success of a project.

In ecology, the long-term value of data has been recognised early. It was proposed that appropriate descriptive metadata has the potential to save much of the value of a study for future generations of research (Fegraus et al. 2005; Michener and Jones 2012). Thus, in the last decades, several tools have been created. They enable data management including the description of data using standardised and well-structured metadata schemata (Higgins, Berkley, and Jones 2002; Nadrowski et al. 2013; Berkley et al. 2001). Apart from the long-term preservation aspect, metadata can support other functions as well. These functions may include the exchange and discovery of data as well as it can enable a better understanding of the content and the context of data which finally allows for better analysis (Michener and Jones 2012). In the context of data analyses, the metadata is particularly interesting. It allows for efficient processing and the integration of the data. It provides context (e.g., sampling methods) and content (e.g., the meaning of variables) related information. The information can capture and describe similarities (e.g., methods, variables) as well as subtle differences between datasets which need to be levelled out before an integration (Fegraus et al. 2005; Pfaff et al. 2017).

Scientometrics is a scientific discipline which uses quantitative methods along the goal to study and understand patterns, dynamics and trends which appear in various scientific disciplines (Hood and Wilson 2001; Garfield 2009). Scientometric analyses typically include the productivity (e.g., count of publications), the collaboration (Otte and Rousseau 2002; Hou, Kretschmer, and Liu 2008) and the impact achieved ranging from single individuals up to whole discipline (Hirsch 2005) or an overview about the historical development of topics (Pollack and Adler 2015). The analyses typically leverage publication metadata such as those collected by databases like Scopus, Science Citation Index or Web of Science (Bar-Ilan 2008). The increasing use of metadata and the installation of online research-data repositories open up growing resources of information for scientometric analyses. In that context, we suggest exploiting metadata produced by ecological projects in

order to create tools which are focused on the visualisation and evaluation including internal processes and project resource. These tools can be utilised finally for better project management and evaluations. We exemplify the use of descriptive metadata by describing a selection of aspects along a decade of research carried out in the BEF-China project (Bruelheide et al. 2014). We describe some selected characteristics of the project and its resources and discuss how the metadata can provide feedback to principal investigators and funding agencies to better achieve their project goals and finally increase the potential outcome and overall value which is produced by ecological projects.

Material and methods

BEF-China

BEF-China is an international research project with the goal to disentangle the influences of plant diversity on functions and services of ecosystems in subtropical forest ecosystems (Bruelheide et al. 2014) The project was set up across two sites which are located in provinces of southeastern China (Zhejiang and Jiangxi). 147 individual researchers from China, Germany, and Switzerland were involved, structured into 16 sub-projects. Two of the sub-projects are responsible for the coordination of “Central Projects”, and the other 14 are working on a wide range of biological and ecological objectives (c.f. appendix Table 1).

The project developed an own data management platform which is called BEF-data in order to manage, document, share and curate all of its datasets (Nadrowski et al. 2013). Also, the application provides a mechanism to initiate and guide upcoming collaborations. This feature is enabled by allowing project partners to request data from each other along with all the relevant metadata like a description of the new research ideas via a so-called paper proposal (Nadrowski et al. 2013). These paper proposals serve as a single point of reference which finally aggregates and collects information about the research idea. This documentation further involves the

authors, the included datasets and is linking to products in the form of publications (c.f. <https://bit.ly/2K1aFXj>).

The Essential Annotation Schema for Ecology (EASE)

The recently developed EASE annotation schema (Pfaff et al. 2017) consists of an annotation vocabulary and a metadata schema. Both of these components are organised and structured around eight categories of information. Further, the categories and vocabulary are based on various vocabulary standards, books and expert knowledge (Pfaff et al. 2017). The main categories of information are “Time” (e.g., the temporal extent and the resolution), “Space” (location names, the spatial extent and resolution), “Sphere” (e.g. layers or parts of the pedosphere, hydrosphere, and the atmosphere where measurements have been made), “Biome” (e.g. type of biomes, latitudinal zones and climatic influences on the seasonality), “Organism” (full names and taxonomy), “Process” (e.g. the names of processes and interactions), “Method” (e.g. the general study approach and the variables which are manipulated to span gradients), and “Chemical” (elements, compounds and biological functions of chemicals). Further, the schema includes a part covering administrative metadata which includes, for example, a title, an abstract, the name of authors and the hosting data repository. EASE has been designed with the goal to provide a consistent basis for a fast and sophisticated annotation of ecological research data in order to improve their visibility and reuse.

EASE is accompanied by a web-based annotation tool (<https://git.io/v5wWe>). The tool is agnostic to data formats and thus allows the annotation of typical research data, e.g., tables, images, videos or audio files via an intuitive graphical user interface (c.f. Figure 4). The application does not only provide the visual support for the annotation (e.g., navigation menus and forms to fill) but also provides access to the vocabulary of EASE during the annotation process. An auto-completion mechanism helps to select terms during the annotation and provides the individual term definitions. The selection enables a harmonised use of terms, prevent spelling

mistakes and speeds up the annotation process. Also, the annotation tool is supporting the import and export of different metadata formats. This mechanism allows for a high degree of compatibility with relevant standards such as EML, ABCD or DwC. The compatibility is achieved through the use of XSLT stylesheets which allow defining meaningful mappings of different but similar concepts in between the metadata standards. This mapping can be used for conversion of information between instances of the schemata and thus allows new annotations in the EASE tool to be based on already existing metadata even if it has been stored in a format not native to the application (Pfaff et al. 2017).

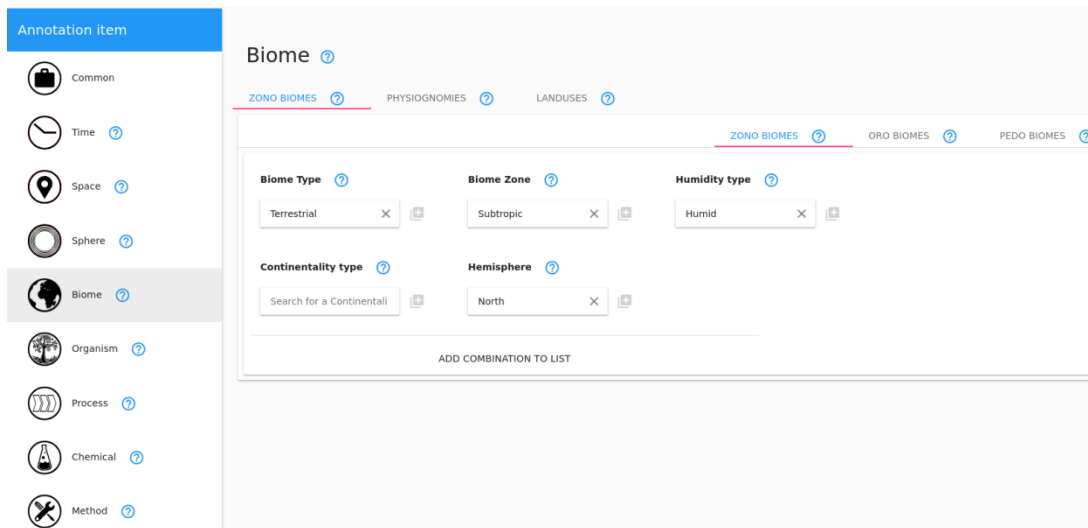


Figure 4 The user interface of the EASE annotation tool. Here it shows the concepts summarised under the “biome” category with an ongoing annotation of a dataset from the BEF-China experiment. The forms hold information about the type of biome, the latitudinal zone, the water availability, the continentality and the hemisphere (in this example the dataset describes a terrestrial, sub-tropic, humid, biome in the northern hemisphere, continentality is not applicable for the biome described and thus left empty).

The annotation process and complementing data

First, the metadata of a total of 250 datasets from the BEF-data portal was downloaded (<https://bit.ly/2JjfILX>) in the Ecological Metadata Language (EML) format (Fegraus et al. 2005; Nadrowski et al. 2013). An XSLT stylesheet was developed next in order to be able to convert the downloaded EML files into the format of EASE. This, however, was only possible for a part of the information (e.g., names of the researchers and variables, coordinates of the bounding box).

Subsequently, the created stylesheet was used to import the information from the EML files into the EASE annotation tool. The annotation of the datasets then was carried out manually using the user interface of the annotation tool. The full-text descriptions with information about the data contained in the EML files were further used as a reference to guide the annotation (including, e.g., study setup, environmental conditions, chemicals and methods, processes observed). In order to carry out an analysis, the information finally was exported from the annotation tool in the EASE format (Pfaff et al. 2017).

The analysis of the metadata was carried out using the R language for statistical computing (R Development Core Team 2015). A parser was written to import the EASE formatted annotation files into the R environment. The import function yielded a data frame where each row is representing one of the datasets which were annotated. The columns contain variables which are either of the type date (e.g., the start/end dates of data collection event), continuous (e.g., elevation or soil depth) or binary. Of the “binary” columns, each informs about the presence or absence of a particular term in the annotation across all the datasets (e.g., was a dataset annotated with the term “Carbon” for chemical elements or not). In the subsequent paragraphs, we regularly use the expressions “annotation category” and “annotation feature”. With the first, we refer to categories of the EASE schema (e.g., Time, Space, Sphere) and with the latter to the instances of terms which make up such a category (e.g., carbon as an element would be one of the annotation features which makes up the Chemical annotation category).

For a better interpretation of the annotation data, it has been complemented with some of the publicly available metadata extracted from the data portal of the BEF-China project. This comprised information about individual researchers and the sub-projects of BEF-China. For the sub-projects, the title and a short description were extracted. Further, information about the paper proposals was extracted (c.f. Methods section) and for each of these, the identity and the number of individuals

involved, the associated datasets and the name of the journal in which the proposal was finally published. We also reached out beyond the BEF-Data portal for the H index of journals from the database of the SCImago Journal & Country Rank (<http://www.scimagojr.com>, 2017) in order to further complement the information about the proposals.

Selected aspects for the analysis and their background

The temporal dynamics of data collection in the BEF-China project was the first aspect which has been selected for analysis. The EASE annotations of the datasets were used to derive the start and end dates which represent the time frame for the data acquisition of each dataset. The years along the lifetime of the project were used as a grouping factor for the count of data collection events. This count then was further split in each year into the count of collection events which were starting, ending and running in the year. The turnover of datasets then was calculated as well using the start and the end date of the data collections. All together this information is providing insight into how the project is moving forward with the data acquisition but also into how a project organises these events over time. It might also serve as an indicator highlighting if research ideas and their related data collection events tend to accumulate in the project, or if they are finished rather timely.

The topics which have been covered by the project and their related dynamics were selected as a second aspect for the analysis. The annotations of the datasets were used to detect the first appearance of each separate annotation feature along the lifetime of the project (e.g., finding the date on which "Carbon" first appeared in the annotation body). The broader annotation categories (i.e., the top level of EASE) consecutively were used as grouping factor to create cumulative sums of their associated features over time. The element "Carbon" for example is part of the "Chemical" category. The cumulative counts have been scaled before they were combined in a single plot for a better comparison.

On top of this broader overview along major topics, two more detailed examples were created based on the same principles as above. These examples were possible due to the hierarchically structured nature of EASE which finally allows discovering more detailed parts of the annotation succinctly. The two examples show the details about chemicals which were measured in the project and the processes which have been observed over time. For this, the annotation category “Chemical” has been dissolved into its component categories which are chemical elements, chemical compounds, and biological functions of the particular chemicals. Beyond the visualisation of the cumulative annotation features in the component categories, the absolute count of the chemical features for each year in these categories has been visualised. For the second detailed example, the annotation category of “Process” was used to observe the processes measured over time (tracked by their names).

Using information along the structured annotation schema has the potential to shed light on the thematic focus of a project and to show how it is developing over time. We postulate that it finally allows a detailed evaluation of projects to answer, e.g., if the project has covered specific topics or when this happened (e.g., did they measure certain variables or did they cover a specific temporal or spatial resolution). Additionally, using the information could finally help to find gaps and provide hints on possible future directions of research.

The public perception of the project was selected as the third aspect of the analysis. It was approximated by comparing the H index (Hirsch 2005) distribution of potential journals in 2017 as an example reference with the distribution of H indexes of the journals in which the BEF-China project published papers in. The journals used for comparison have been filtered along their topic keywords for biology, ecology and general purpose journals as well as to topics targeted by the project (e.g., ecology, evolution, behaviour and systematics, genetics or geography, c.f. appendix Table 2). The narrower focus of journals allowed for a better

comparison as the H index likely depends on the scientific genre as some areas are citing more than others (e.g., they have more individual scientists). The comparison of potential journal H indices with the ones achieved by the whole project has the potential to be used as a measure evaluating the publicly perceived value and the quality of the research. Thus it could serve as an indicator of the success of a project. In order to investigate what drives the H factor of the paper proposals of the BEF-China project, a Simpson-diversity index has been calculated for each dataset based on the EASE annotations. The index was calculated separately for each top-level category per dataset (e.g., diversity of Location, Organism, Process). Along with the id of the proposal, the count of datasets per proposal and the involved persons per proposal has been fed into a random forest (regression type). The variable importance (i.e., the influence of the variables onto the accuracy of the prediction, no matter whether it is positive or negative) has been calculated for the predictors of the H index which the proposals achieved. The importance finally allows gaining a first idea into what are the strongest predictors for the impact achieved in a project.

The structure of collaboration in the project was selected as the last aspect of the analysis. The collaboration was approached on two different levels being (i) the interactions of the sub-projects and (ii) the interactions of individual researchers. Several network analyses have been carried out where the nodes represent either sub-projects or individual researchers. The connections between the nodes were determined based on the fact if there was a joint data collection effort, which means a common data ownership. For nodes in these networks, two measures were calculated. First, the “authority” was calculated. This centrality measure increases for nodes that have many connections to nodes which are well connected themselves and thus highlights strongly connected clusters of nodes. Secondly, the “eccentricity” was calculated (this one only for the individual-based networks). Eccentricity is measuring the shortest distances from each node to all the other

nodes in the network. Thus, it allows to express how close the nodes are to each other. The eccentricity was calculated further for two different scenarios of individual-based networks. First for a network of the individuals formed based on collecting data together. Second for a network of individuals that were co-authors on a publication. Finally, the eccentricity of the nodes in the two scenarios has been compared using a Wilcoxon test. The network analyses do not only give an insight into the project structure but are useful for example for principal investigators to detect new collaboration potential. The insights could finally drive the project management towards the benefit of the whole research network of the project.

Results

Data collection activity and throughput

A peak of data collection events in the project appeared with the start of the first funding phase in 2008. The count of new starting data collection events then is decreasing from that point in time continuously towards the end of the third funding phase. The most intensive data collection activity happened between the years of 2008 and 2012 with a peak of 96 collection events in the year 2012. We found an increasing amount of data collection events to be finished from 2008 to 2012 with a peak in the year 2012 ($n = 75$, Figure 5). The total turnover of the datasets (i.e., new datasets appearing and old disappearing based on their collection time frame) in the project is positive, and thus it is characterised mainly by data collection events which are finalised (c.f. Figure 6).

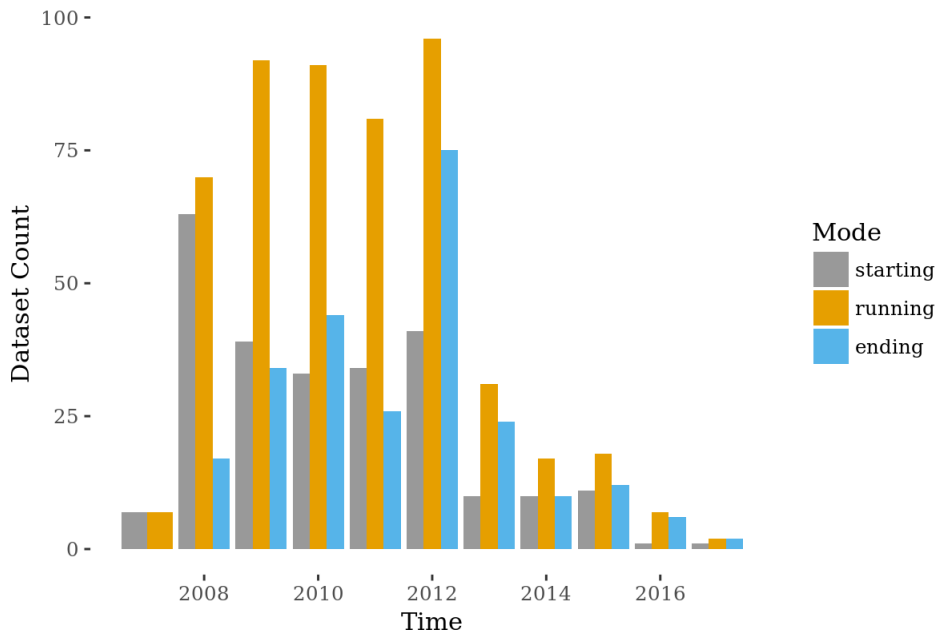


Figure 5 An overview about the data acquisition effort per year along the lifetime of the project which is ranging from 2007 up to 2017. It shows the count of data collection efforts starting, ending and running per year. A majority of datasets were started in 2008, and the highest number of parallel data collection was observed between the years 2008 and 2012 with a peak in 2012. In 2012 there is also a peak of data collections ending.

