Biodiversity Data Journal 6: e29616
doi: 10.3897/BDJ.6.e29616

Forum Paper

Incentivising use of structured language in biological descriptions: Author-driven phenotype data and ontology production

Hong Cui[‡], James A. Macklin[§], Joel Sachs[§], Anton Reznicek[|], Julian Starr[¶], Bruce Ford[#], Lyubomir Penev[¤], Hsin-Liang Chen[«]

‡ University of Arizona, TUCSON, United States of America
§ Agriculture and Agri-Food Canada, Ottawa, Canada
| University of Michigan, Ann Arbor, United States of America
¶ University of Ottawa, Ottawa, Canada
# University of Manitoba, Winnipeg, Canada
¤ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
« University of Massachusetts at Boston, Boston, United States of America

Corresponding author: Hong Cui (hongcui@email.arizona.edu)

Academic editor: Sarah Faulwetter

Received: 07 Sep 2018 | Accepted: 23 Oct 2018 | Published: 07 Nov 2018

Citation: Cui H, Macklin J, Sachs J, Reznicek A, Starr J, Ford B, Penev L, Chen H (2018) Incentivising use of structured language in biological descriptions: Author-driven phenotype data and ontology production. Biodiversity Data Journal 6: e29616. https://doi.org/10.3897/BDJ.6.e29616

Abstract

Phenotypes are used for a multitude of purposes such as defining species, reconstructing phylogenies, diagnosing diseases or improving crop and animal productivity, but most of this phenotypic data is published in free-text narratives that are not computable. This means that the complex relationship between the genome, the environment and phenotypes is largely inaccessible to analysis and important questions related to the evolution of organisms, their diseases or their response to climate change cannot be fully addressed. It takes great effort to manually convert free-text narratives to a computable format before they can be used in large-scale analyses. We argue that this manual curation approach is not a sustainable solution to produce computable phenotypic data for three reasons: 1) it does not scale to all of biodiversity; 2) it does not stop the publication of free-text phenotypes that will continue to need manual curation in the future and, most

importantly, 3) It does not solve the problem of inter-curator variation (curators interpret/convert a phenotype differently from each other). Our empirical studies have shown that inter-curator variation is as high as 40% even within a single project. With this level of variation, it is difficult to imagine that data integrated from multiple curation projects can be of high quality. The key causes of this variation have been identified as semantic vagueness in original phenotype descriptions and difficulties in using standardised vocabularies (ontologies). We argue that the authors describing phenotypes are the key to the solution. Given the right tools and appropriate attribution, the authors should be in charge of developing a project's semantics and ontology. This will speed up ontology development and improve the semantic clarity of phenotype descriptions from the moment of publication. A proof of concept project on this idea was funded by NSF ABI in July 2017. We seek readers input or critique of the proposed approaches to help achieve community-based computable phenotype data production in the near future. Results from this project will be accessible through https://biosemantics.github.io/author-driven-production.

Keywords

Controlled Vocabulary, Computable Phenotype Data, Data Quality, Phenotype Ontologies

Introduction

Phenotypes are paramount for describing species, studying function and understanding organismal evolution. Recent advancements in computation technology have enabled large-scale, data-driven research, but its full potential has not been realised due to lack of data. High impact research, such as studying trait evolution and its relationship to phylogeny and the environment (e.g. Zanne et al. 2013; Pender 2016), identifying candidate causal genes based on known genotype-phenotype relationships in other taxa (e.g. Edmunds et al. 2015) and resolving taxon names through analysing the relationships between taxonomic concepts with character-based evidence (e.g. Franz et al. 2015; Cui et al. 2016) cannot be realised at this scale without computable phenotype data being available for every clade and taxonomic group.

