Explorative Analysis of Heterogeneous, Unstructured, and Uncertain Data

A Computer Science Perspective on Biodiversity Research

C. Beckstein[1], S. Böcker[1], M. Bogdan[2], H. Bruelheide[3,4], H. M. Bücker[1], J. Denzler[1], P. Dittrich[1], I. Grosse[4,5], A. Hinneburg[5], B. König-Ries[1,4], F. Löffler[1], M. Marz[1], M. Müller-Hannemann[5], M. Winter[4] and W. Zimmermann[5]

[1]*Institute for Computer Science, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany*

[2]*Institute of Computer Science, Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany*

[3]*Institute of Biology, Martin Luther University Halle Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany*

[4]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*

[5]*Institute of Computer Science, Martin Luther University Halle Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle (Saale), Germany*

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Keywords: Data Analysis, Biodiversity Informatics, Research Strategy.

Abstract: We outline a blueprint for the development of new computer science approaches for the management and analysis of big data problems for biodiversity science. Such problems are characterized by a combination of different data sources each of which owns at least one of the typical characteristics of big data (volume, variety, velocity, or veracity). For these problems, we envision a solution that covers different aspects of integrating data sources and algorithms for their analysis on one of the following three layers: At the data layer, there are various data archives of heterogeneous, unstructured, and uncertain data. At the functional layer, the data are analyzed for each archive individually. At the meta-layer, multiple functional archives are combined for complex analysis.

1 BIG DATA IN BIODIVERSITY

What is the benefit of biological diversity? What are the follow-up costs of the disappearance of certain species? How much diversity is needed in an ecosystem to maintain its function in the long term? What benefits provides an ecosystem for us humans? What actions must be taken for ecosystem services to guarantee stable provisioning of food or drinking water on a continuing basis? What impact do climate change and land use have on biodiversity? What does it cost us today to take measures for biodiversity protection? What does it cost us tomorrow if we do not do it today?

These and similar questions are increasingly brought into the light of public consciousness due to their enormous socio-economic and social relevance. Given the potential impact of climate change on biodiversity, the evaluation of biodiversity and ecosystem services is of central importance to society. To more effectively protect biodiversity from overexploitation and destruction, it is indispensable to understand the variability of living organisms, their interactions, and their ecological function. Biodiversity research establishes the foundations for the description and modeling of diversity on different scales and thus for the understanding of ecosystem functioning as one of the ultimate goals. To this end, a wealth of data capturing different aspects of diversity is collected and maintained in various large and heterogeneous data archives.

In the ongoing and lively discussion on big data, biodiversity data and their huge potential for the future of mankind have so far been largely neglected within the computer science community (Manyika et al., 2011). This is particularly surprising because biodiversity data has *all* of the standard properties of big data—in particular, volume, variety, velocity, and veracity. Moreover, biodiversity data with its coverage of multiple scales and its high complexity are a big challenge for algorithm and software development in the big data field (Hampton et al., 2013; Marx, 2013; Goff et al., 2011).

The big data problems in biodiversity science are

Beckstein C., Böcker S., Bogdan M., Bruehlheide H., M. Bücker H., Denzler J., Dittrich P., Grosse I., Hinneburg A., König-Ries B., Löffler F., Marz M., Müller-Hannemann M., Winter M. and Zimmermann W..
Explorative Analysis of Heterogeneous, Unstructured, and Uncertain Data - A Computer Science Perspective on Biodiversity Research.
DOI: 10.5220/0005098402510257
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (DATA-2014), pages 251-257
ISBN: 978-989-758-035-2
Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

defined not only by their sheer size and extreme heterogeneity, but also by their complex interaction on different scales. Hence, in this position paper, we argue that it is necessary to develop a new meta-layer of data analysis as a novel methodological aspect of computer science. This meta-layer offers a wide scope of applicability beyond the area of biodiversity research. Similar big data problems arise from the life sciences, economics, and decision support. New innovative techniques for data analysis need to be developed on the meta-layer. In addition, the plausibility of these analyses needs to be evaluated. In particular, the credibility of the individual data sources needs to be assessed carefully. We believe this to be the only way to ensure that complex solutions to complex big data problems are comprehensible for scientists, policy makers, and other stakeholders.