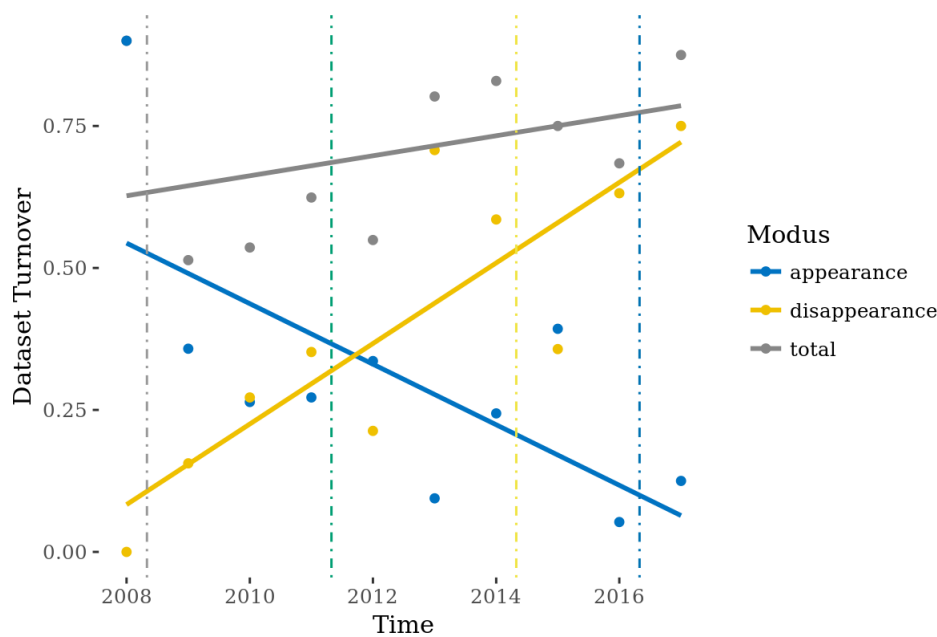


Figure 6 Turnover of dataset collection events in the BEF-China project in total and faceted into the components of turnover being appearance (data collection started) and disappearance (data collection ended). The total turnover highlights a positive trend. The dashed lines represent the beginning and the end of the funded project phases (... , 2008-05-01 = gray, 2011-04-30 = green, 2014-04-30 = yellow, 2016-04-30 = blue, ...).

Coverage and dynamics of topics

The cumulative sum of unique annotation features used under each of the annotation categories along time shows that information about organisms dominates the project, followed by information about chemicals, methods and processes (c.f. Figure 7). The dynamics of the categories highlight that some annotation categories are saturating faster (e.g., Biome, Space, c.f. Figure 7), whereas other categories are growing more slowly but receive new contributions along the full lifetime of the project (e.g., Method). We also see that some categories are more dominant in the first project phase (Time, Sphere, Space, Biome) whereas others are taking over later (Organism, Chemical, Process, Method).

The first of the detailed examples using the components of the "Chemical" annotation category shows that the project has a focus on chemical compounds which is followed by elements and biological functions (c.f. Figure 8; For a more detailed explanation of these concepts see Pfaff et al. 2017). The chemical elements in the project reach approximately half the abundance of chemical compounds. Chemical compounds reach their saturation (i.e., the maximum number of different compounds) in the second project phase whereas the biological functions of chemicals come into play in later phases. The second detailed example shows the processes, which have been covered along the time represented by their names. According to the graph, the project is mainly focusing on the growth of plants, dissimilation, and processes related to nutrient cycling (c.f. Figure 10).

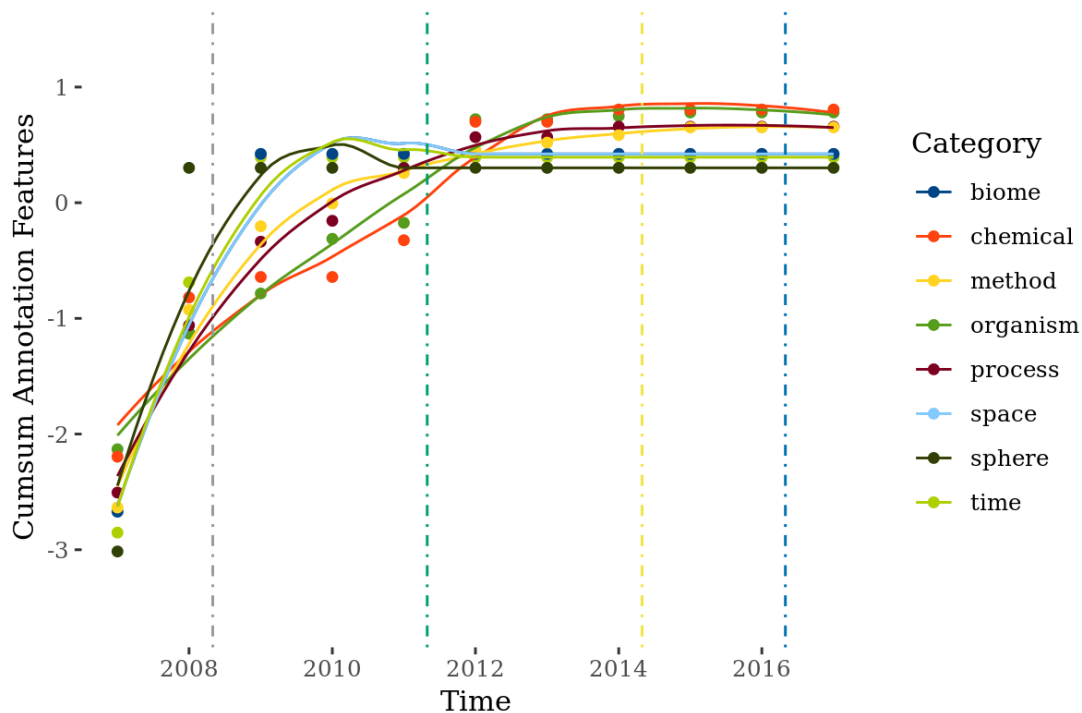


Figure 7 The cumulative count of unique features in the annotation contributing to the respective annotation categories (e.g., Time, Space) along the lifetime of the project from 2007 up to 2017 (scaled for comparison reasons). It highlights some aspects to be more important in the first project phase as they were accumulating and saturating faster (e.g., biome and spatial information) than others (e.g., Method or Processes). The dashed lines are representing the beginning and the end of project phases (... , 2008-05-01 = grey, 2011-04-30 = green, 2014-04-30 = yellow, 2016-04-30 = blue, ...) (graphics created with ggplot2: L. Wilkinson 2011).

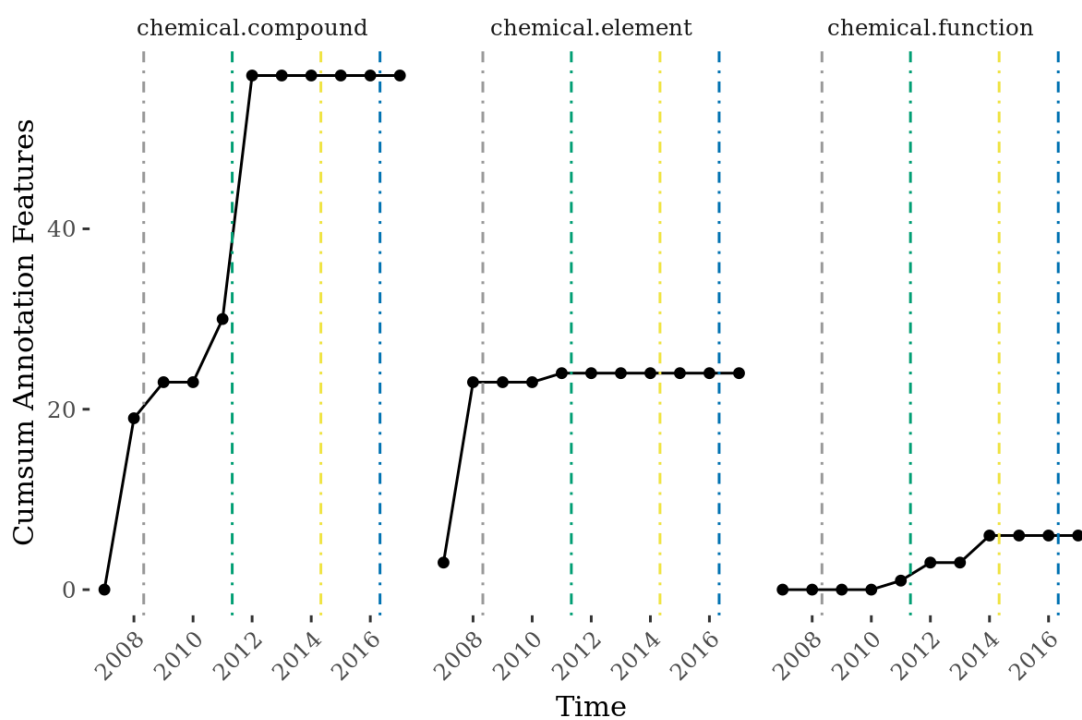


Figure 8 The cumulative count of the unique annotation features contributing to the chemical annotation category from 2007 up to 2017. It shows that a high count of chemical compounds directly followed by chemical elements and the biological functions of chemicals dominate the chemical aspects. The dashed lines designate the beginning and the end of the project phases (... , 2008-05-01 = gray, 2011-04-30 = green, 2014-04-30 = yellow, 2016-04-30 = blue, ...).

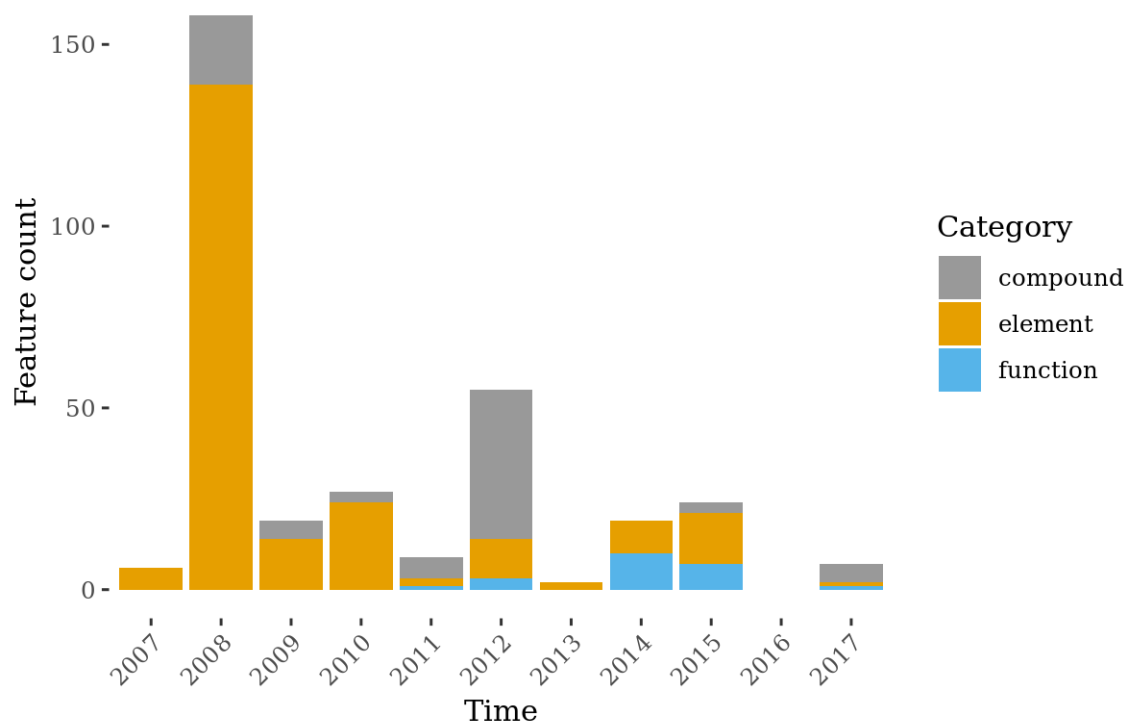


Figure 9 The count of the appearances of chemical features along the project lifetime by years separated by categories they belong to. This overview provides insights into when the topics have been dealt with throughout the project and to which extent. Chemical elements have been measured across all years and are typically dominating. The compounds are measured more sporadically and take over the dominance only in two years being 2012 and 2017. Biological functions of the particular chemicals are measured the least and the most prominent in later phases.

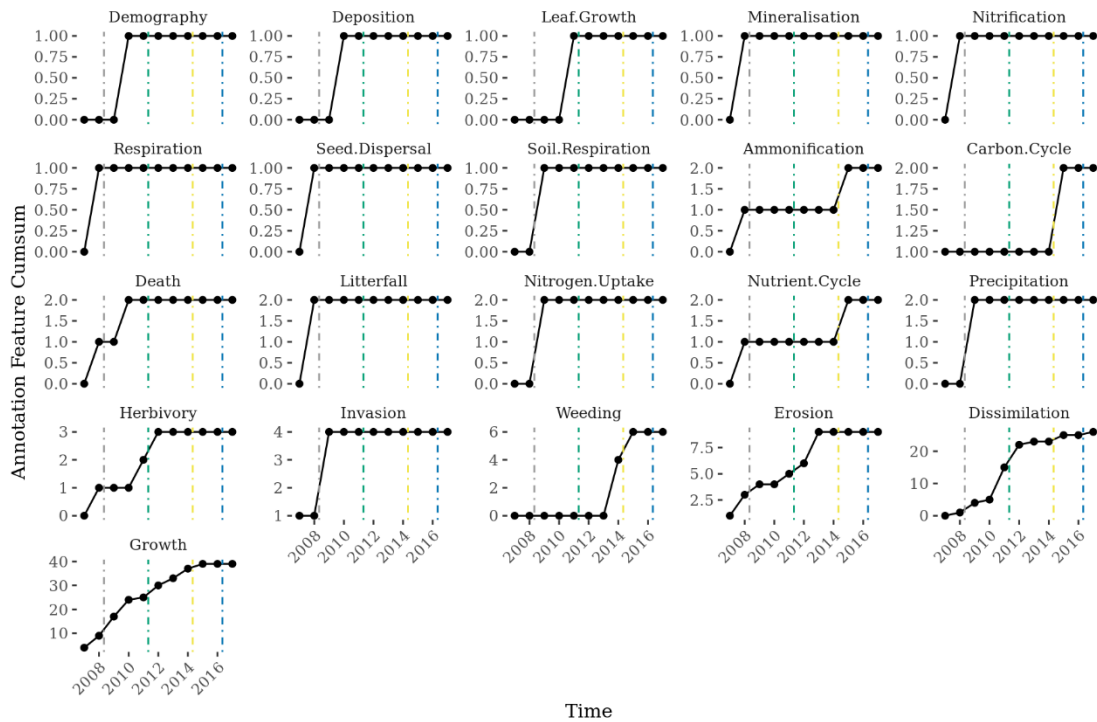


Figure 10 The annotation category “Processes” represented and tracked by the process names and their cumulative sum of mentions over time from 2007 to 2017 (in ascending order). This overview shows when and how often specific processes have been targeted and when they reach a point of saturation in the project. The project seems to focus on the processes of “Growth” directly followed by “Dissimilation” and “Erosion” of which all are mentioned in the main objectives of the project.

Public perception of the project

In BEF-China 147 researchers have been involved of which 85% are owners of data according to the metadata; 72% have been involved in at least one paper proposal and finally in the resulting publications. 176 research proposals were created over the lifetime of the project, out of which 108 finally were accepted for publication by peer-reviewed scientific journals. The proposals differed widely in the count of datasets on which they were based on ranging between one and 43 (mean = 11, this includes published and unpublished proposals). The proposals were accepted in 50 different journals. The H indexes of these journals ranged between 15 and 240 (mean = 143, per reference of 2017). The distribution of the H indexes of journals targeted by the project's publications compared to the global H index of potentially relevant journals from 2017 (including biological ,ecological journals and

multipurpose journals) shows that publications which are produced by the BEF-China project are above the overall mean of H indexes (c.f. Figure 11); the majority of the publications is even placed inside the third quartile of the potential H indexes ($n = 76$; 70,3%). A Kruskal-Wallis test revealed a highly significant difference between the two groups of h indices ($p < 2.2e-16$). The random forest along the paper proposals revealed that the most influential predictor for the H index achieved by the proposals is the count of datasets which are used. The count is potentially an indicator of targeting more complex research questions. The count was followed then by the diversity measures led by organisms ahead of processes, space and methods. The diversity index of biomes only had a marginal predictive impact on the H index whereas the persons even had a negative impact on the prediction quality (c.f. Figure 12).