Textual phenotype descriptions that hold valuable information are continuously being published, yet they are not amenable to computation. When added to the massive amount of phenotype data sitting in older publications, these free-text character descriptions represent a major, under-utilised resource for integrating phenotypic data into modern, large-scale biological research projects that typically involve genomic, climatic and habitat data. These descriptive data are often variable in expression and terminology. Different descriptions of the same character may appear to describe two different traits or two different characters might be interpreted as one. Transforming various natural language expressions into computable data requires a process, called ontologising, where the semantics (meaning) of varied expressions are mapped to terms in an ontology and therefore made explicit (Mabee et al. 2007). An ontology holds a set of well-defined terms

and their relationships, for example, *leaf* and *petiole*, have a relationship: all *petioles* are *part of* some *leaf*. Ontologising ensures “apples are compared to apples” and forms the foundation for meaningful data integration and machine inference and reasoning (i.e. inferring new facts from given facts). For example, if *leafstalk* is equivalent to *petiole*, then all *leafstalks* are *part of* some *leaf* as well.

Currently, making free-text phenotype information computable requires highly trained post-doctoral researchers manually ontologising the descriptions, facilitated by some software applications. However, the manual curation of legacy descriptions is not a sustainable solution for phenotype data production because it does not stop the continued publication of free-text phenotype descriptions that need semantic curation before use. If we assume that each of the estimated 750,000 biomedical papers published in English in 2014 (Ware and Mabe 2015) mentions just one phenotypic character and each character takes about 5 minutes to curate (Dahdul et al. 2015, personal communication with Dahdual), one year’s worth of English biomedical journal publications alone would take a full-time postdoc over 30 years to curate.

Manual curation also does not address the fundamental causes of large (~ 40%) variations in the phenotype data manually curated by different workers (e.g. Cui et al. 2015; Manda et al. in press). This level of variation is concerning because ontologised characters must be highly accurate for computers to produce sound inferences or support data integration. In detailed analyses, two major underlying causes of variation were revealed: incomplete, hard-to-use ontologies and semantic ambiguities in source descriptions (Cui et al. 2015; Huang et al. 2015). Neither of these problems can be adequately addressed by manual curation or text-mining techniques because computers are at their weakest with semantic and pragmatic analyses and cannot be expected to perform better than highly trained humans.

As long as phenotype descriptions continue to be produced as free text, computable phenotype data will remain a major bottleneck holding back large-scale biological research. Given the varied usages of phenotype terms/expressions by different authors and given the fact that the meanings of a term evolve over time, it is evident the semantics of phenotypic characters (categorical or continuous characters) can be most accurately captured at the time of writing by their authors. Any downstream process risks information loss or even misinformation.

Author-Driven Phenotype Data and Ontology Production

We have been awarded funding to investigate a new paradigm of phenotype data production centred on description authors and supported by intuitive software tools to allow them to compose semantically clear descriptions while contributing their vocabularies/ expressions to a shared ontology for their taxon groups. It brings authors to the forefront of ontology construction, promotes clear expressions and exposure of all valid meanings of technical terms and encourages open collaboration and consensus building amongst scientists. While the proposed approach presents a major conceptual change in phenotype

data authoring, the change can be introduced via software environments with which authors are already familiar, for example, Google Docs and Wikis. We will approach the project from the perspectives of social and software engineering, examining human social and collaborative behaviour (e.g. attribution and motivation) and software usability to identify factors that encourage or discourage users from adopting the approach. Although we will start with a test case using the plant genus *Carex* L. ("sedges", family Cyperaceae), the project has the potential to change how biodiversity is described in general and dramatically ease the production of computable phenotype data at a large scale.