The structure of this position paper is as follows. In Sect. 2, we outline a target application scenario for the meta-layer of data analysis. By summarizing related work in Sect. 3, we give a sketch of the current state of scientific knowledge in this area. In Sect. 4, we draw a blueprint for data analysis of heterogeneous, unstructured, and uncertain data that are typical for biodiversity science and, finally, we draw conclusions in Sect. 5.

2 APPLICATION SCENARIO

Due to global warming, more and more species that live in southern Europe, Asia Minor, and Africa arrive in Germany. Take mosquito species as an illustrating example. Some mosquito species carry diseases from one part of the world to another part where they currently do not occur. How large is the risk that these species and their diseases become established in Germany? Which roles do the existing mosquito species in facilitating or preventing the spread of these diseases? Can we expect that predation will take care of the problem or do we have to find a technical solution, such as developing vaccines?

The answers to these questions are of high social and economic interest. From a computer science perspective, they ask for a solution that is also applicable to other complex big data problems.

Biodiversity data sources available to answer these questions are in transition: They are complemented with data arising from ongoing digitization processes of museum collections. This also includes data originally collected for a different purpose, such as mobile phone photos taken by citizens being publicly available in social networks (“Citizen Science”). Climate data are another part of complex and constantly changing data relevant for biodiversity research. The degree of this change can have a significant impact on the methods used to handle the data. Returning to the above example, linking genome data and climate data with current data collected from social networks on cumulative occurrences of disease-carrying mosquitoes, e.g., by geo-referenced photos from mobile phones combined with automatic determination software (if applicable), may allow better decisions on overarching questions (Graham et al., 2011). Ideally, this is done by a complete quantitative integration of different sets of heterogeneous data and different sets of algorithms for their analysis. The underlying mathematical methods such as hierarchical Bayesian modeling are evolving rapidly. At the same time, algorithms and hardware are now sufficiently developed to make a quantitative integration possible.

A long-term goal of biodiversity research is to bring together a plethora of different big data sources. The data differ in terms of their spatial scale, ranging from the molecular scale such as genomic data to the global scale such as remote sensing data. Different time scales are also relevant such as femtoseconds for molecular dynamics and decades for CO_2 concentrations, or even centuries for changing species distributions or evolutionary processes. In addition, biodiversity data differ in terms of biological structures. They range from relatively homogeneous, uniformly-structured data such as genomic or remote sensing data to very heterogeneous, semi-structured data.

We envision a software infrastructure that allows us to exploratively analyze such complex data distributed over multiple data sources. This architecture must (1) weigh different data sources according to their relative reliability and (2) connect them as required for the question of interest. An important requirement is to enable a fast solution of biodiversity problems when the underlying data change dynamically. To this end, we suggest scientific workflows that, once defined for the solution of a particular biodiversity problem, may be (semi-)automatically re-executed when data are updated or new data sources become available.

Furthermore, a simple adaptation of this workflow allows a seamless integration of new data analysis methods. This functionality is crucial to ensure that decision makers always find support on the basis of current facts and data analysis methods and that no recommendations are given based on outdated data or methods. From a computer science perspective, the challenge is to enable solutions that are independent of the data sources and the evaluation methods. This way, we envision an architecture that is also applica-

ble to similar application domains beyond biodiversity science.

3 RELATED WORK

Due to its interdisciplinary nature, the problem is relevant to both computer science and biodiversity science. In computer science, there is numerous previous work that considers the problem more or less independent from the application domain. Here, semantic web services and in particular semantic annotations are important. However, to date, these approaches perform unsatisfactorily when scaling to larger problems sizes and therefore have not yet gained any important practical significance.