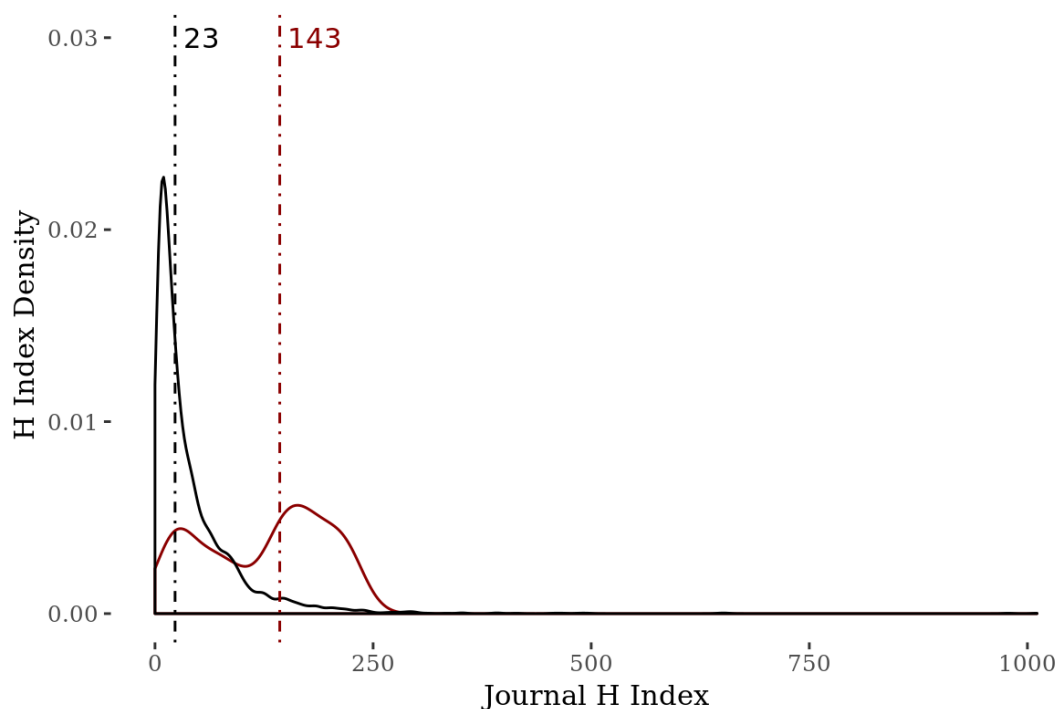


Figure 11 The journal H indices of 2017 including 3687 journals (black line data and the black dashed line = mean). In red the frequency of H indices of the journals in which the BEF-China project published in (50 different journals, the dashed line shows the mean). The BEF-China project published mainly in the third quartile compared to the H indices of potentially relevant journals. A Kruskal-Wallis test shows a highly significant difference between the groups ($p < 2.2e-16$).

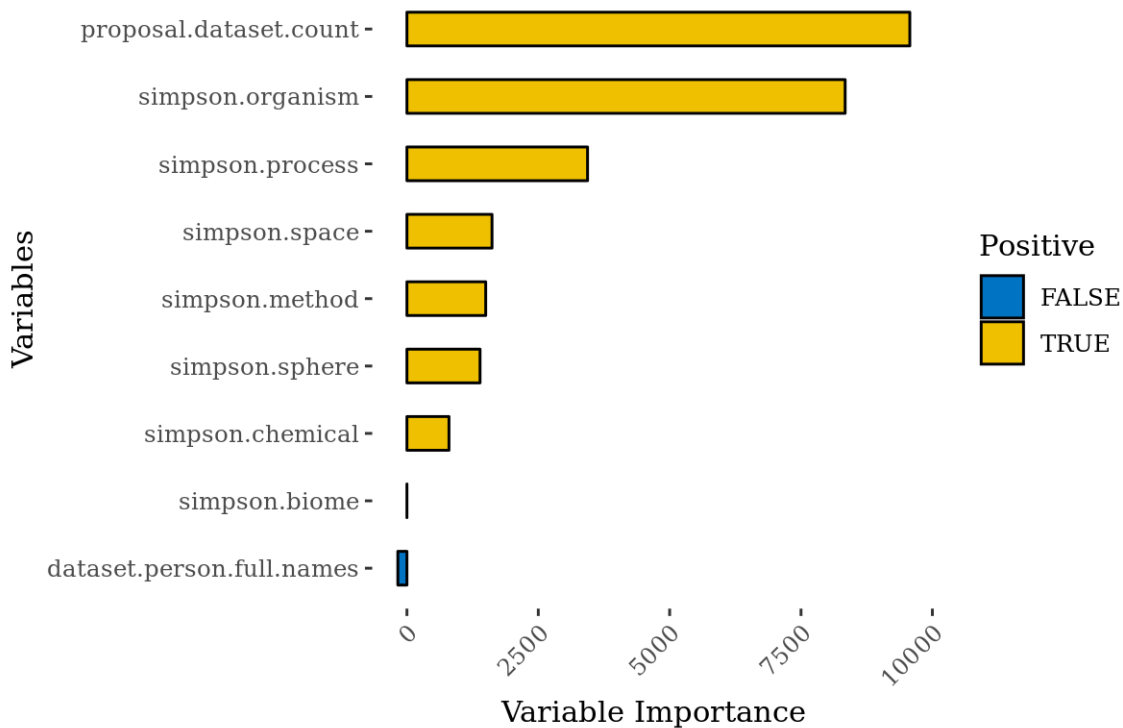


Figure 12 The variable importance derived from a random forest (regression type). The input variables are shown on the left-hand side (y-axis) and the importance of the variables onto the h-index of journals on the x-axis which has been achieved by the publications of the BEF-China project. The random forest input variables explained 64.37 per cent of the variance in the data. The count of datasets has the highest impact on the prediction of the H index achieved by the proposal. This is followed by different parts of the diversity of the data captured by the EASE annotation and finally the individual persons involved in the publication.

Collaboration structure

On average, the number of individual researchers collecting data in a joint effort is around three, with an absolute range from one to nine researchers (c.f. appendix Figure 16). The collaboration network of sub-projects shows that all of the projects are well connected, except the sub-projects 11 and 12, which indicates that they do not have any joint data collection with other sub-projects (c.f. Figure 13). Based on joint data collections the individual researchers in the network show up as well connected except for a few ones which only form a single mutual relationship with one other researcher (c.f. Figure 14 and Appendix Figure 17). The Wilcoxon test comparing the node eccentricity across the two different individual-based

networks shows a significant difference (collaboration along the data collection versus on publication). The node eccentricity in the BEF-China project is higher during data collection and lower in the publication related network. Taken together there are less, close collaborations in the project during data collection and significantly more, close collaboration during the publication (c.f. Figure 15).

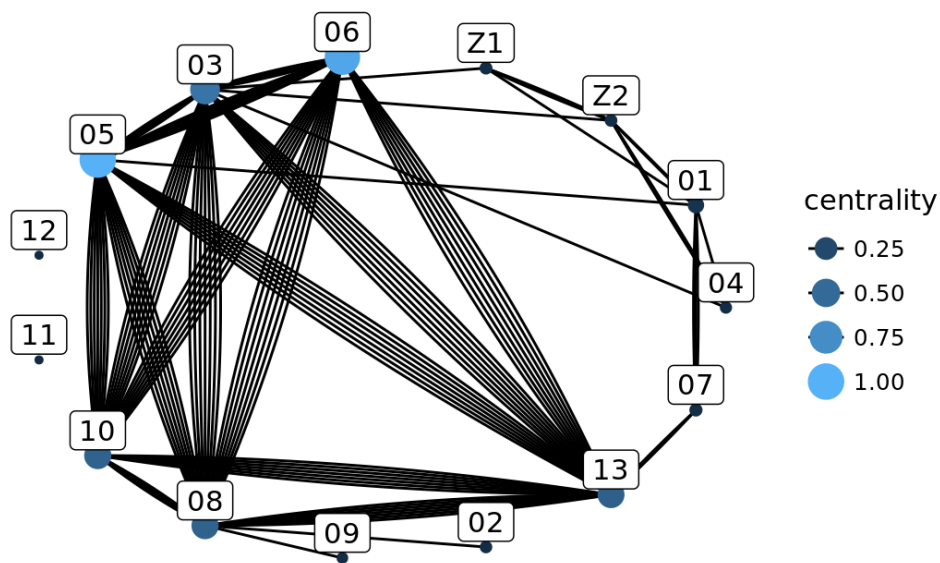


Figure 13 The sub-projects and the collaborations based on datasets collected under a joint effort from 2007 up to 2017. The size and the colour represent the centrality, which highlights how well the sub-projects are connected and to how many nodes they are connected with which have good connections as well thus forming strong clusters showing synergies in data collection. The project 05 and 06, covering processes like soil carbon fluxes, decomposition, nutrient cycling, soil erosion and water resource management, show the highest centrality in the network. The figure also highlights projects which have no documented datasets in collaboration with other projects (11, 12). Please, note that SP12 was only funded in phase 1 and 2, while SP14 only started in the third phase. There was no dataset available from project 14 thus it does not appear in the plot. For more information about the sub-project's objectives see Appendix Table 1.

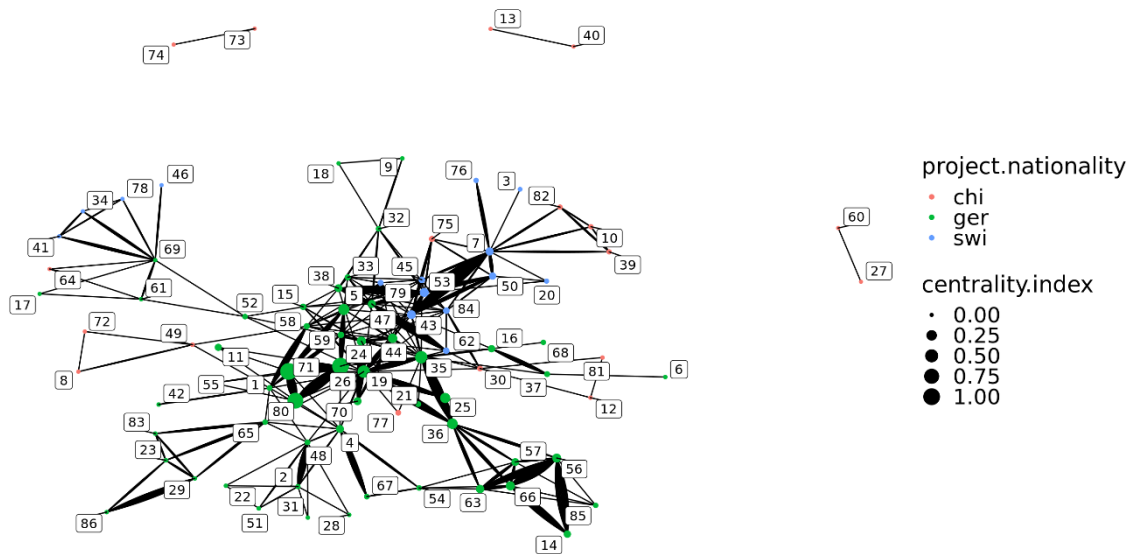


Figure 14 Overview over the personal level connections in the BEF-China project. The nodes are representing individuals and the edges each a collaboration based on working together on the collection of an individual dataset. The nodes here are sized according to the centrality in the network. Here a centrality measure was selected, that finds strongly connected individuals whom themselves are again well connected. Thus, the algorithm detects well-connected clusters which have a strong synergy with joint data collection. The colour code highlights the nationality of the institution of the main PI the individuals are associated with (based on being organised in the same sub-project chi = China, ger = Germany, swi = Swiss).

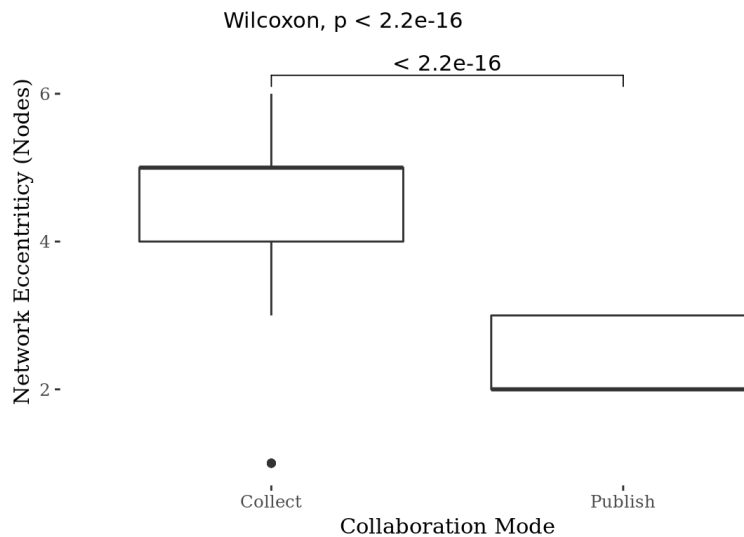


Figure 15 The comparison of the eccentricity of the nodes (i.e., researchers) between the two different personal collaboration scenario or networks (data collection versus publication). The research network of the BEF-China project collecting data on the left and publishing articles together on the right. The eccentricity in the research project is significantly higher in the network, which the individual researchers form during the data collection. However, it decreases for the publication network where researchers are overall more well connected.

Discussion

Our examples highlight the untapped and hidden potential of information in metadata of ecological projects. It can be exploited in many different ways even beyond the original purpose of the metadata (Fegraus et al. 2005; Pfaff et al. 2017). Information from the metadata can be utilised along developing scientometric instruments which in turn can be used to improve the self-awareness of ecological projects and for project management (Atkinson 1999). However, the ideas that are presented here do not claim completeness. They instead represent use-cases to serve as the foundation to extend upon. In the following, we are discussing separate parts of the analysis before wrapping them up into a broader context, indicating how we think their usefulness and interpretability can be maximised in the future.

Data collection activity and throughput

The bulk of data collection activity in the BEF-China project was starting with the beginning of the first funding phase in 2008. Afterwards, there were continuously fewer data collection campaigns starting. The decline reflects the fact that the project has to calculate with the received funding and plan with the time that is available to finish objectives before the funding period is ending. The positive turnover of the data collection events in the project can be seen as an indicator for how the project works or how it has been organised. Thus, it might show that the project was well organised from the beginning. Small and modularly defined data collection events are carried out which are finished continuously along project phases. New collection events are coming in rather sparsely and complement the existing data without starting to dominate the project (c.f. Figure 6). Both of the graphs that we have developed here are shedding light on aspects which might be particularly interesting for principal investigators. Continuous monitoring of the data collection dynamics could help to pinpoint potential problems and serve as an early indicator of the success of a project. For example, if data collection events are not finished in time, this could be an indicator for problems with the data collection and finally trigger a meeting between, e.g., the responsible data collectors and the

principal investigator to help to find a solution. Further, if new collection events are aggregating and dominate a project, this could indicate problems in management or the overall communication which might need some attention, e.g., developing better project structures or get expert advice from outside.

Coverage and dynamics of topics

The analysis of the topics provides an insight into the thematic development of a project (Rip and Courtial 1984). It does not only show the overall dominance of specific topics and with this the orientation or the focus of the project but also indicates the dynamics of separate topics along time (Pollack and Adler 2015). While the overview along the broad annotation categories of EASE (e.g., Time, Space, Sphere) is already helping to get a general understanding of the project, there is even more potential hidden in the individual metadata categories. The two more detailed examples exhibited that the fine-grained metadata could potentially help answer questions like, e.g., if the project carried out research related to specific topics as well as when this exactly happened and to which extent. This information is interesting for researchers in a project as it can help to prevent redundant efforts, strengthen the synergies with new collaborations or beyond this allowing detection of yet untapped topics for potential future directions of research. Additionally, this information could be used in the context of an evaluation or defence of the project as it allows to better pin down if the project reached certain thematic milestones or to what extent (e.g., did they measure the carbon content, how often and when?).

Public perception and collaboration structure

Based on the information whether the sub-projects have a documented interaction in the data collection or not, we see that all reasonably well connected. This form of network analysis can help to detect if the sub-projects in a larger consortium are separated or on the opposite how well they are integrated. This insight could finally allow taking action if needed for an improvement of the collaboration for the benefit of the whole project, e.g., incentivise or stimulate integrative research to

strengthen the synergies. However, sometimes sub-projects might be set up to be separated on purpose. Disconnected projects can contain innovative pilot studies testing, e.g., a new methodology. These need to be established first before the project can be connected in collaborations with the other research going on in a project. Based on whether the separate individuals in the project have worked together on a particular dataset or not we can see that there are connections between most of them. There are some researchers, however, which are more actively involved with the data collection. Thus, they are stronger connected and form clusters in the network. These clusters are of importance for the overall amount of data collection which happens in the project. On the other hand, there are some researchers, which have been working only with a hand full of others on a data collection. They might stand for individuals establishing new ideas or represent less connected and separated projects, e.g., supervisors and their PhD students. In concert with the measure of node eccentricity and the comparison between the two collaboration scenarios of individuals being data collection versus publication, the networks provide valuable insights into the performance and the structures of a project. We see that the eccentricity of the nodes (i.e., individual researchers) for the data collection is significantly higher in comparison to the network of publications. It shows that the collaboration between the researchers in the project is much tighter when it comes to publications compared to the data collection. The shown difference might be an indicator of a nice organised ecological project which is distributing the data collection across interdisciplinary specialists before the data finally is discussed and integrated with diverse colleagues analysing and publishing it together.

While the journal impact factor has been criticised as a measure of scientific success, the H index has been proposed as a straightforward alternative (Seglen 1997; Hirsch 2005). The H indices of the BEF-China project indicates that it publishes over-average compared with a global picture of H indices from potentially suitable

target journals. Such a comparison might be useful as an indicator for the overall success of the project. When comparing multiple projects with each other, this gets even more interesting. It could allow approaching the question about the causes of a higher H index (e.g., project size, lifetime, spatial extent). The random forest analysis and the importance based on the variables that have been derived from the EASE annotation are indicating that the count of included datasets has the highest impact on the H index. The count of datasets is directly followed by different aspects of the thematic diversity of the data lead by organisms. Taken together it indicates that the diversity of the data which is included in an analysis has a strong influence on the impact which is finally achieved with the publication. An interesting addition to this for the future would be to take a closer look at the published papers of a project, e.g., using topic models to extract actionable data from the text. This analysis could go hand in hand with investigating into the topics of a broad set of publications to elucidate the “zeitgeist” in the domain of ecology. This overview could allow showing what topics are *en vogue* during the lifetime of a project to finally better see where the papers with a higher and with a lower impact factor are located (topic-wise, Neff and Corley 2009).