Using the ongoing *Carex* revisionary work as the evaluation case for this project is an excellent choice because: 1). *Carex* is one of the largest genera in flowering plants, with close to 2000 species containing considerable variation. 2). A network of *Carex* experts already work closely to prepare the revisions. 3). *Carex* is treated in Flora of North America and Flora of China, from which we have previously extracted over 1200 *Carex* morphological terms and will be used to build the initial *Carex* Phenotype Ontology (CPO) for scientists to improve and 4). Scientists on this project will use the large amount of characters produced from this approach to expand their past research to a scale not possible before (Pender 2016). We are not advocating the creation of more ontologies as randomly creating ontologies will only create new challenges for the end users. What we are arguing is that any phenotypic ontologies created must be *directly* useful to the scientists. If some of these usages are out of the scope of existing ontologies (e.g. in the case of plants, the Planteome Consortium Ontologies), they need to be addressed by more specific domain ontologies, in consultation with the exisiting ontologies. In the *Carex* case, the Author's project and the Planteome project have made a clear roadmap in terms of when to reuse terms and relations from the Plant Ontology and when to create new terms for the *Carex* Ontology. We feel that getting the buy-in at this time from the authors is the most critical mission, while developing successful ontology development strategies, a valuable side product, is of a secondary concern, at least for this project.

We also note that the larger academic and scientific research environment support the premises of the proposed approach. The importance of computable phenotype data is widely recognised and data silos are being actively dissolved. Ontologies and other data publications are valued and attribution methods are being actively examined to credit intellectual contributions to digital resource curation, such as the efforts by the International Society for Curation (http://biocuration.org), OBO Foundry (http://www.obofoundry.org) and THOR (http://project-thor.eu). Publishers like Pensoft are actively seeking and welcoming new methods to stop the continued publication of legacy descriptions. Having years of experience with using digital tools/devices, scientists are expert users of digital collaborative environments (e.g. Wikis, Google Docs). The time is right to investigate a long-term solution to phenotype data production.

Proposed System Design

Fig. 1 illustrates the prototype that will be developed and evaluated in the project. Analogous to Google Doc or Microsoft Word Editor's Spell Checker and personal dictionary, a semantic-aware Description Editor can be used to check semantics using a shared ontology. By making it easy for all authors to add their term usages to a shared ontology and to relate new terms to existing terms in the ontology, a taxon-specific phenotype ontology then comprehensively covers the terms and relationships of the descriptors (i.e. the domain) used by the author community. By revealing how terms are used within a community setting, authors are encouraged to converge to best practices in describing certain characters. This process offers two key benefits: (1) the authors are free to use their terms of choice and (2) author terms are related explicitly to other terms in the ontology so the meaning of the terms is clear. This results in descriptions with clear semantics, making ontologisation of the characters a straightforward step for composing formal statements by harvesting the semantics already expressed in the descriptions. This is then a task that computers can do more efficiently than curators. In addition, the community will quickly have a comprehensive ontology that is tested by use.

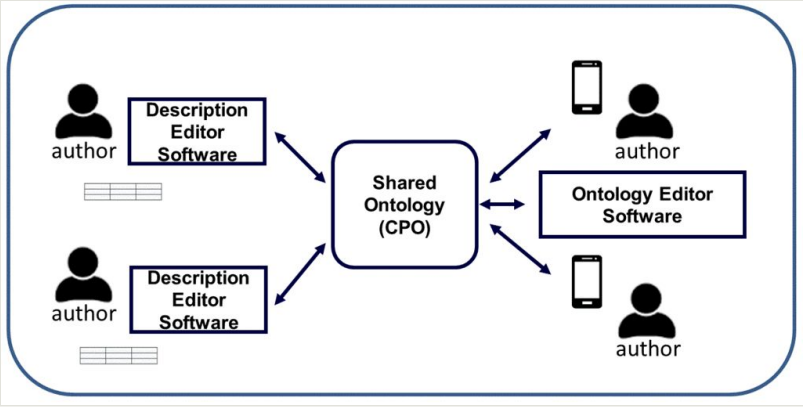

Figure 1. doi

The Integrated Description and Open Collaborative Ontology Editing Platform, with taxon-character matrices by-products. Notice that description authors are also ontology authors.