The connection between computer science and the application domain biodiversity science is established by a subdiscipline called biodiversity informatics. This field is concerned with the application of information technologies to improve the management of biodiversity data. In particular, it facilitates the access to biodiversity data archives and performs relevant data analyses. In recent years, this discipline has intensively dealt with the support, integration, and combination of biodiversity data (Hardisty et al., 2013). Due to the high level of difficulty, which results from the previously described data heterogeneity, this integration process is not yet sufficiently advanced. Today, the combination of data in most cases “only” applies to spatial information. Examples include the Global Biodiversity Information Facility (GBIF, www.gbif.de), Species 2000 (www.sp2000.org), and the European Biodiversity Observation Network (EU BON, eubon.eu). GBIF is a portal with 400 million entries on the occurrence of species, Species 2000 plans to build an integrated species list, and the EU project EU BON seeks to increase the interoperability of electronic biodiversity platforms.

Recently, much effort has been made in biodiversity informatics to formalize established functional relationships. This way, several Knowledge Organization Systems (KOS) (Catapano et al., 2011) have emerged. Their aim is to make the semantic relationship between data machine-readable. Unfortunately, these systems are currently not mutually compatible. Information systems in biodiversity research increasingly consider alternative subsequent usages of the data in the collection phase (Nadrowski et al., 2013). However, the list of projects that follow an integrated approach to data analysis is less extensive. Here, the Map of Life project (Jetz et al., 2012), which brought together a variety of different data types and data analyses, has played a pioneering role. For plant research, there is the iPlant Collaborative (iPlant) (Goff et al., 2011), a virtual organization that provides a platform for storing and analyzing large data sets collaboratively. There are also numerous tools for processing individual big data types, such as Trinity (http://trinityrnaseq.sourceforge.net), Velvet (Zerbino and Birne, 2008), segemehl (Hoffmann et al., 2009; Otto et al., 2012), Jstacs (Grau et al., 2012), or NGS read trimming (Hedtke et al., 2014). Further, text mining tools are useful to process text documents of scientific publications that describe generation, collection, reliability and scope of data in biodiversity. Such text-mining tools include implementations for particular tasks of natural language processing (OpenNLP, 2008; Manning et al., 2014), document clustering and topic modeling (McCallum and Mimno, 2002; Gohr et al., 2009; Blei, 2012).

Existing initiatives from biodiversity informatics typically focus on the mere application of information technology techniques, not on their development. Thus, they lack a dedicated focus on fundamental issues of computer science. In the future, however, computer science can be expected to become a driving force for the development of new methods that enable data integration and data analysis in the first place. We believe that the following areas will play a crucial role in the solution of big data problems from biodiversity science: algorithm engineering, bioinformatics, biosystems analysis, data mining, data management, high performance computing, image and signal processing, knowledge representation, machine learning, parallel processing, software engineering, and visualization.

On the political side, there is a great need for direct support in decision processes, as formulated by the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) (Secretary of IPBES, 2012):

> “However, biodiversity and ecosystem services are declining at an unprecedented rate, and in order to address this challenge, adequate local, national and international policies need to be adopted and implemented. To achieve this, decision makers need scientifically credible and independent information that takes into account the complex relationships between biodiversity, ecosystem services, and people. They also need effective methods to interpret this scientific information in order to make informed decisions. The scientific community also needs to understand the needs of decision makers better in order to provide them with the relevant in-

> formation. In essence, the dialogue between the scientific community, governments, and other stakeholders on biodiversity and ecosystem services needs to be strengthened."

Since there are no such techniques available today, we argue that a blueprint for data analysis as suggested in the following section needs to be implemented urgently.

4 DATA ANALYSIS BLUEPRINT

Our main objective is the development of domain-independent techniques from computer science to improve the support of complex big data solutions and their application to biodiversity research. The blueprint that we sketch in this position paper is biased by our own previous work. Of course, there are other options for fulfilling this goal. Our approach is based on a three-layer architecture as shown in Fig. 1.