Wrap up and outlook

Here we have shown that there is much potential in the metadata of ecological projects to be unleashed. It can help ecological projects to become more self-aware, potentially more successful and to create more value in the long run. Together this benefits can be enabled by less redundancy, stronger utilisation of the synergies and by the detection of gaps in the covered topics in order to decide about future research directions. The examples that we have shown are providing insights into a long-term ecological project which are useful on their own. However, several aspects would benefit from a broader comparison taking into account a range of different ecological projects. That would relativise the absolute value of the data collection and turnover, the impact factors and the influences of predictors as well as the project structures in the collaboration networks. Extending the analysis for

more projects, in the end, would then potentially enable the generalizability and finally a projection into new projects. That, in turn, could guide decisions on the project management (e.g., how long does it typically take to account for a specific topic appropriately?). Thus we plan to extend the presented ideas in the near future to a broader set of ecological projects and towards a general framework which is further detailed in the following.

A workshop could help to bring board members of funding agencies, researchers from Ecology, project management and scientometrics together. They have to discuss which information is most useful as transparent feedback and for evaluation and feedback purposes. Based on the outcome of the workshop the methods which are presented in this article could be extended and improved, e.g., to track broader sets of topics, term co-occurrences or changes in the interest of the researchers along the course of time. A manual analysis with the improved tools across multiple projects could deliver further insights. A broader data basis could potentially allow answering questions like, e.g.: Are projects failing when they are dominated by the aggregation of new data collection events over their lifetime? Is the centrality measure in the collaboration networks in comparison between the data collection and the publication an indicator for the success of a project; and is it reverse to what we have found in our analyses in unsuccessful projects? Further, the information could help to evaluate if specific topics need more time to be addressed appropriately by comparing different projects with similar topics or goals with each other. The outcome could finally be taken into account during the project planning or in the decisions made about the funding of projects with a particular focus or scope (Healey, Rothman, and Hoch 1986; Landreth and Silva 2013).

Beyond all of this, the Essential Annotation Schema for Ecology and the ideas presented in this article could be readily implemented into existing data management platforms like, e.g., BEF-Data or BEXIS. This integration would finally

bring the resulting benefits into the data- and project-management workflows ecological projects are already used to. Beyond this, the integration could further serve as a motor driving the expansion of the presented ideas into new projects increasing the interpretability of the results in relation across a broader data basis.

Acknowledgement

Thanks to the group Spezielle Botanik und Funktionelle Biodiversität at the University of Leipzig and all their members for the continuous support regarding proofreading the manuscript and for the intense discussions bringing this work forward. Special thanks go to Ronny Richter and Daniel Marra for the detailed discussion of ideas for statistical methods for the analyses. We also thank the DFG for funding BEF-China as a project and all the hard working data managers (Karin Nadrowski, Anne Lang) on the BEF-Data portal for curating the datasets and metadata.

Appendix

Graphics

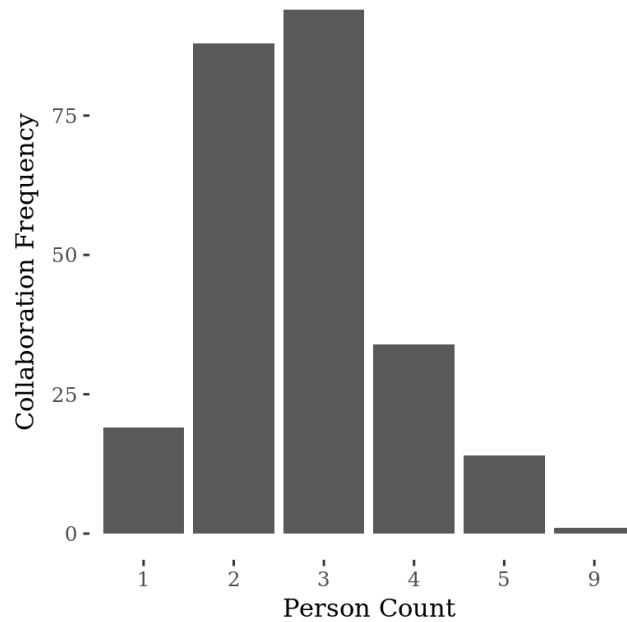


Figure 16 The collaboration structure based on individuals working together collecting a dataset ($n = 250$) along the whole project lifetime from 2007 up to 2017. The collaboration frequency follows a normal distribution, centred around 3 and reaching out to a minimum of 1 and a maximum of 9 persons.

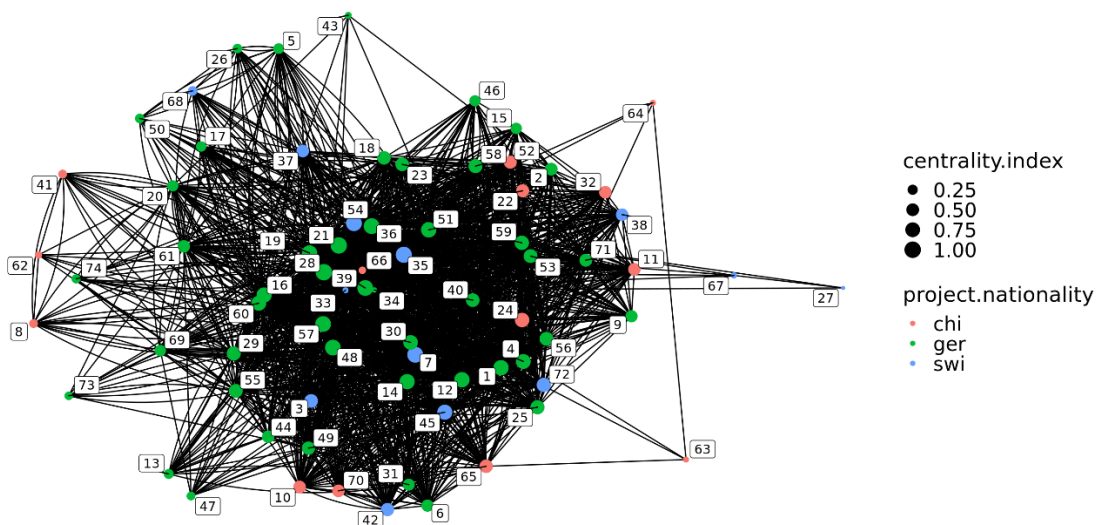


Figure 17 Overview over the personal level connections in the BEF-China project. The nodes are representing individuals and the edges each a collaboration based on publishing a paper together. The size of the nodes represents their centrality in the network. This centrality increases for strongly connected individuals which themselves are connected with others that have many connections. Thus, the algorithm detects well-connected clusters with a strong synergy in publication effort. The colour code highlights the nationality of the institution of the main PI who led the sub-projects the individual was affiliated with.

Tables

Table 1 It shows the sub-project with their id and a short description derived from the information scraped from the BEF-China data portal. Please, note that SP12 was only funded in phase 1 and 2, while SP14 only started in the third phase.

Project id	Short description
01	Below ground primary production, demography, production
02	Seasonal growth, demography
03	Functional diversity (traits)
04	Insect genetic diversity
05	Soil carbon fluxes and decomposition, nutrient cycling (carbon, nitrogen)
06	Soil erosion, water resource management (plant diversity)
07	Soil microbes (mycorrhiza)
08	Microhabitat litter layer, Functional role of herbivores, predators, and saprophytics
09	Plant-insect interaction
10	Deadwood (dynamics)
11	Succession and invasibility
12	Root traits and plasticity, phosphorous availability and cycling
13	Soil microbe physiology profile, biomass, and activity
14	Pathogens (fungal) (no data has been available at the date starting with the

	annotation for this paper)
Z1	Coordination and project management
Z2	Data management

Table 2 The full overview of journal categories the BEF-China project published in. It has been used to filter journals and their H indices from the list of all journals in the Scimago database for the comparison against the H indices achieved by the BEF-China project.

Journal Categories
ecology, evolution, behavior and systematics, forestry, nature and landscape, conservation, agricultural and biological sciences, plant science, environmental, chemistry, earth and planetary sciences, global and planetary change, agronomy and, crop science, environmental science, earth-surface processes, physiology, soil, science, ecological modeling, medicine, microbiology, genetics, insect, science, geography, planning and development, animal science and, zoology, multidisciplinary, biochemistry, genetics and molecular biology, biology

General Discussion

Chapter One

Closing the gap between primary data repositories and the knowledge which is finally presented in publications is crucial for a sustainable scientific culture (Poisot, Mounce, and Gravel 2013). The link not only allows to preserve invaluable datasets by making them more visible but also helps to build up a concrete fundament for future research (Tenopir et al. 2011). Old data can be a breeding ground for new ideas and potentially help to solve the most critical questions of our time (Sala et al. 2000). An introduction of policies at journals demanding the curation and the publication of data are only a first step in the right direction. More reliable mechanisms need to be installed at journals in order to check if authors finally adhere to data publication and documentation guidelines (Roche et al. 2015). The checklist could comprise the use of standard data formats, the provision of primary research data and the completeness of the documentation, but also rigorous checks for the data quality (White et al. 2013; M. D. Wilkinson et al. 2016). Linking the data stewardship policy with funding and making data citable can further contribute motivating researchers to create exhausting documentation and to publish primary data for sharing and reuse amongst a broader audience of scientists (e.g., European Science Foundation 2008). However, providing detailed metadata as documentation is highly time intensive. Thus native support for documentation built into the software, researchers use in ecology for their daily analyses could be beneficial (e.g., R, Python, BExIS, BEF-Data). The tools could help to organise, collect and prepare information in a transparent way. A good example is workflow tools like Kepler or Pegasus (Altintas et al. 2004; Deelman 2005). They allow creating so-called provenance records which keep track of the data manipulation starting from the import of research data into the environment down to all the finally derived data products. (Buneman, Khanna, and Wang-Chiew 2001; Bowers et al. 2006). The collected provenance information can be used in the end to

create a data manipulation and analysis report in text form or a graph for visualisation (Simmhan, Plale, and Gannon 2005). Only preserving R scripts as a provenance record as it is suggested in chapter one is coming with the downside that scripts can be hard to read and understand in some cases. While R has lacked tools for provenance documentation for a long time (Silles and Runnalls 2010) over the recent years, there have been attempts to create such functionality, e.g., with the RDataTracker package (Lerner and Boose 2015). With a little bit of preparation, this package enables the creation of documentation and graphs, similar to the workflow tools mentioned before. The documentation can finally contribute to the understanding of what an R analysis script does. It can substantially improve the suggested form of documentation from chapter one and should be considered as a crucial complement to saving the executable script.

Another downside of the suggested documentation workflow is that it involves a “private” data repository (i.e., BEF-Data). There, the provenance record it is likely to befall the same fate as data which tends to get lost over time (P. Bryan Heidorn 2008). Thus, a more holistic approach to preserving the documentation is required. It could include handing over the full set of documentation to a publisher or scientific data repository who are in charge of taking care of the data and documentation in the long run (e.g., DataDryad, FigShare: Singh 2011). Along those lines, it would also be beneficial to develop standards for the transfer procedure including, e.g., the exchange of the documentation or the structure and mandatory content of such a documentation package. It could include for example the publication as pdf and text format, the primary data, analysis scripts, the derived graphs and other data products (e.g., jpg images and tables). A standard could help publishers to create generic interfaces for the publishing process. Private repositories with data, in turn, could implement new routines which allow submitting a finished project or dataset on the press of a button. A standardised format could further help with semi-automated quality checks on both sides (i.e.,

researcher/journal). Taken together, this could help to ensure the proper documentation and preservation of data as well as it helps ensure faster reuse of data (e.g., data processing and analysis tools can provide functionality to import available documentation packages).

While chapter one focused on the idea to close an important gap in the documentation along with the analysis of ecological data; it also indicates the urgent need for a more holistic solution for the documentation of data and in ecology and the resulting benefits. The presented `rBEFdata` is the first package implementing an ecological metadata standard into the widely used R environment (Touchon and Mccoy 2016). The implementation, however, is far from complete. It only covers a small subset of the Ecological Metadata Language (EML, Fegraus et al. 2005) to provide the essential information relevant to the processing of the data. In the meanwhile, a new package has been implemented by the `rOpenScience` community covering the full schema of EML (I contributed to this package as well, particularly the part of organisational metadata for the authors and the management of citations). Beyond the access to the information in EML, this new package allows creating documentation using the EML format while analysing data in the R environment. This can be the prerequisite for new tools which help to derive documentation transparently as far as possible without bothering the researcher (e.g. collect categories from data and ask for a definition if it is unknown). Further, it allows accessing EML described datasets and facilitates their analyses and synthesis providing relevant information similar as to what is shown in chapter one.

Chapter Two

Next, to the preservation and documentation of data, the discovery of data is equally important (Ryen W. White, Bill Kules, Steven M. Drucker 2006) for the long-term success in ecology. Much effort has been put into the development of metadata standards preserving information about the content and context of data

sets in ecology (Fegraus et al. 2005; Wieczorek et al. 2012; Holetschek et al. 2012; Pfaff et al. 2017). The information in metadata schemata in use is ranging from text-based descriptions down to more fine-grained annotations using precisely defined attributes (e.g., a name of a geological age or a time zone, Strohmaier, Körner, and Kern 2012). Fine-grained and precisely defined information is more accessible from the metadata while information in the full-text descriptions is rather laborious to utilise in a data discovery (Greenberg 2005; Beall 2008). While full-text descriptions contain much valuable information for a human reader, a computer, and with this, the most widely used search algorithms cannot make too much sense of it (i.e., full-text search; Beall 2008). Although the access problem to information in full-text can be approached somewhat by machine learning techniques and natural language processing (e.g., extracting abstract topics), specific ontologies are finally required for the detection of meaning to make efficient use of the extracted content (Alani et al. 2003; Chowdhury 2005; Walls et al. 2014).

However, the development of good quality ontologies is time and labour intensive. The development breaks down to three fundamental issues. The first is the upper-level ontology problem. Before any concept can be modelled semantically, it needs various building blocks of a rather generic nature. For ecology that could include concepts of “planet” and “space” (e.g. with “latitude”, “longitude”) and “location” (adding e.g. “elevation” and “name”) but also “boundary” to finally be able formulating a model which describes e.g. what a “geographic region” is. Beyond this, a concept of “time” would be required as well in order to be able to model ecological processes. These upper-level ontologies are abstract, and there are many of these today which are built on different philosophical backgrounds and individual perceptions of the real world (Mascardi, Cordì, and Rosso 2007). Neither choosing one of these available ontologies as the right one and basis of an own ontology nor connecting domain-specific concepts, e.g., from ecology with the abstract descriptions in an upper-level ontology is a trivial task. It requires highly

interdisciplinary expertise (e.g., philosophy, ecology, biology, informatics) on the one hand as well as a solid understanding of the chosen ontology on the other hand.

The choices when developing an own ontology are 1. Selecting an existing upper-level ontology as a foundation and use it, 2. Develop an own upper-level ontology, or 3. not to use any upper-level ontology at all. Because there is not “the” single upper-level ontology all research domains agreed upon so far, domain ontologies are often created using the latter choice; And this brings us to the second major problem. There are many ontologies available describing knowledge in the context of ecology in different detail and quality (e.g., Degtyarenko et al. 2007; Buttigieg et al. 2013). However, if they are not connected via an upper-level ontology, they are disconnected, and with this, they are not interoperable. In other words, they become separated “islands” of knowledge. Integrating these “islands” in order to gain a bigger picture represents the third, as of yet, unsolved problems in ontology engineering. Merging existing ontologies into a single larger one or a specific smaller one is challenging. It comes with many potential conflicts of rather subtle nature (Bench-Capon et al. 1997). These conflicts are including, e.g., that the same terms are used in ontology “A” and “B” while their meaning is differing either by how they are modelled or even only by their text-based definition (c.f. example in Table 3). These problems, however, are unlikely to be resolved automatically shortly. Thus detangling differences between ontologies while reusing and merging the similarities in a new compound ontology remains a manual effort which is a time intensive and error-prone endeavour.

Table 3 Two ontology rudiments in comparison modelling organisms and processes in ecology. In ontology A "Organism" is located in "Living Thing". The "Agent" part is modelled as a separate entity in this representation and organisms are not necessarily expected to be an "Agent" in any process. In ontology "B" all "Organism" are in "Agent" and thus expected to be involved in a biological process.

Ontology A:	Ontology B:
<ul style="list-style-type: none"> • Thing <ul style="list-style-type: none"> ○ Agent ○ Process ○ Living Thing <ul style="list-style-type: none"> ▪ Organism 	<ul style="list-style-type: none"> • Thing <ul style="list-style-type: none"> ○ Process ○ Living Thing <ul style="list-style-type: none"> ▪ Agent ▪ Organism

While the vocabulary which stands behind EASE is not yet modelled as an ontology, the whole framework has been based on a theoretical model. This model represents the perspective of a typical researcher in ecology on data and potential analyses. Thus the framework includes, e.g., the name of variables and if they were measured or manipulated in an experimental setup, their temporal and spatial resolution as well as the environments from which they are originating from as a context (for further details c.f. chapter two). The annotation with EASE is asking for detailed information which might be hidden in full-text fields in metadata schemata (if not entirely forgotten as they are not explicitly asked for). It does not only help to preserve valuable bits of information but also makes details explicitly available for the use in data discovery. The information allows narrowing down the results by using a faceted navigation mechanism. The selection helps to find a specific dataset or compatible data more efficiently. While the upper-level ontology problem remains to be resolved on a more global scale, EASE could potentially be based on the OBO foundry and its design principles in the future (Smith et al. 2007). It consists of an upper-level ontology and a collection of ontologies which are based on common principles for the context of life-sciences. This framework could help to finally utilise the vocabulary of EASE beyond faceted navigation, e.g., in

extracting the knowledge hidden in text-based descriptions of ecological metadata or for the advanced integration of ecological data (Michener and Jones 2012).

EASE is not a full-featured metadata schema as it lacks the typical full-text description elements. Thus it cannot and does not want to replace the existing metadata schemata (Fegraus et al. 2005; Holetschek et al. 2012; Wieczorek et al. 2012). Instead, it can be seen as a complement and merging its ideas with established metadata standards like EML (Ecological Metadata Language), ABCD (Access to Biological Collection Data) or DwC (Darwin Core) could be a fruitful task. It would bring together human-readable full-text based documentation with fine-grained attributes and the designed ecological vocabulary of EASE; thus combining proper human-readable documentation with the potential of an advanced discovery.

The EASE framework was published as open source software (<https://bit.ly/2OoWBII>). This publication potentially allows a broader audience and of ecological projects or individuals to test and use the annotation application and the underlying framework. The EASE vocabulary is extendable and new vocabulary created by projects which use the framework would be of particular interest for the improvement of EASE. Sharing and discussing added terms in dedicated events or with the help of an online platform (e.g., the vocabulary service of GFBio) would be valuable for the ecological research community (Weller and Peters 2008). It could help shed light onto the diversity of language which is used across ecology and to spark discussions to develop agreement on terms which are used to communicate scientific findings better. This insight could, e.g., help to mediate or even overcome general problems in communication (e.g., Bush et al. 1997) and the use of clear vocabulary could speed up scientific progress enabling the better integration of existing resources (e.g., datasets) even across sub-disciplines of ecology.

Chapter Three

Ecological projects have grown into large consortia and complex networks which are acting on a global scale to collect information relevant for nature conservation and land management (e.g., Baeten et al. 2013; Bruelheide et al. 2014). Efforts in ecology are most often funded by governmental investments of tax money (e.g., DFG in Germany, NSF in the USA). Different funding mechanisms are provided, e.g. in Germany by the DFG which promote interdisciplinary research and collaboration focusing on the resolution of particular problems over mid- to longterm periods and differing project structures and sizes. These funding schemes include research units and collaborative research centres which provide up to 12 years of funding and research centres which have the final goal to establish themselves as internationally visible research instances bundling competences along a particular focus (e.g. iDiv, c.f. Homepage of the DFG). The public funding of projects in ecology comes along a particular responsibility which is to maximise the value that is produced based on the investment to finally pay back the stakeholders (i.e., societies) in the form of solved problems or knowledge. To control for a project's progress mechanisms are installed at funding agencies like the DFG like, e.g., reaching defined milestones and detailed reporting in between project phases. In order to achieve an optimal output from a project in regards to data and knowledge, two things are essential. It requires a good overview of the project and its resources on the one hand and project management on the other hand which is carefully planning and guiding the project along its lifetime (Atkinson 1999).

However, with the increasing size and complexity of ecological projects, keeping an overview of their resources and internal processes has become an increasing challenge. It is hard to keep an overview of the involved people, their expertise and

collaborations, datasets and collected variables as well as over the publications and the topics which have been covered. Investigating the structures and dynamics of science has been of broad interest for a long time, which has been documented in publications of scientometric analyses (Hood and Wilson 2001). These analyses are dealing with research networks based on publications, emerging trends and the impact of research (Hou, Kretschmer, and Liu 2008; Garfield 2009). Project management however and particularly the analysis of the project's resources and processes so far lead a rather miserable existence behind the scenes of the projects, conducted only by responsible investigators (e.g. to report back to the funding agency).

Chapter three indicates the potential of metadata and other information produced by ecological projects. It shows how it can be utilised in developing instruments which are providing an overview of a project and feedback to researchers. The information can be used not only along with project management but potentially for the evaluation of projects as well. While the instruments that have been shown already allow the examination of ecological projects from a detailed topic-based perspective, they are far from complete. Developing the tools into a more general framework could be driven forward in workshops inviting researchers from ecology and the field of scientometry. They could discuss and agree on what they think is the most helpful or what is further promoting the transparency in a project to help finally maximise its value and output.

Currently, the tools are showing that the BEF-China project finishes its data collection continuously, that information about organisms and chemicals dominate the project and that researchers are slightly separated during their data collection. In publications, however, they are more connected to each other integrating the diverse collection of datasets in order to derive new knowledge. Further, the project reaches a high amount of good impact journals. However, the tools could be extended for example by creating clusters that are based on annotations of the

datasets of separate researchers to suggest potential new collaborations. Another interesting addition could be instruments which are including details about the funding of a project and its expenses. This insight could allow linking the funding and the expenses directly with the diversity of the covered topics and the thematic orientation or with the achieved impacts in the project. Also, the instruments could finally be linked more closely amongst each other including for example an overview about the variables which have been measured versus the ones which have been used in a publication to help further detect unutilized potential.

The graphs which have been created in chapter three are already interesting on their own. However, they are of limited use at the moment as they are based on the numbers of a single project only. Analysing more projects with the same tools could help to increase the interpretability of the results. Finally, this could also provide evidence for general trends and patterns existing in ecological projects. This information might help to provide certain predictability which in turn could be used in project management influencing the overall planning and setup of new projects. The planning could involve decisions to be made on the amount of required funding or the length of the period as well as on the number of workers which is needed (e.g., Do projects covering specific topics need more time, money, workforce?). Implementing a package for the R environment providing the tools could finally provide access to the presented ideas for a broader audience. Further, integrating the tools into data management platforms like BExIS (Lotz et al. 2012) and BEF-Data (Nadrowski et al. 2013) could finally help increasing not only their visibility but also the acceptance, demonstrating the increase of transparency and efficiency within a project, utilizing the tools that researchers are already used to.

Structural Synthesis

The vocabulary in the form of a thesaurus developed along with chapter one had a significant influence on the following chapters. While we failed to construct a well-designed vocabulary from the extracted folksonomies in a bottom-up approach.

The experiences that have been made in the process shaped the top-down vocabulary and the annotation schema developed in chapter two. In that way the first chapter also links with the third one which is applying the designed vocabulary to a real-world scenario, analysing project related resources and processes. While the implementation of EML into the R environment in the first chapter is not covering the full schema, it highlights the potential of such tools and routines in the analysis environment. It is covering access to relevant information during the analysis while promoting better documentation. In that particular context chapter one also links into the third chapter. It provides ideas to help with the documentation of data and the processing and gives an insight into a potentially useful tool to finally improve on the depth and the detail of information which can impact the project evaluation and management. Chapter two is providing the basis which is utilised in chapter three. It develops a standard for structured documentation of ecological data using clearly defined attributes. While the vocabulary embedded into the tool from chapter two is already a good start, it is likely far from complete. Thus in the future, it would be interesting to see an exchange of information between projects picking up the schema contributing to the vocabulary with their specific annotations. While the basic framework can be utilised for a better discovery in a faceted search, chapter three picks up on the idea and extends upon it. It shows that the fine-grained information can be used for tools which can be utilised in the project evaluation and management to finally allow for improving on the project management while generating an increased value by better utilisation of resources.

Conclusion

It is known that collected ecological data has a value which is going beyond its original research idea (Fegraus et al. 2005). Along those lines, it has become more apparent that it is vital to preserve as much data as possible for its reuse in the future (Diepenbroek et al. 2014). The preservation of data depends on proper

documentation along with the life-cycle on the one hand and reliable cyberinfrastructure on the other hand (White et al. 2013). Overall it requires technical and software related solutions as well as institutions which sign responsible for the data curation in the long run (Diepenbroek et al. 2014, or the National Scientific Data Infrastructure, NFDI). While the data curation has been addressed over time in various efforts along with software, infrastructure, and policy; in ecology, the use of metadata and particularly the documentation beyond it are still underutilised. While standardised metadata schemata are existing which are suitable for being used in the context of ecology, they are mainly focused on human-readable, object-based documentation that is capturing, e.g., the content and context of datasets (Fegraus et al. 2005; Holetschek et al. 2012; Wieczorek et al. 2012). With that particular focus, they are addressing a critical aspect of the documentation in ecology, but also only a part of the needs in documentation along with the full-lifecycle of data.

Chapter one targets bridging the gap in documentation between raw research data and the finally derived knowledge. This is important as it allows to track down knowledge to the roots it has been derived from (i.e. data) which finally enables the repeatability, and checks for correctness and quality of results. Chapter one introduces ideas on how to narrow down the gap improving documentation by utilising the R environment and its statistical scripts. The analysis scripts here are functioning as the link between the raw data and the finally derived data products. While storing scripts along with the data is the first step in the right direction, the presented ideas also indicate that there is a need for a far more holistic solution. Storing the documentation with the data in a private repository is particularly problematic as the smaller repositories are typically inaccessible to a broader audience and their information is more likely to get lost (P. Bryan Heidorn 2008). Small research data repositories rather can be seen as a short-term solution for the management of research data, and it should be treated more like a scratch pad. That

means that they are suitable for maintaining the data and its documentation only until a project is ending (e.g. Nadrowski et al. 2013). After that, the repositories are in need of robust counterparts which are taking over responsibility for the data management in the long run (Diepenbroek et al. 2014). The transition of the data to a publisher or particularly into long-term repositories is a time intensive and complicated task. A standardised documentation package could offer a viable solution to enable better interoperability between small research repositories, publishers and the repositories for permanent storage. These packages could include metadata but also the raw research data, data products in the form of tables, graphs, images and the publication in text form. Such a package would require broad discussions and agreement between researchers, publishers and the data repositories bundled in a coordinated effort defining a robust standard. These standards could then serve as the basis to create new data publication and reuse mechanisms built into small data repositories and analysis software as well as on the publisher and data repository side to enable tools which help to check the quality of the documentation and data.

While established metadata schemata offer several relevant attributes which could be utilised in the context of data discovery, e.g., faceted navigation to filter along the dates of the data collection and the author names, they also hide much of the information in full-text (Fegraus et al. 2005). In the EASE framework presented in chapter two, the amount of explicitness has been maximised in favour over using full-text descriptions. Thus, an annotation with EASE allows building filters for the discovery of data along the idea if datasets are compatible with each other or if they are suitable for a particular analysis which is essential for reusing data. The schema comprises, e.g., the temporal resolution of measured variables or the fact if variables have been measured or manipulated in the study. In the future, the existing schemata may find inspiration in EASE. An effort merging ideas of EASE with existing metadata schemata could finally bring together proper human-

readable documentation on the one hand with a framework of carefully selected and well-defined annotation attributes and ecological vocabulary, on the other hand, improving the discoverability of data. While the annotation framework is helpful on its own, modelling the vocabulary of EASE as a proper ontology in the future would allow for new opportunities. The ontology could finally help along with the extraction of meaningful information from full-text in ecological metadata to further improve on the discoverability of data or even to help with the better integration of diverse data.

Chapter three is indicating that metadata of ecological projects is coming along with untapped potential. While the internal information about a project has likely been used by responsible scientists behind the scenes of ecological projects, e.g. to prepare reports for funding agencies or to solve project management related issues; the information has the potential for being used more transparently. Turning the documentation into useful tools to be available for all researchers in a project could finally help to raise the overall awareness in a project about, e.g., underutilised data, options for collaborations or even provide ideas on potential new topics. Such an overview could lead to better utilisation of all available resources in a project (e.g. datasets, variables, collaborations) and along those lines finally increase the overall value gained from the project. As the presented analyses along chapter three are currently limited to a single project only, it would be of interest to extend them across more projects in the future. This could help with increasing the interpretability of the results allowing to use the resulting insights for projections like, e.g., better project management even before a project starts, e.g. guiding decisions on the length of the project phases and the required funding.

While the current data management solutions along the life-cycle of data are on a good way, the creation and use of metadata and other forms of documentation need more attention. Creating proper documentation, however, is a highly time-intensive process. The creation of tools which help with the collection of

documentation in a transparent and unobtrusive form could be key to improve on available details and coverage with important information. New documentation standards and quality checks for data and projects are urgently needed in ecology. Their establishment could finally help with maintaining a sustainable scientific culture along the improved exchange, discovery and reuse of data while also maintaining a transparent record about how our knowledge has been derived.

General Appendix

The EASE XSD

The following series of images highlight a little bit of the structure which stands behind the Essential Annotation Schema for Ecology (XSD). This overview is an addition to allow the reader to understand better how the schema and the application on top of it work together. It also allows to point out the effort which has been put into that framework which cannot be seen from the user-friendly and straightforward surface of the graphical application interface of the annotation application. The overview starts from the top-level container of the XSD schema and walks along the hierarchical structure to succinctly reveal more details. Exemplary one branch of the schema has been selected which is broken down into its components down towards one leaf of the structure which hosts predefined terms from the annotation vocabulary. A full version of the schema encompasses almost 20.000 lines of XSD code which can be found online in following GitHub repository: <https://github.com/cpfaff/ease>.

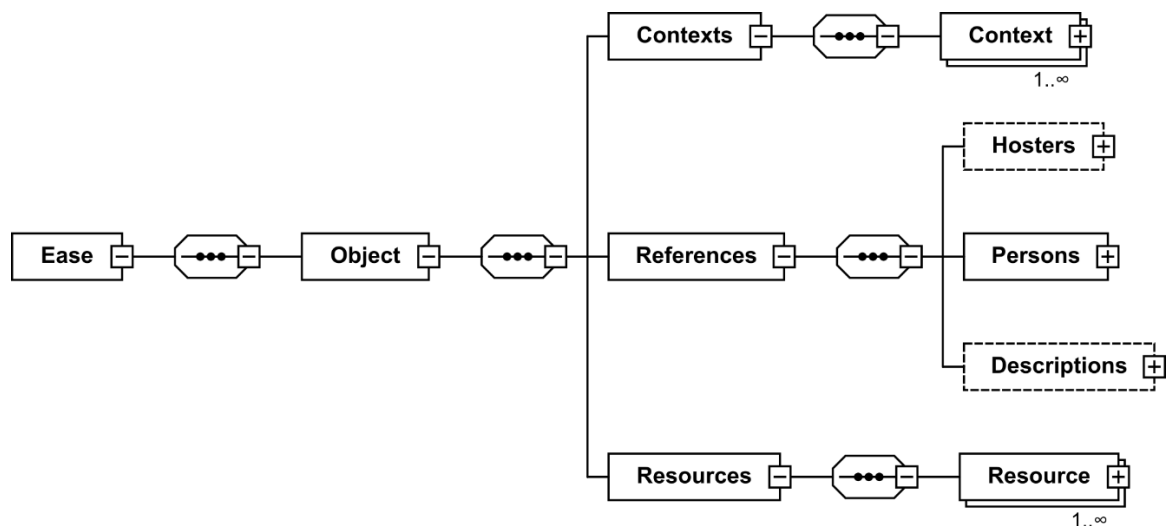


Figure 18 The schematic structure of the EASE standard. The root element of the XSD is named "Ease". This element is comprised of the generic term "Object" which indicates that we describe an EASE object. The object is further separated into three main parts which are "Resources" (it ensures access to the data object described, e.g., by a download URL), "References" (it contains full-text

descriptions and information about institutions and persons) and “Contexts” which is the primary container that stores the faceted annotation.

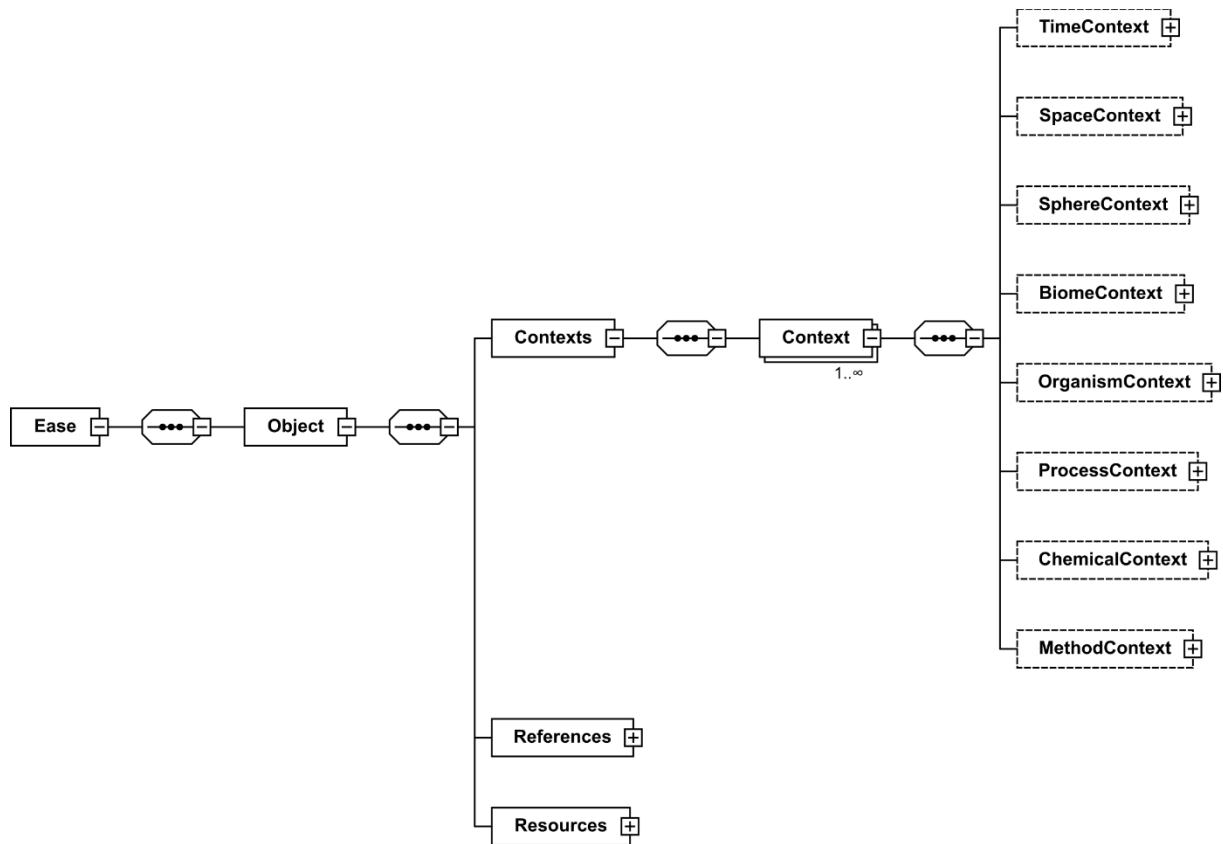


Figure 19 The schematic structure of the EASE standard. The “Context” contains the most important parts for the faceted annotation and the bulk of the vocabulary for the annotation (~1600 concepts). The “Context” unfolds into the areas of time, space, sphere, biome, organism, process, chemical and methods.