On the (green) data layer, different data sources are arranged along spatial and biological scales. Techniques of data selection and analysis for an individual big data source can be applied from the (yellow) functional layer. The blue arrows from the data layer to the functional layer identify these requests to a single big data source. The (red) meta-layer, the question answering layer, then allows the combination of different big data sources. It is this layer where higher-level questions are asked and solutions to these problems are returned. The arrows between the functional layer and the question answering layer symbolize these multiple requests that serve to gain combined knowledge from different big data sources or to confirm, refute, or even invent hypotheses.

All requests must be translated into scientific workflows that answer these requests. Partial results from various big data sources can lead to contradictions although the evaluation algorithms operate correctly. To respond to an overall problem, these contradictions must be detected and weighted in a suitable manner on the meta-layer. For this purpose, tools and methods must be developed that support complex, multi-scale research processes on the basis of spatially distributed, heterogeneous, and unstructured big data sources. To the best of our knowledge, there is no other approach that uses this three-layer approach in biodiversity research or in the general field of big data.

An intensive interdisciplinary collaboration between computer scientists and scientists from application domains is essential to consider the three layers in an integrated way. The task of the applied sciences is the formulation of an actual problem, the development of domain-specific solutions, and the evaluation of proof-of-concept implementations. The task of computer science is the generalization of the problems from a specific application science, the development of domain-independent solutions, and the realization of proof-of-concept implementations. In this interdisciplinary collaboration, the following contributions are necessary on the three layers.

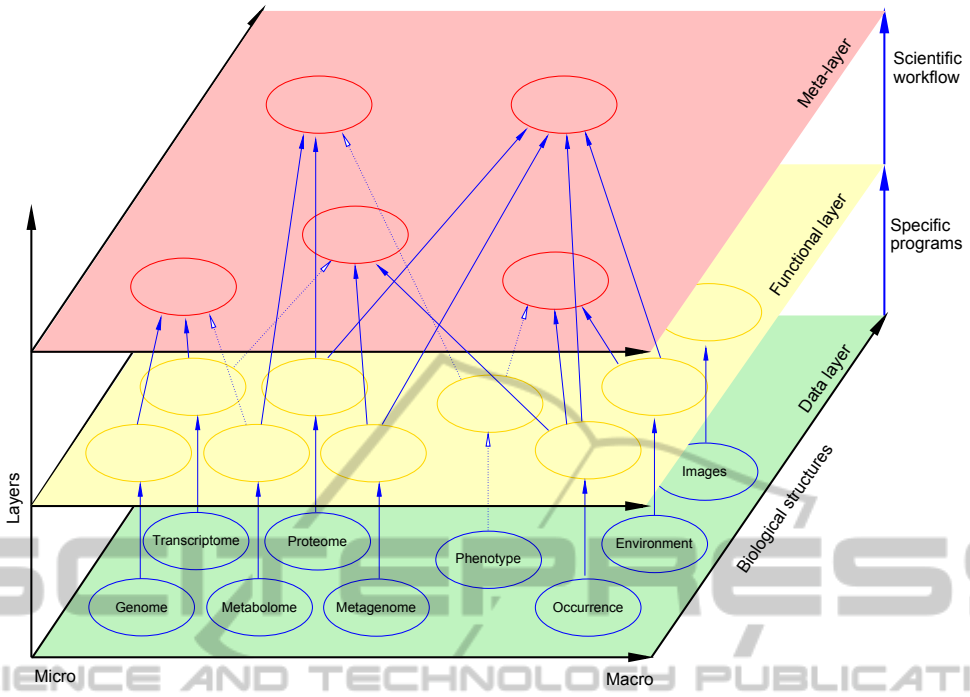
4.1 Meta-layer

On the meta-layer, domain-specific languages are used to formulate a higher-level problem without detailed knowledge of computer science. At the same time an efficient implementation of the question answering layer needs to be developed. Implementations will not only involve model-driven code generation, but rather compiler technology, which is commonly considered to be more powerful (Berg and Zimmermann, 2014). Given our current understanding of application domains of similar complexity as biodiversity science, workflow engines are required that support (i) nested workflows and (ii) packaging of single workflow steps into larger groups that can be automatically combined to build entire families of related analysis software packages.