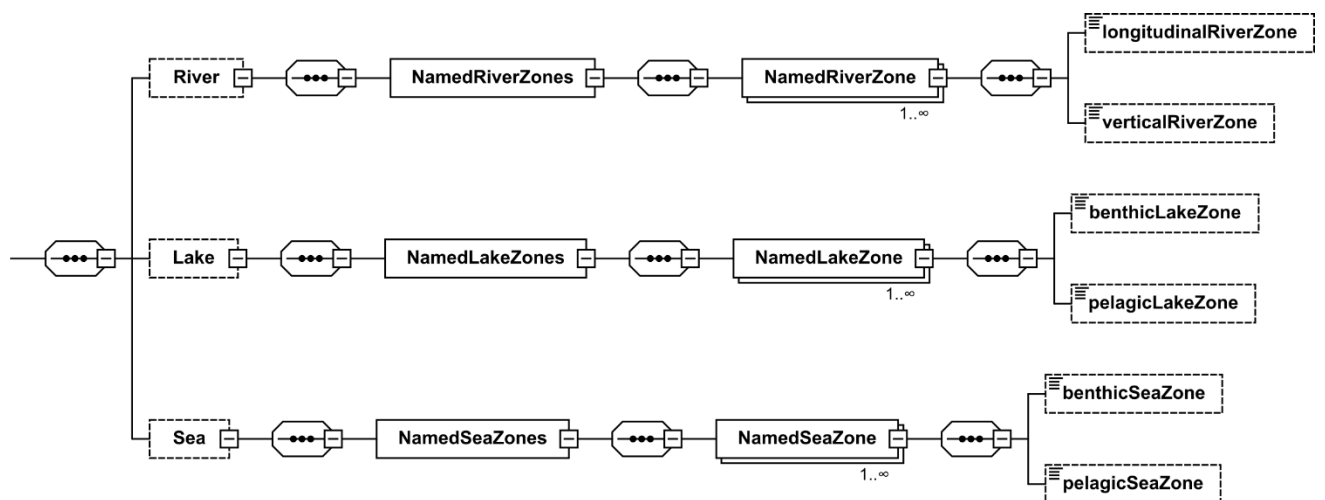


Figure 20 The schematic structure of the EASE standard. Here the “Context” of “Sphere” has been unfolded and in there the “Hydrosphere”. It further divides into aspects related to rivers, lakes and seas wherein it covers the names of zones or areas which allows the localisation or contextualization

to explain from where precisely the samples have been taken. In the user interface this part is also supported with a graphical helper to select from the vocabulary (e.g., in *benthicSeaZone* =, e.g., *Abyssal*, *Hadal*) to speed up the annotation process c.f. Figure 21.

The screenshot shows the gfbio annotation application interface. At the top, there is a navigation bar with 'gfbio' logo and links for 'Dashboard', 'New Annotation Item', 'Data search', and 'Administration'. Below this is a sidebar menu titled 'Annotation item' with icons and labels for 'Common', 'Time', 'Space', 'Sphere', 'Biome', 'Organism', 'Process', 'Chemical', and 'Method'. The 'Sphere' option is selected and highlighted. The main content area is titled 'Sphere' and contains a list of categories: 'Atmosphere', 'Ecosphere', 'Pedosphere', and 'Hydrosphere'. The 'Hydrosphere' category is selected, and a sub-tab 'SEA' is active. Below the tabs is a diagram showing ocean zones. The vertical axis is labeled 'Depth in m' with markers at 200, 1000, 4000, 8000, and 10000. The zones are represented by horizontal bars and checkboxes: Litoral (0-200m), Neritic (0-200m), Bathyal (200-1000m), Abyssal (1000-8000m), Epipelagic (0-200m), Mesopelagic (200-1000m), Bathypelagic (1000-4000m), Abyssopelagic (4000-8000m), Hadopelagic (8000-10000m), Benthic (0-10000m), and Hadal (10000m+). Each term has a question mark icon next to it.

Figure 21 Screenshot of the annotation application. Here it shows the sphere part of the faceted annotation with the hydrosphere. Selected we see the "Sea" part with its visual selection helper to pick from the vocabulary for the annotation of a dataset. Hovering over the question marks will pop up detailed description or definitions (to understand the meaning of the term) and a reference for the source of each term.

Tables chapter two

Table 4 It shows the conceptual topics of time in EASE in relation to how the topics are covered in EML, ABCD and DwC metadata standards (X = not explicitly available as an element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented, e.g., using XSLT transformations.

EASE	EML	ABCD	DwC
The time range for data acquisition with ISO conform start and end date and the time zone (Olson time zone names)	The time range for data acquisition with ISO conform start and end date (coverage module)	X (But a time frame capturing a collection unit identification event)	Time range of a data acquisition event
Geological time frames (International Chronostratigraphic Chart)	Time ranges with an alternative timescale in the coverage module	Geological time frames along bio-, chrono- and lithostratigraphy	Geological context with upper and lower boundaries specifying a geological time frame
Temporal extent (second, minute, ...)	X	X	X
Temporal resolution (second, minute, ...)	X	X	X

Table 5 It shows the conceptual topics for space in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as an element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented, e.g., using XSLT transformations.

EASE	EML	ABCD	DwC
Location name and type (e.g., River, Ocean) as well as the hierarchical relationship to a country and continent (GeoNames)	X (But potentially the location names and the relation information can be provided as the full-text geographic description in the coverage module)	Location name and hierarchical relation as well as a way to specify close by locations	Location name and type in form of specific elements (e.g. island = xxx, country = xxx) , Hierarchical relation of the location
Bounding box and elevation as well as coordinates	Bounding box in decimal degrees, elevation and complex polygons	X (But has a field which allows specifying a download URL for polygon information)	Arbitrary complex polygons in "Well-known Language" markup format
Spatial extent (point, plot, ...)	X	X	X
Spatial resolution (point, plot, ...)	X	X	X

Table 6 It shows the conceptual topics for biomes in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as an element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented, e.g., using XSLT transformations.

EASE	EML	ABCD	DwC
Parameterised biome information, e.g., latitudinal and longitudinal zonation, water availability and physiognomy.	X (But potentially can be provided as the full-text description in the geographic coverage)	X (But captures the Biotope in the context of gathering a collection unit)	X (But captures the Habitat in the context of a data acquisition event)
The condition of biome and land use type	X	X	X

Table 7 It shows the conceptual topics for organisms in EASE in relation to how the topics are covered in the EML, ABCD and DwC metadata standards (X = not explicitly available as an element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented, e.g., using XSLT transformations.

EASE	EML	ABCD	DwC
Taxonomy restricted to elements along the main ranks of the Linnean topology. Scientific species names are captured separately for fungi, viruses, plants, and animals	Taxonomy with a free to specify rank and value for the taxon	Taxonomy with free to specify higher taxon name of the organism. Scientific species names are captured separately for fungi, viruses, plants, animals	Taxonomy along the elements of the main ranks of the Linnean topology and free to define taxonomic classification (e.g., Animalia, Chordata)

Table 8 It shows the conceptual topics for methods in EASE in relation to how the topics are covered in the EML, ABCD, and DwC metadata standards (X = not explicitly available as an element in the schema). This mapping also provides an idea on how future ingestion of information from the schemata to EASE can be implemented, e.g., using XSLT transformations.

EASE	EML	ABCD	DwC
General study approach by type and localisation	X (But allows to specify detailed step by step method protocols in the methods module)	X (But a way to describe a method used to make a collection or observation)	X (But a description of the measurement methods, e.g., a reference to a protocol)
Variables by name and, a modifier that designates if they have been measured or modified	Variables and units and a direct link to tabular data also allowing the detailed description of categories in data	A generic way to specify a measurement or fact including information like, e.g., date and time and a unit of the measurement	The name of a variable, the accuracy and the unit of a measurement

Author Contributions

Dissertation

Claas-Thido Pfaff

A framework to support the annotation, discovery, and evaluation of data in ecology, for better visibility and reuse of data and an increased societal value gained from environmental projects.

Author Contribution Statement

Title: rBEFdata: documenting data exchange and analysis for a collaborative data management platform

Journal: Ecology and Evolution

DOI: 10.1002/ece3.1547

Authors: Claas-Thido Pfaff, Birgitta König-Ries, Anne C. Lang, Sophia Ratcliffe, Christian Wirth, Xingxing Man, Karin Nadrowski

Claas-Thido Pfaff

Designed research, Contributed to the development of the rBEFdata R package, Conceptualized the paper, Wrote the paper, Edited the paper, Submitted the paper

Birgitta König-Ries

Conceptualized the paper, Edited the paper

Anne C. Lang

Analysed data, Conceptualized the paper, Edited the paper

Sophia Ratcliffe

Conceptualized the paper, Edited the paper

Christian Wirth

Designed research, Wrote grant proposal, Conceptualized the paper, Edited the paper

Xingxing Man

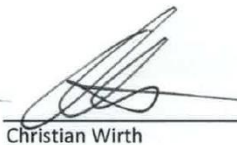
Contributed to the development of the rBEFdata package

Karin Nadrowski

Designed research, Conceptualized the paper, Edited the paper, Contributed to the development of the rBEFdata R package



Claas-Thido Pfaff



Christian Wirth



Karin Nadrowski

Dissertation

Claas-Thido Pfaff

A framework to support the annotation, discovery, and evaluation of data in ecology, for better visibility and reuse of data and an increased societal value gained from environmental projects.

Author Contribution Statement

Title: Essential Annotation Schema for Ecology (EASE)—A framework supporting the efficient data annotation and faceted navigation in ecology

Journal: PLOS ONE

DOI: <https://doi.org/10.1371/journal.pone.0186170>

Authors: Claas-Thido Pfaff, David Eichenberg, Mario Liebergesell, Birgitta König-Ries, Christian Wirth

Claas-Thido Pfaff

Designed research, Created vocabulary, Designed schema and web application, Conceptualized the paper, Wrote the paper, Edited the paper, Submitted the paper

David Eichenberg

Created vocabulary, Conceptualized the paper, Edited the paper

Mario Liebergesell

Created vocabulary, Edited the paper

Birgitta König-Ries

Designed research, Conceptualized paper, Edited the paper

Christian Wirth

Created vocabulary, Designed research, Wrote grant proposal, Conceptualized the paper, Edited the paper



Claas-Thido Pfaff



Christian Wirth

Dissertation

Claas-Thido Pfaff

A framework to support the annotation, discovery, and evaluation of data in ecology, for better visibility and reuse of data and an increased societal value gained from environmental projects.

Author Contribution Statement/Declaration of Authorship (Unpublished Manuscript)

Title: On the evaluation of ecological projects using their metadata

Target Journal: Methods in Ecology and Evolution

Authors: Claas-Thido Pfaff, Helge Bruelheide, David-Eichenberg, Birgitta König-Ries, Christian Wirth

Claas-Thido Pfaff

Herewith I declare my authorship on this unpublished manuscript. The contributions of the other authors are declared separately (see below). My contribution:

Designed research, Conceptualized statistics, Carried out statistics, Conceptualized the manuscript, Wrote the manuscript

Helge Bruelheide

Conceptualized the manuscript, Edited the manuscript

David Eichenberg

Conceptualized statistics, Conceptualized the manuscript, Edited the manuscript

Birgitta König-Ries

Conceptualized statistics, Conceptualized the manuscript, Edited the manuscript

Christian Wirth

Designed research, Conceptualized statistics, Conceptualized the manuscript, Edited the manuscript, Wrote grant proposal



Claas-Thido Pfaff



Christian Wirth

Chaired sessions

1. Juliana Steckel, Claas-Thido Pfaff, and Sophia Ratcliffe. 9th September 2014. Hosting a session at the GfÖ annual meeting in Hildesheim with the topic “Prospective benefits of knowledge transfer in ecology”.
2. Juliana Steckel, Claas-Thido Pfaff, and Michael Owonibi. 31st August – 4th September 2015. Hosting a session at the GfÖ annual meeting 2015 in Göttingen with the topic “Supportive data management tools for integrated ecological studies – best practices and smart services”.
3. Claas-Thido Pfaff and Marco Schmidt. 5th – 9th September 2016. Hosting a session at the GfÖ annual meeting 2016 in Marburg with the topic “Revitalizing the long tail of science – increasing data visibility, access, and fitness for use”.

Workshops

1. Claas-Thido Pfaff and Karin Nadrowski. 7th – 8th September 2013. Hosting a workshop at the GfÖ annual meeting, about “Managing heterogeneous data for ecological analyses”.
2. Claas-Thido Pfaff and Karin Nadrowski. 6th – 7th September 2014. Hosting a workshop at the GfÖ annual meeting, about “Managing heterogeneous data in ecology”.
3. Claas-Thido Pfaff and Julianne Steckel. 29th – 30th August 2015. Hosting a workshop at the GfÖ annual meeting with the title “Data Management Workshop (GFBio)”.
4. Claas-Thido Pfaff (GFBio) and Rudolf May (BfN). 20th May 2015. Hosting a workshop at the meeting of the Network-Forum Biodiversity Germany, about biodiversity informatics in the context of how important environmental data can be preserved and how the visibility and reuse can be increased in the future.

Talks

1. Pfaff, Claas-Thido et al. 11th September 2013. Talk with the topic “rBEFdata, a package for analysing data stored on a BEF-data portal” at the GfÖ annual meeting in the session Management of Ecological Data through its life-cycle.
2. Pfaff, Claas-Thido et al. 23rd January 2014. Talk with the topic “Knowledge Organization Systems in Biology” at the University of Leipzig, Institut für Spezielle Botanik und Funktionelle Biodiversität.
3. Pfaff, Claas-Thido et al. 3rd July 2014. Talk with the topic “The EML R package – Metadata integration into R” at the University of Leipzig, Institut für Spezielle Botanik und Funktionelle Biodiversität.
4. Pfaff, Claas-Thido et al. 9th September 2014. Talk with the topic “The EML metadata package for R” held along the GfÖ annual meeting 2014 in Hildesheim.
5. Pfaff, Claas-Thido et al. 10th December 2014. Invited talk with the topic “The EML R package – Metadata integration into R” held along the R Users Meeting hosted by the Wageningen University.
6. Pfaff, Claas-Thido et al. 8th January 2015. Talk with the topic “Towards a vocabulary to aid data annotation and discovery in ecology” at the University of Leipzig, Institut für Spezielle Botanik und Funktionelle Biodiversität.
7. Pfaff, Claas-Thido et al. 25th June 2015. Talk with the topic “A vocabulary to guide data annotation and discovery in ecology (a status update)” at the University of Leipzig, Institut für Spezielle Botanik und Funktionelle Biodiversität.
8. Pfaff, Claas-Thido et al. 8th June 2016. Talk with the topic “Comprehensive Annotation for Ecology (CAFE)” at the BEXIS 2 developer conference in Jena.

9. Pfaff, Claas-Thido et al. 8th September 2016. Talk with the topic “Comprehensive Annotation for Ecology (CAFE)” at the GfÖ annual meeting 2015 in Marburg.
10. Pfaff, Claas-Thido et al. 15-16th June 2017. Talk with the topic “Essential Annotation Schema for Ecology (EASE): A vocabulary supporting faceted navigation driven data discovery in ecology” at the BEXIS 2 developer conference in Jena.

Posters

1. Pfaff, Claas-Thido et al. 31st August – 4th September 2015. Poster with the title “Towards a vocabulary to aid data annotation and discovery in ecology” at the GfÖ annual meeting 2015 in Göttingen.
2. Pfaff, Claas-Thido et al. 11th – 14th December 2017, Poster with the title “Essential Annotation Schema for Ecology (EASE)” at the Joint BES-GFÖ-Necov and EEF conference held at the ICC in Ghent, Belgium.

Related publications

1. Felicitas Löffler, Claas-Thido Pfaff, Naouel Karam, David Fichtmüller, Friederike Klan: What do Biodiversity Scholars Search for? Identifying High-Level Entities for Biological Metadata. S4BioDiv@ISWC 2017

Curriculum Vitae

Name:

Claas-Thido Pfaff

Aktuelle Position:

Wissenschaftlicher Angestellter

Arbeitgeber:

Institut für Spezielle Botanik und Funktionelle Biodiversität
Universität Leipzig
Johannisallee 21
04103 Leipzig

Seit 2013 bis heute:

Promotionstätigkeit am Institut für Spezielle Botanik und Funktionelle Biodiversität an der Universität Leipzig.