Solid explanations or justifications of a result include answers to questions like the following: How was the result or a part of it produced? Which data and information were used and how were they used? Why were certain data classified as relevant or irrelevant and therefore explicitly excluded from the analysis? Which assumptions were based on the process and at which point? Typical questions for more complex processes are: How were the results of sub-processes combined to those of the overall process? How well-founded is the result and how does it depend on secondary decision-making processes? How sensitive is a result with respect to a variation of its input parameters? Where and when does a slight change to the input yield a qualitatively different result?

Evaluation methods and data sources may change over time. Whenever this happens, it is necessary to re-evaluate results that were produced before the change. Given that the decision-making process is encoded as a scientific workflow, processes on the meta-layer can (semi-)automatically analyze which parts of the process are effected by the change. On the one hand this allows to generate updated results for big data problems efficiently. On the other hand it is then possible to algorithmically document, visualize, and explain the impact of this change to human decision-makers in a comprehensible way.

Without any doubt, some questions asked on the



SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 1: Schema of a three-layer architecture for explorative analysis of heterogeneous, unstructured, and uncertain data from biodiversity science. Data layer contains originally data and quality information. Existing and novel programs transform extracted knowledge to useable units in the functional layer. A workflow combines a subset of knowledge units to answer a overarching question in the meta-layer.

meta-layer will not be answerable with the data at hand, or only with unacceptable large error margins. One aim of the meta-layer will then be gap identification, informing the functional layer, which type and combination of data would be required to solve the problem at hand. Such identified gaps would then either result in data mining strategies to fill in the required information, or more probably, instigate the collection of new biological primary data.

4.2 Functional Layer

Scientific workflows on the meta-layer can only be executed efficiently if the functional layer provides results from necessary tools for data analysis and also a sufficient annotation of these tools. The data layer consists of a variety of big data sources whose characteristics may vary dramatically. Therefore, there is a need to investigate to what extent known algorithms for small problem sizes are also applicable to large problem sizes. Are there convergence problems with iterative numerical algorithms when increasing the problem size? Is it sufficient to use existing algorithms with an increased number of digits used in floating-point arithmetic, or are completely new algorithms necessary? Which algorithms do not scale to a high number of processors or are inadequate for high-latency networks?

Thus, specific programs are needed to extract knowledge that is then transferred to the meta-layer. In Fig. 1, this transfer from the data layer to the functional layer is represented by blue arrows. These programs might be existing software tools with specific parameters, but we also need novel techniques addressing scalability and reliability (Fortmeier et al., 2013), in particular for algorithms for the identification and calculation of relevant data (Freytag et al., 2012; Hoffmann et al., 2012; Spüler et al., 2012). These techniques are to be developed using modern methods of algorithm engineering (Müller-Hannemann and Schirra, 2010). From these studies one can expect to find generalized results for classes of algorithms. At the interface between the functional and the meta-layer, the formal description of the algorithms and their properties plays a crucial role. Only when algorithms are adequately described, automated reasoning on the meta-layer is possible at all.