2010-2012:

Master of Science in Biologie, Universität Leipzig, Gesamtnote: 1,5.

Abschlussarbeit.: „The usage of ecological concepts for the extraction of data on the example of a carbon dynamics analysis“ (Note: 1,1).

2007-2010:

Bachelor of Science in Biologie, Universität Hohenheim, Gesamtnote: 1,9.

Abschlussarbeit.: „Biotische und abiotische Einflüsse auf die Verbreitung der Erzwespe *Ixodiphagus hookeri* in Baden-Württemberg“ (Note: 1,0).

Selbstständigkeitserklärung

Ich versichere, die Anforderungen nach §8 (2) der Promotionsordnung der Fakultät für Lebenswissenschaften an der Universität Leipzig eingehalten zu haben.

Insbesondere versichere ich, dass die vorliegende Arbeit ohne unzulässige Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind in der Arbeit als solche kenntlich gemacht worden.

Ich versichere hiermit auch, dass, außer den in der Liste der Co-Autoren genannten Personen, keine weiteren Personen bei der geistigen Herstellung der vorliegenden Arbeit beteiligt waren.

Zudem versichere ich, dass die vorgelegte Arbeit in gleicher oder in ähnlicher Form keiner anderen wissenschaftlichen Einrichtung zum Zwecke einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt oder veröffentlicht wurde.

Leipzig, den _____

Claas-Thido Pfaff _____

Bibliography

- Adams, Jonathan. 2012. "Collaborations: The Rise of Research Networks." *Nature*. Nature Publishing Group. <https://doi.org/10.1038/490335a>.
- Alani, H., Sanghee Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and N.R. Shadbolt. 2003. "Automatic Ontology-Based Knowledge Extraction from Web Documents." *IEEE Intelligent Systems* 18 (1): 14–21. <https://doi.org/10.1109/MIS.2003.1179189>.
- Altintas, I., C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. 2004. "Kepler: An Extensible System for Design and Execution of Scientific Workflows." *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, 423–24. <https://doi.org/10.1109/SSDM.2004.1311241>.
- Anderson, Karen, and Kevin J. Gaston. 2013. "Lightweight Unmanned Aerial Vehicles Will Revolutionize Spatial Ecology." *Frontiers in Ecology and the Environment* 11 (3): 138–46. <https://doi.org/10.1890/120150>.
- Arnqvist, Göran, and David Wooster. 1995. "Meta-Analysis: Synthesizing Research Findings in Ecology and Evolution." *Trends in Ecology & Evolution* 10 (6): 236–40. [https://doi.org/10.1016/S0169-5347\(00\)89073-4](https://doi.org/10.1016/S0169-5347(00)89073-4).
- Atkins, Daniel E, Kelvin K Droegemeier, Stuart I Feldman, Hector García Molina, Michael L Klein, David G Messerschmitt, Paul Messina, et al. 2003. "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure." *Science*, 84. <http://www.nsf.gov/od/oci/reports/atkins.pdf>.
- Atkinson, Roger. 1999. "Project Management: Cost, Time and Quality, Two Best Guesses and a Phenomenon, Its Time to Accept Other Success Criteria." *International Journal of Project Management* 17 (6): 337–42.

[https://doi.org/10.1016/S0263-7863\(98\)00069-6](https://doi.org/10.1016/S0263-7863(98)00069-6).

Attwood, T. K., D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne.

2011. "Utopia Documents: Linking Scholarly Literature with Research Data."

Bioinformatics 27 (13): i568–74. <https://doi.org/10.1093/bioinformatics/btq383>.

Baeten, Lander, Kris Verheyen, Christian Wirth, Helge Bruelheide, Filippo Bussotti,

Leena Finér, Bogdan Jaroszewicz, et al. 2013. "A Novel Comparative Research

Platform Designed to Determine the Functional Significance of Tree Species

Diversity in European Forests." *Perspectives in Plant Ecology, Evolution and*

Systematics 15 (5): 281–91. <https://doi.org/10.1016/J.PPEES.2013.07.002>.

Bar-Ilan, Judit. 2008. "Which H-Index? — A Comparison of WoS, Scopus and

Google Scholar." *Scientometrics* 74 (2): 257–71. [https://doi.org/10.1007/s11192-](https://doi.org/10.1007/s11192-008-0216-y)

[008-0216-y](https://doi.org/10.1007/s11192-008-0216-y).

Beall, Jeffrey. 2008. "The Weaknesses of Full-Text Searching." *Journal of Academic*

Librarianship 34 (5): 438–44. <https://doi.org/10.1016/j.acalib.2008.06.007>.

Bench-Capon, T J M, T J M Bench-Capon, M J R Shave, M J R Shave, Pepijn R S

Visser, Pepijn R S Visser, Dean M Jones, and Dean M Jones. 1997. "Analysis of

Ontology Mismatches." *AAAI Spring Symposium on Ontological Engineering,*

164–72.

Berkley, Chad, Matthew Jones, Jivka Bojilova, and Daniel Higgins. 2001. "Metacat:

A Schema-Independent XML Database System." *Proceedings of the International*

Conference on Scientific and Statistical Database Management, SSDBM, 171–79.

<https://doi.org/10.1109/SSDM.2001.938549>.

Blomquist, Gary J., and Anne Geneviève Bagnères. 2010. *Insect Hydrocarbons*

Biology, Biochemistry, and Chemical Ecology. Insect Hydrocarbons Biology,

Biochemistry, and Chemical Ecology. <https://doi.org/10.1017/CBO9780511711909>.

Boettiger, Carl, Scott Chamberlain, Edmund Hart, and Karthik Ram. 2015.

“Building Software, Building Community: Lessons from the ROpenSci Project.” *Journal of Open Research Software* 3 (1). <https://doi.org/10.5334/jors.bu>.

Borer, Elizabeth T., James B. Grace, W. Stanley Harpole, Andrew S. MacDougall, and Eric W. Seabloom. 2017. “A Decade of Insights into Grassland Ecosystem Responses to Global Environmental Change.” *Nature Ecology and Evolution* 1 (5): 0118. <https://doi.org/10.1038/s41559-017-0118>.

Borgman, Christine L., Jillian C. Wallis, and Noel Enyedy. 2007. “Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries.” *International Journal on Digital Libraries* 7 (1–2): 17–30. <https://doi.org/10.1007/s00799-007-0022-9>.

Bowers, Shawn, Timothy McPhillips, Bertram Ludäscher, Shirley Cohen, and Susan B. Davidson. 2006. “A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows.” In , 133–47. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11890850_15.

Brin, Sergey, and Lawrence Page. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” *Computer Networks and ISDN Systems* 30 (1–7): 107–17. [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X).

Brovkin, Victor, Thomas Raddatz, Christian H. Reick, Martin Claussen, and Veronika Gayler. 2009. “Global Biogeophysical Interactions between Forest and Climate.” *Geophysical Research Letters* 36 (7): 1–5. <https://doi.org/10.1029/2009GL037543>.

Bruelheide, Helge, Karin Nadrowski, Thorsten Assmann, Jürgen Bauhus, Sabine Both, François Buscot, Xiao Yong Chen, et al. 2014. “Designing Forest Biodiversity Experiments: General Considerations Illustrated by a New Large Experiment in Subtropical China.” *Methods in Ecology and Evolution* 5 (1): 74–89. <https://doi.org/10.1111/2041-210X.12126>.

- Buneman, Peter, Sanjeev Khanna, and Tan Wang-Chiew. 2001. "Why and Where: A Characterization of Data Provenance." In , 316–30. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_20.
- Bush, Albert O., Kevin D. Lafferty, Jeffrey M. Lotz, and Allen W. Shostak. 1997. "Parasitology Meets Ecology on Its Own Terms: Margolis et Al. Revisited." *The Journal of Parasitology* 83 (4): 575. <https://doi.org/10.2307/3284227>.
- Buttigieg, Pier, Norman Morrison, Barry Smith, Christopher J Mungall, and Suzanna E Lewis. 2013. "The Environment Ontology: Contextualising Biological and Biomedical Entities." *Journal of Biomedical Semantics* 4 (1): 43. <https://doi.org/10.1186/2041-1480-4-43>.
- Cardinale, Bradley J., J. Emmett Duffy, Andrew Gonzalez, David U. Hooper, Charles Perrings, Patrick Venail, Anita Narwani, et al. 2012. "Biodiversity Loss and Its Impact on Humanity." *Nature* 486 (7401): 59–67. <https://doi.org/10.1038/nature11148>.
- Chowdhury, Gobinda G. 2005. "Natural Language Processing." *Annual Review of Information Science and Technology* 37 (1): 51–89. <https://doi.org/10.1002/aris.1440370103>.
- Collins, Scott L., Luís MA Bettencourt, Aric Hagberg, Renee F. Brown, Douglas I. Moore, Greg Bonito, Kevin A. Delin, et al. 2006. "New Opportunities in Ecological Sensing Using Wireless Sensor Networks." *Frontiers in Ecology and the Environment* 4 (8): 402–7. [https://doi.org/10.1890/1540-9295\(2006\)4\[402:NOIESU\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2006)4[402:NOIESU]2.0.CO;2).
- Darwin, Charles. 1859. *On the Origin of the Species*. Darwin. [https://doi.org/10.1016/S0262-4079\(09\)60380-8](https://doi.org/10.1016/S0262-4079(09)60380-8).
- Deelman, E. 2005. "Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems." *Scientific Programming Journal* 13: 219–

- Degtyarenko, K., P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. 2007. "ChEBI: A Database and Ontology for Chemical Entities of Biological Interest." *Nucleic Acids Research* 36 (Database): D344–50. <https://doi.org/10.1093/nar/gkm791>.
- Díaz, Sandra, Sebsebe Demissew, Julia Carabias, Carlos Joly, Mark Lonsdale, Neville Ash, Anne Larigauderie, et al. 2015. "The IPBES Conceptual Framework - Connecting Nature and People." *Current Opinion in Environmental Sustainability* 14: 1–16. <https://doi.org/10.1016/j.cosust.2014.11.002>.
- Diepenbroek, Michael, Frank Oliver Glöckner, Peter Grobe, Anton Güntsch, Robert Huber, Birgitta König-Ries, Ivaylo Kostadinov, et al. 2014. "Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)." *Informatik 2014 – Big Data Komplexität Meistern. GI-Edition: Lecture Notes in Informatics (LNI) - Proceedings*, 1711–24.
- Egerton, Frank N. 2012. *Roots of Ecology: Antiquity to Haeckel*. University of California Press. https://books.google.de/books?hl=de&lr=&id=hhKyJWPezD0C&oi=fnd&pg=PR9&dq=roots+of+ecology&ots=MVjr-UEAAc&sig=PF1s_tK6vVelOX4RaHpPoIU3KK8#v=onepage&q=roots+of+ecology&f=false.
- Egerton, Frank N. 2001. "A History of the Ecological Sciences, Part 2: Aristotles And Theophrastos." *Bulletin of the Ecological Society of America*.
- English, Jennifer, Marti Hearst, Rashmi Sinha, Kirsten Swearingen, and Ka-ping Yee. 2002. "Flexible Search and Navigation Using Faceted Metadata." *Matrix*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.6.8556&rep=re>

p1&type=pdf.

European Science Foundation. 2008. *Shared Responsibilities in Sharing Research Data: Policies and Partnerships*. www.esf.org.

Fallside, David C, and Priscilla Walmsley. 2004. "XML Schema Part 0: Primer Second Edition." W3C. 2004. <https://doi.org/http://www.w3.org/TR/xmlschema-0/>.

Fecher, Benedikt, Sascha Friesike, and Marcel Hebing. 2015. "What Drives Academic Data Sharing?" *PloS One* 10 (2): e0118053. <https://doi.org/10.1371/journal.pone.0118053>.

Fegraus, Eric H., Sandy Andelman, Matthew B. Jones, and Mark Schildhauer. 2005. "Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation." *Bulletin of the Ecological Society of America* 86 (3): 158–68. [https://doi.org/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2).

Fischer, Markus, Oliver Bossdorf, Sonja Gockel, Falk Hänsel, Andreas Hemp, Dominik Hessenmöller, Gunnar Korte, et al. 2010. "Implementing Large-Scale and Long-Term Functional Biodiversity Research: The Biodiversity Exploratories." *Basic and Applied Ecology* 11 (6): 473–85. <https://doi.org/10.1016/J.BAAE.2010.07.009>.

Friederichs, K. 1958. "A Definition of Ecology and Some Thoughts About Basic Concepts." *Ecology* 39 (1): 154. <https://doi.org/10.2307/1929981>.

Garfield, Eugene. 2009. "From the Science of Science to Scientometrics Visualizing the History of Science with HistCite Software." *Journal of Informetrics* 3 (3): 173–79. <https://doi.org/10.1016/J.JOI.2009.03.009>.

Giles, Jeremy R.A. 2011. "Geoscience Metadata—No Pain, No Gain." In , 29–33. [https://doi.org/10.1130/2011.2482\(03\)](https://doi.org/10.1130/2011.2482(03)).

- Greenberg, Jane. 2005. "Understanding Metadata and Metadata Schemes." *Cataloging & Classification Quarterly* 40 (3–4): 17–36. https://doi.org/10.1300/J104v40n03_02.
- Groot, Rudolf de, Luke Brander, Sander van der Ploeg, Robert Costanza, Florence Bernard, Leon Braat, Mike Christie, et al. 2012. "Global Estimates of the Value of Ecosystems and Their Services in Monetary Units." *Ecosystem Services* 1 (1): 50–61. <https://doi.org/10.1016/J.ECOSER.2012.07.005>.
- Hackett, Edward J., John N. Parker, David Conz, Diana Rhoten, and Andrew Parker. 2008. "Ecology Transformed: The National Center for Ecological Analysis and Synthesis and the Changing Patterns of Ecological Research." In *Scientific Collaboration on the Internet*, 277–96. The MIT Press. <https://doi.org/10.7551/mitpress/9780262151207.003.0016>.
- Haeckel, E. 1866. *Generelle Morphologie Der Organismen. Allgemeine Grundzüge Der Organischen Formen-Wissenschaft, Mechanisch Begründet Durch Die von Charles Darwin Reformirte Descendenztheorie*. G. Reimer.
- Healey, Peter, Harry Rothman, and Paul K. Hoch. 1986. "An Experiment in Science Mapping for Research Planning." *Research Policy* 15 (5): 233–51. [https://doi.org/10.1016/0048-7333\(86\)90024-7](https://doi.org/10.1016/0048-7333(86)90024-7).
- Hearst, Marti. 2008. "UIs for Faceted Navigation: Recent Advances and Remaining Open Problems." *International Journal of Machine Learning and Computing* 1 (4): 337–343. <http://research.microsoft.com/en-us/um/people/ryenw/hcir2008/doc/H CIR08-Proceedings.pdf>.
- Higgins, D., C. Berkley, and M.B. B. Jones. 2002. "Managing Heterogeneous Ecological Data Using Morpho." *Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM 2002–Janua*: 69–76. <https://doi.org/10.1109/SSDM.2002.1029707>.

- Hirsch, J E. 2005. "An Index to Quantify an Individual's Scientific Research Output." *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569–72. <https://doi.org/10.1073/pnas.0507655102>.
- Hobbie, JOHN E., STEPHEN R. CARPENTER, NANCY B. GRIMM, JAMES R. GOSZ, and TIMOTHY R. SEASTEDT. 2009. "The US Long Term Ecological Research Program." [Http://Dx.Doi.Org/10.1641/0006-3568\(2003\)053\[0021:TULTER\]2.0.CO;2](Http://Dx.Doi.Org/10.1641/0006-3568(2003)053[0021:TULTER]2.0.CO;2), January. [https://doi.org/10.1641/0006-3568\(2003\)053\[0021:TULTER\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2003)053[0021:TULTER]2.0.CO;2).
- Holetschek, J., G. Dröge, A. Güntsch, and W. G. Berendsohn. 2012. "The ABCD of Primary Biodiversity Data Access." *Plant Biosystems - An International Journal Dealing with All Aspects of Plant Biology* 146 (4): 771–79. <https://doi.org/10.1080/11263504.2012.740085>.
- Hood, William W., and Concepción S. Wilson. 2001. "The Literature of Bibliometrics, Scientometrics, and Informetrics." *Scientometrics*. <https://doi.org/10.1023/A:1017919924342>.
- Hotho, Andreas, Robert Jäschke, Christoph Schmilz, and Gerd Stumme. 2006. "Information Retrieval in Folksonomies: Search and Ranking." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4011 LNCS: 411–26. https://doi.org/10.1007/11762256_31.
- Hou, Haiyan, Hildrun Kretschmer, and Zeyuan Liu. 2008. "The Structure of Scientific Collaboration Networks in Scientometrics." *Scientometrics* 75 (2): 189–202. <https://doi.org/10.1007/s11192-007-1771-3>.
- Hughes, Lesley. 2000. "Biological Consequences of Global Warming: Is the Signal Already Apparent?" *Trends in Ecology & Evolution* 15 (2): 56–61. [https://doi.org/10.1016/S0169-5347\(99\)01764-4](https://doi.org/10.1016/S0169-5347(99)01764-4).