4.3 Data Layer

An equally important task is the annotation of the characteristics of data sources and individual data at the data layer. Here, it is necessary to cover both the meaning of data, e.g., by mapping to existing tax-

onomies, as well as the quality of data sources, e.g., its coverage and error rate. It is essential to make use of extensive expertise in the areas of information management for biodiversity data (Lotz et al., 2012), semantic web (Nadrowski et al., 2013), and user-generated annotations (Gohr et al., 2010; Gohr et al., 2011). Due to the sheer size of many data sources it will only be possible to evaluate them efficiently if the evaluating programs are brought to the locations of the data (function shipping) rather than, as usual, copying the data to the respective processing locations. This could be achieved using declarative data description and transformation languages. Examples of such languages are SQL, SparQL, DataLog and Map-Reduce. We envision scientific workflows that compose large data queries and transformation jobs. Using todays database infrastructure, this would result into jobs consisting of multiple dependent SQL queries or Map-Reduce jobs. In the text-mining domain, such architectures are used in the TopicExplorer-System (Hinneburg et al., 2012). For biodiversity science, the spectrum ranges from molecular biological data such as genome, transcriptome, proteome, and metabolome data, to observed data (species, traits, habitats etc.), remote sensing, and climate data.

4.4 Combining the Layers

The building blocks to be developed at the three layers are to be merged into a software architecture that supports scientists in the whole process of knowledge discovery and that records the analysis and the decision-making processes as follows. First, the system searches for the appropriate data, programs, and its parameters. This is followed by finding a suitable combination of multiple data sources and by solving the big data problem. Finally, the system selects a suitable way to visualize the results. During this procedure, all the steps are documented and stored so that they can finally be visualized graphically. This way, the origin of the data and the results are reproducible. The results can be exported for further processing or archiving using different ways. In addition, it is possible to repeat an explorative analysis with minimal human effort and to easily integrate new scientific workflows. Note that explorative visual analysis goes beyond mere result presentation. It allows interactive explanations and justifications of results. Therefore, fast and easy design and composition of visualization views are necessary to provide interfaces to users that allow them to explore analysis results, relate them to observed data and understand their impact.

5 CONCLUDING REMARKS

Biodiversity research has a high societal and economic relevance. Many key questions of this discipline can only be answered using big data. However, up to now, support of big data in this field is limited. Existing approaches address individual aspects, but not the problem as a whole. We believe that an innovative computer science approach is needed here. For this purpose, we proposed a three-layer architecture connecting data sources and function implementations to scientific workflows supporting domain-specific problem solving.

REFERENCES

Berg, C. and Zimmermann, W. (2014). Evaluierung von Möglichkeiten zur Implementierung von Semantischen Analysen für Domänenspezifische Sprachen. In *Software-Engineering 2014, Workshopband Arbeitstagung Programmiersprachen ATPS 2014*, volume 1129, pages 112–128. CEUR Workshop Proceedings.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Catapano, T., Hobern, D., Lapp, H., Morris, R. A., Morrison, N., Noy, N., Schildhauer, M., and Thau, D. (2011). Recommendations for the use of knowledge organization systems by GBIF. Global Biodiversity Information Facility (GBIF), Copenhagen. Available at http://www.gbif.org/orc.

Fortmeier, O., Bücker, H. M., Fagginger Auer, B. O., and Bisseling, R. H. (2013). A new metric enabling an exact hypergraph model for the communication volume in distributed-memory parallel applications. *Parallel Computing*, 39(8):319–335.

Freytag, A., Rodner, E., Bodesheim, P., and Denzler, J. (2012). Rapid uncertainty computation with Gaussian processes and histogram intersection kernels. In *Proc. Asian Conf. Comput. Vis.*, pages 511–524.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W. H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S.-j., Kvilekval, K., Manjunath, B., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S. M., Cranston, K., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D. J., White, J. W., Leebens-Mack, J., Donoghue, M. J., Spalding, E. P., Vision, T. J., Myers, C. R., Lowenthal, D., Enquist, B. J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L., and Stanzione, D. (2011). The iPlant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science*, 2(34).

Gohr, A., Hinneburg, A., Schult, R., and Spiliopoulou, M. (2009). Topic evolution in a stream of documents.

In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30–May 2, 2009, Sparks, Nevada, USA*, pages 859–872. SIAM.