- Jones, Matthew B., Mark P. Schildhauer, O.J. Reichman, and Shawn Bowers. 2006. "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere." *Annual Review of Ecology, Evolution, and Systematics* 37 (1): 519–44. <https://doi.org/10.1146/annurev.ecolsys.37.091305.110031>.
- Kattge, J., S. Díaz, S. LAVOREL, I. C. Prentice, P. Leadley, G. BÖNISCH, E. GARNIER, et al. 2011. "TRY - a Global Database of Plant Traits." *Global Change Biology* 17 (9): 2905–35. <https://doi.org/10.1111/j.1365-2486.2011.02451.x>.
- Kuhnert, Petra M., Tara G. Martin, and Shane P. Griffiths. 2010. "A Guide to Eliciting and Using Expert Knowledge in Bayesian Ecological Models." *Ecology Letters* 13 (7): 900–914. <https://doi.org/10.1111/j.1461-0248.2010.01477.x>.
- Kwa, Chunglin. 2005. "Local Ecologies and Global Science: Discourses and Strategies of the International Geosphere-Biosphere Programme." *Social Studies of Science* 35 (6): 923–50. <https://doi.org/10.1177/0306312705052100>.
- Lamere, Paul. 2008. "Social Tagging and Music Information Retrieval." *Journal of New Music Research* 37 (2): 101–14. <https://doi.org/10.1080/09298210802479284>.
- Landreth, Anthony, and Alcino J Silva. 2013. "The Need for Research Maps to Navigate Published Work and Inform Experiment Planning." *Neuron* 79 (3): 411–15. <https://doi.org/10.1016/j.neuron.2013.07.024>.
- Lang, Anne C., Goddert von Oheimb, Michael Scherer-Lorenzen, Bo Yang, Stefan Trogisch, Helge Bruelheide, Keping Ma, and Werner Härdtle. 2014. "Mixed Afforestation of Young Subtropical Trees Promotes Nitrogen Acquisition and Retention." Edited by Paul Kardol. *Journal of Applied Ecology* 51 (1): 224–33. <https://doi.org/10.1111/1365-2664.12157>.
- Lerner, Barbara S., and Emery R. Boose. 2015. "RDataTracker and DDG Explorer." In , 288–90. Springer, Cham. https://doi.org/10.1007/978-3-319-16462-5_36.
- Lotz, Thomas, Jens Nieschulze, Jörg Bendix, Maik Dobbermann, and Birgitta

- König-Ries. 2012. "Diverse or Uniform? - Intercomparison of Two Major German Project Databases for Interdisciplinary Collaborative Functional Biodiversity Research." *Ecological Informatics* 8 (March): 10–19. <https://doi.org/10.1016/j.ecoinf.2011.11.004>.
- Madin, Joshua, Shawn Bowers, Mark Schildhauer, Serguei Krivov, Deana Pennington, and Ferdinando Villa. 2007. "An Ontology for Describing and Synthesizing Ecological Observation Data." *Ecological Informatics* 2 (3 SPEC. ISS.): 279–96. <https://doi.org/10.1016/j.ecoinf.2007.05.004>.
- Maes, Joachim, Benis Egoh, Louise Willemen, Camino Liqueste, Petteri Vihervaara, Jan Philipp Schägner, Bruna Grizzetti, et al. 2012. "Mapping Ecosystem Services for Policy Support and Decision Making in the European Union." *Ecosystem Services* 1 (1): 31–39. <https://doi.org/10.1016/J.ECOSER.2012.06.004>.
- Mankovskii, Serguei, Martin Gogolla, Susan D. Urban, Suzanne W. Dietrich, Susan D. Urban, Suzanne W. Dietrich, Ming-Hsuan Yang, et al. 2009. "Ontology." In *Encyclopedia of Database Systems*, 1963–65. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_1318.
- Mascardi, Viviana, Valentina Cordì, and Paolo Rosso. 2007. "A Comparison of Upper Ontologies." *Woa*, 55–64. <https://doi.org/10.1.1.107.1689>.
- Michener, William K., and Matthew B. Jones. 2012. "Ecoinformatics: Supporting Ecology as a Data-Intensive Science." *Trends in Ecology and Evolution* 27 (2): 88–93. <https://doi.org/10.1016/j.tree.2011.11.016>.
- Millenium Ecosystem Assessment. 2010. "Ecosystems and Human Well-Being: Biodiversity Synthesis." *Ecosystems*. <https://doi.org/10.1057/9780230625600>.
- Nadrowski, Karin, Sophia Ratcliffe, Gerhard Bönisch, Helge Bruelheide, Jens Kattge, Xiaojuan Liu, Lutz Maicher, et al. 2013. "Harmonizing, Annotating and Sharing Data in Biodiversity-Ecosystem Functioning Research." *Methods in*

Ecology and Evolution 4 (2): 201–5. <https://doi.org/10.1111/2041-210x.12009>.

Neff, Mark William, and Elizabeth A. Corley. 2009. “35 Years and 160,000 Articles: A Bibliometric Exploration of the Evolution of Ecology.” *Scientometrics* 80 (3): 657–82. <https://doi.org/10.1007/s11192-008-2099-3>.

Oren, Eyal, Renaud Delbru, and Stefan Decker. 2006. “Extending Faceted Navigation for RDF Data,” 559–72. https://doi.org/10.1007/11926078_40.

Otte, Evelien, and Ronald Rousseau. 2002. “Social Network Analysis: A Powerful Strategy, Also for the Information Sciences.” *Journal of Information Science* 28 (6): 441–53. <https://doi.org/10.1177/016555150202800601>.

P. Bryan Heidorn. 2008. “Shedding Light on the Dark Data in the Long Tail of Science.” *Library Trends* 57 (2): 280–99. <https://doi.org/10.1353/lib.0.0036>.

Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, and Gregor Hagedorn. 2011. “Pensoft Data Publishing Policies and Guidelines for Biodiversity Data.” *Natural History*.

Pereira, Henrique M, Paul W Leadley, Vânia Proença, Rob Alkemade, Jörn P W Scharlemann, Juan F Fernandez-Manjarrés, Miguel B Araújo, et al. 2010. “Scenarios for Global Biodiversity in the 21st Century.” *Science (New York, N.Y.)* 330 (6010): 1496–1501. <https://doi.org/10.1126/science.1196624>.

Peters, Debra P.C. 2010. “Accessible Ecology: Synthesis of the Long, Deep, and Broad.” *Trends in Ecology & Evolution* 25 (10): 592–601. <https://doi.org/10.1016/J.TREE.2010.07.005>.

Peters, Debra P C, Peter M. Groffman, Knute J. Nadelhoffer, Nancy B. Grimm, Scott L. Collins, William K. Michener, and Michael A. Huston. 2008. “Living in an Increasingly Connected World: A Framework for Continental-Scale Environmental Science.” *Frontiers in Ecology and the Environment* 6 (5): 229–37. <https://doi.org/10.1890/070098>.

- Petrovskii, S., and N. Petrovskaya. 2012. "Computational Ecology as an Emerging Science." *Interface Focus* 2 (2): 241–54. <https://doi.org/10.1098/rsfs.2011.0083>.
- Pfaff, Claas Thido, David Eichenberg, Mario Liebergesell, Birgitta König-Ries, and Christian Wirth. 2017. "Essential Annotation Schema for Ecology (EASE) - A Framework Supporting the Efficient Data Annotation and Faceted Navigation in Ecology." *PLoS ONE* 12 (10): 1–13. <https://doi.org/10.1371/journal.pone.0186170>.
- Pierotti, Raymond. 2010. *Indigenous Knowledge, Ecology, and Evolutionary Biology*. Routledge. <https://doi.org/10.4324/9780203847114>.
- Poisot, Timothee, Ross Mounce, and Dominique Gravel. 2013. "Moving toward a Sustainable Ecological Science: Don't Let Data Go to Waste!" *Ideas in Ecology and Evolution* 6 (2): 11–19. <https://doi.org/10.4033/iee.2013.6b.14.f>.
- Pollack, Julien, and Daniel Adler. 2015. "Emergent Trends and Passing Fads in Project Management Research: A Scientometric Analysis of Changes in the Field." *International Journal of Project Management* 33 (1): 236–48. <https://doi.org/10.1016/J.IJPROMAN.2014.04.011>.
- Porter, John H., Eric Nagy, Timothy K. Kratz, Paul Hanson, Scott L. Collins, and Peter Arzberger. 2009. "New Eyes on the World: Advanced Sensors for Ecology." *BioScience* 59 (5): 385–97. <https://doi.org/10.1525/bio.2009.59.5.6>.
- R Development Core Team. 2015. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing*. <https://doi.org/10.1007/978-3-540-74686-7>.
- R Development Core Team. 2016. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing Vienna Austria*. <https://doi.org/10.1038/sj.hdy.6800737>.
- Ratcliffe, Sophia, Mario Liebergesell, Paloma Ruiz-Benito, Jaime Madrigal

- González, Jose M. Muñoz Castañeda, Gerald Kändler, Aleksi Lehtonen, et al. 2016. "Modes of Functional Biodiversity Control on Tree Productivity across the European Continent." *Global Ecology and Biogeography* 25 (3): 251–62. <https://doi.org/10.1111/geb.12406>.
- Raupach, M. R., P. J. Rayner, D. J. Barrett, R. S. DeFries, M. Heimann, D. S. Ojima, S. Quegan, and C. C. Schmullius. 2005. "Model-Data Synthesis in Terrestrial Carbon Observation: Methods, Data Requirements and Data Uncertainty Specifications." *Global Change Biology* 11 (3): 378–97. <https://doi.org/10.1111/j.1365-2486.2005.00917.x>.
- Reich, Peter B, David Tilman, Forest Isbell, Kevin Mueller, Sarah E Hobbie, Dan F B Flynn, and Nico Eisenhauer. 2012. "Impacts of Biodiversity Loss Escalate through Time as Redundancy Fades." *Science (New York, N.Y.)* 336 (6081): 589–92. <https://doi.org/10.1126/science.1217909>.
- Reichman, O J, Matthew B Jones, and Mark P Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science (New York, N.Y.)* 331 (6018): 703–5. <https://doi.org/10.1126/science.1197962>.
- Rip, A., and J. -P. Courtial. 1984. "Co-Word Maps of Biotechnology: An Example of Cognitive Scientometrics." *Scientometrics* 6 (6): 381–400. <https://doi.org/10.1007/BF02025827>.
- Roche, Dominique G., Loeske E. B. Kruuk, Robert Lanfear, and Sandra A. Binning. 2015. "Public Data Archiving in Ecology and Evolution: How Well Are We Doing?" *PLOS Biology* 13 (11): e1002295. <https://doi.org/10.1371/journal.pbio.1002295>.
- Rüegg, Janine, Corinna Gries, Ben Bond-Lamberty, Gabriel J. Bowen, Benjamin S. Felzer, Nancy E. McIntyre, Patricia A. Soranno, Kristin L. Vanderbilt, and Kathleen C. Weathers. 2014. "Completing the Data Life Cycle: Using Information Management in Macrosystems Ecology Research." *Frontiers in*

- Ecology and the Environment* 12 (1): 24–30. <https://doi.org/10.1890/120375>.
- Ryen W. White, Bill Kules, Steven M. Drucker, m.c. schraefel. 2006. “Supporting Exploratory Search: Introduction.” *Communications of the ACM, Volume 49, Number 4 (2006), Pages 36-39*. <https://doi.org/http://doi.acm.org/10.1145/1121949.1121978>.
- Sala, O E, F S Chapin, J J Armesto, E Berlow, J Bloomfield, R Dirzo, E Huber-Sanwald, et al. 2000. “Global Biodiversity Scenarios for the Year 2100.” *Science (New York, N.Y.)* 287 (5459): 1770–74. <https://doi.org/10.1126/SCIENCE.287.5459.1770>.
- Salton, G. 1980. “Automatic Term Class Construction Using Relevance-A Summary of Work in Automatic Pseudoclassification.” *Information Processing and Management* 16 (1): 1–15. [https://doi.org/10.1016/0306-4573\(80\)90002-3](https://doi.org/10.1016/0306-4573(80)90002-3).
- Savage, Caroline J., and Andrew J. Vickers. 2009. “Empirical Study of Data Sharing by Authors Publishing in PLoS Journals.” Edited by Chris Mavergames. *PLoS ONE* 4 (9): e7078. <https://doi.org/10.1371/journal.pone.0007078>.
- Seglen, P O. 1997. “Why the Impact Factor of Journals Should Not Be Used for Evaluating Research.” *BMJ (Clinical Research Ed.)* 314 (7079): 498–502. <http://www.ncbi.nlm.nih.gov/pubmed/9056804>.
- Silles, Chris A., and Andrew R. Runnalls. 2010. “Provenance-Awareness in R.” In , 64–72. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17819-1_8.
- Silvertown, Jonathan. 2009. “A New Dawn for Citizen Science.” *Trends in Ecology & Evolution* 24 (9): 467–71. <https://doi.org/10.1016/J.TREE.2009.03.017>.
- Simmhan, Yogesh L., Beth Plale, and Dennis Gannon. 2005. “A Survey of Data Provenance in E-Science.” *ACM SIGMOD Record* 34 (3): 31. <https://doi.org/10.1145/1084805.1084812>.

- Singh, Jatinder. 2011. "FigShare." *Journal of Pharmacology & Pharmacotherapeutics* 2 (2): 138–39. <https://doi.org/10.4103/0976-500X.81919>.
- Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration." *Nature Biotechnology* 25 (11): 1251–55. <https://doi.org/10.1038/nbt1346>.
- Strohmaier, Markus, Christian Körner, and Roman Kern. 2012. "Understanding Why Users Tag: A Survey of Tagging Motivation Literature and Results from an Empirical Study." *Web Semantics: Science, Services and Agents on the World Wide Web* 17 (December): 1–11. <https://doi.org/10.1016/J.WEBSEM.2012.09.003>.
- Sy, Mohameth-François, Sylvie Ranwez, Jacky Montmain, Armelle Regnault, Michel Crampes, and Vincent Ranwez. 2012. "User Centered and Ontology Based Information Retrieval System for Life Sciences." *BMC Bioinformatics* 13 Suppl 1 (Suppl 1): S4. <https://doi.org/10.1186/1471-2105-13-S1-S4>.
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." Edited by Cameron Neylon. *PLoS ONE* 6 (6): e21101. <https://doi.org/10.1371/journal.pone.0021101>.
- Touchon, Justin C, and Michael W Mccoy. 2016. "The Mismatch between Current Statistical Practice and Doctoral Training in Ecology" 7 (August): 1–11.
- Trant, J. 2009. "Studying Social Tagging and Folksonomy: A Review and Framework." *Journal of Digital Information* 10 (1): 1–44.
- United Nations. 1992. "8th Convention on Biological Diversity. Rio de Janeiro, 5 June 1992" 2 (June): 214. https://treaties.un.org/doc/Treaties/1992/06/1992060508-44 PM/Ch_XXVII_08p.pdf.
- Venter, J C, M D D Adams, E W W Myers, P W W Li, R J J Mural, G G G Sutton, H

- O O Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51. <https://doi.org/10.1126/science.1058040>.
- Vines, Timothy H., Arianne Y K Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24 (1): 94–97. <https://doi.org/10.1016/j.cub.2013.11.014>.
- Walls, Ramona L., John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, et al. 2014. "Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies." Edited by Vladimir B. Bajic. *PLoS ONE* 9 (3): e89606. <https://doi.org/10.1371/journal.pone.0089606>.
- Weigelt, Alexandra, Elisabeth Marquard, Vicky M. Temperton, Christiane Roscher, Christoph Scherber, Peter N. Mwangi, Stefanie von Felten, et al. 2010. "The Jena Experiment: Six Years of Data from a Grassland Biodiversity Experiment." *Ecology* 91 (3): 930–31. <https://doi.org/10.1890/09-0863.1>.
- Weisser, Wolfgang W., Christiane Roscher, Sebastian T. Meyer, Anne Ebeling, Guangjuan Luo, Eric Allan, Holger Beßler, et al. 2017. "Biodiversity Effects on Ecosystem Functioning in a 15-Year Grassland Experiment: Patterns, Mechanisms, and Open Questions." *Basic and Applied Ecology* 23 (September): 1–73. <https://doi.org/10.1016/J.BAAE.2017.06.002>.
- Weller, Katrin, and Isabella Peters. 2008. "Seeding , Weeding , Fertilizing – Different Tag Gardening Activities for Folksonomy Maintenance and Enrichment Tag Gardening - Revision and Maintenance of Folksonomies." *Proceedings of I-Semantics'08, International Conference on Semantic Systems Proceedings of I-SEMANTICS '08*, 100–117. http://triple-i.tugraz.at/blog/wp-content/uploads/2008/11/13_seeding-weeding-fertilizing.pdf.

- White, Ethan, Elita Baldrige, Zachary Brym, Kenneth Locey, Daniel McGlinn, and Sarah Supp. 2013. "Nine Simple Ways to Make It Easier to (Re)Use Your Data." *Ideas in Ecology and Evolution* 6 (2): 1–10. <https://doi.org/10.4033/iee.2013.6b.6.f>.
- Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard." Edited by Indra Neil Sarkar. *PLoS ONE* 7 (1): e29715. <https://doi.org/10.1371/journal.pone.0029715>.
- Wilkinson, Leland. 2011. "Ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H." *Biometrics*. <https://doi.org/10.1111/j.1541-0420.2011.01616.x>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March): 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Woodward, F. I., M. R. Lomas, and C. K. Kelly. 2004. "Global Climate and the Distribution of Plant Biomes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 359 (1450): 1465–76. <https://doi.org/10.1098/rstb.2004.1525>.
- Wright, H.E., and P.J. Bartlein. 1993. "Reflections on COHMAP." *The Holocene* 3 (1): 89–92. <https://doi.org/10.1177/095968369300300110>.
- Yee, Ka-Ping, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. "Faceted Metadata for Image Search and Browsing." *Proceedings of the Conference on Human Factors in Computing Systems - CHI '03*, no. 5: 401. <https://doi.org/10.1145/642611.642681>.