Gohr, A., Hinneburg, A., Spiliopoulou, M., and Usbeck, R. (2011). On the distinctiveness of tags in collaborative tagging systems. In *Proc. Int. Conf. Web Intelligence, Mining and Semantics, WIMS*, page 62. ACM.

Gohr, A., Spiliopoulou, M., and Hinneburg, A. (2010). Visually summarizing the evolution of documents under a social tag. In *Proc. of International Conference on Knowledge Discovery and Information Retrieval, KDIR*.

Graham, E. A., Henderson, S., and Schloss, A. (2011). Using mobile phones to engage citizen scientists in research. *Eos, Transactions American Geophysical Union*, 92(38):313–315.

Grau, J., Keilwagen, J., Gohr, A., Haldemann, B., Posch, S., and Grosse, I. (2012). Jstacs: A Java framework for statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, 13:1967–1971.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W., Budden, A., Batcheller, A., Duke, C., and Porter, J. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162.

Hardisty, A., Roberts, D., et al. (2013). A decadal view of biodiversity informatics: Challenges and priorities. *BMC Ecology*, 13(1):16.

Hedtke, I., Lemnian, I., Müller-Hannemann, M., and Große, I. (2014). On optimal read trimming in next generation sequencing and its complexity. In *Proceedings of AlCoB 2014*, volume 8542 of *LNBI*, pages 83–94. Springer.

Hinneburg, A., Preiss, R., and Schröder, R. (2012). Topicexplorer: Exploring document collections with topic models. In Flach, P. A., Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 838–841. Springer Berlin Heidelberg.

Hoffmann, J., Güttler, F., El-Laithy, K., and Bogdan, M. (2012). Cyfield-RISP: Generating dynamic instruction set processors for reconfigurable hardware using OpenCL. In *Artificial Neural Networks and Machine Learning ICANN 2012*, volume 7552 of *LNCS*, pages 169–176. Springer Berlin Heidelberg.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermüller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comp. Biol.*, 5:e1000502.

Jetz, W., McPherson, J. M., and Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution*, 27(3):151–159.

Lotz, T., Nieschulze, J., Bendix, J., Dobbermann, M., and König-Ries, B. (2012). Diverse or uniform?—Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecological Informatics*, 8:10–19.

Manning, C., Jurafsky, D., and Liang, P. (2014). Stanford CoreNLP. http://nlp.stanford.edu/software/corenlp.shtml. [Online; accessed 19-May-2014].

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Report of the Mc Kinsey Global Institute, Mc Kinsey & Company.

Marx, V. (2013). The big challenges of big data. *Nature*, 498:255–260.

McCallum, A. K. and Mimno, D. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. [Online; accessed 19-May-2014].

Müller-Hannemann, M. and Schirra, S., editors (2010). *Algorithm Engineering: Bridging the Gap between Algorithm Theory and Practice*, volume 5971 of *LNCS*. Springer.

Nadrowski, K., Ratcliffe, S., Bönisch, G., Bruelheide, H., Kattge, J., Liu, X., Maicher, L., Mi, X., Prilop, M., Seifarth, D., Welter, K., Windisch, S., and Wirth, C. (2013). Harmonizing, annotating and sharing data in biodiversity–ecosystem functioning research. *Methods in Ecology and Evolution*, 4(2):201–205.

OpenNLP (2008). Apache OpenNLP. https://opennlp.apache.org/. [Online; accessed 19-July-2008].

Otto, C., Stadler, P. F., and Hoffmann, S. (2012). Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, 28:1698–1704.

Secretary of IPBES (2012). Intergovernmental platform on biodiversity and ecosystem services. Bonn. http://www.ipbes.net/about-ipbes.html.

Spüler, M., Rosenstiel, W., and Bogdan, M. (2012). Adaptive SVM-based classification increases performance of a MEG-based Brain-Computer Interface (BCI). In *ICANN 2012, Part I, LNCS 7552*, pages 669–676.

Zerbino, D. R. and Birne, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18:821–829.