It is important to differentiate this approach from a standardisation approach where the authors are limited to using a set of "standardised" terms selected by others. The proposed approach does not limit author's choices, but it requires the authors to register the meaning (i.e. semantics) of the terms in their descriptions in an ontology and relate them to other existing terms to allow accurate interpretations in the future. For example, a standardisation approach might require Joe to use the term *strong* when he wishes to say *stout*. In contrast, our approach might show Joe that *stout* has two related but different meanings: *increased size* and *strong (not fragile)*. This would allow Joe to choose the most precise term to use, *increased size*, *strong* or *stout* and, in turn, allow the reader, human or computer, to obtain

the accurate meaning intended by the author. The key idea of the proposed approach is to make all valid meanings of a term clear and visible to a community of users and to encourage the user to filter and choose terms with the most accurate meaning for their purposes.

When the user adds a term to the ontology, the online open Ontology Editor is invoked, presenting different patterns to relate the terms in semantic ways (e.g. assert *utricle in Carex ≡ perigynium in Carex, spike is_a inflorescence, spikelet ≡ secondary spike or small spike, stout ≡strong and increased size, weak ≡decreased magnitude or decreased strength*). Ontology design patterns (e.g.Egaña et al. 2008; Presutti et al. 2012) can be used to wrap the complexity of the logic in a friendly user interface so that users lacking description logic training can use them. For example, non-specific structures, such as apex, surface and base, that can be part of many different structures need to be treated with several logic assertions. Our software can detect cases like this and automatically generate the complete set of assertions for the user to approve (Fig. 2).

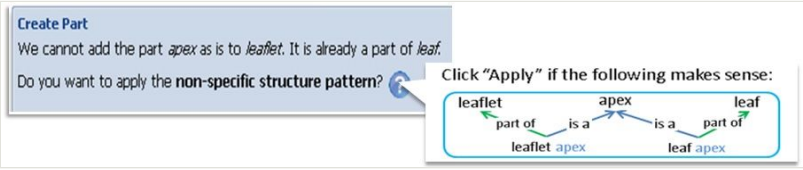

Figure 2. doi

The system detects that the user is attempting to add a substructure (apex) to multiple parent structures (leaf and leaflet). This triggers the system to suggest the non-specific structure pattern to the user. When the user confirms, the system will insert four assertions (4 links in the graph) into the ontology automatically.

These patterns are expected to greatly improve the predictability of the ontology, reduce variation and lower the barrier to entry for biologists. The software will auto-detect situations, whenever possible, for which a pattern may be useful; once the user confirms, the system will carry out what needs to be done on the user's behalf. Fig. 2 illustrates such a scenario for the non-specific structure pattern described above.

Small ontology building tasks such as conflicts amongst term definitions and relationships can be broadcast via a simple mobile app for registered authors to resolve at their leisure. Technical challenging cases can be resolved with help from trained ontology engineers, for example, the Planteome Project (http://planteome.org) or the OBO Foundry.

The rewards to authors who adopt this new workflow include: (1) Narrative descriptions in camera-ready form for publication. (2) A taxon-by-character matrix formulated ontology terms, ready for publication. These can be published in partner journals (e.g.Pensoft journals) in a customisable human readable form (e.g. sentences or matrices) and a variety of new ontologised formats such as EQs (Entity Quality) in the Phenoscape Knowledgebase (http://kb.phenoscape.org) or RDF graphs (Resource Description Framework, a format used widely on the Semantic Web). (3) Formal attributions and

increased citations. On one hand, research has shown that studies that make their data available receive more citations than similar studies that do not (e.g. Piwowar and Vision 2013) and, on the other hand, terms added to the ontology can be linked to the Open Researcher and Contributor ID (ORCID) and the name of the authors and packaged as a micro-publication with a DOI. This could give data consumers another way to include formal data citations in their publications. Even though current data citation practices vary (Robinson-Garcia et al. 2016), the trend is clear as the support for data citations has been widely seen cross disciplines, from science (e.g. Gupta et al. 2017, Cook et al. 2016) to social sciences (e.g. Berez-Kroeker et al. 2017) and from libraries (e.g.Brase et al. 2015) to publishers (e.g.Pavlech 2016). (4) Achievement badges earned based on their contributions within the platform and visible to colleagues.

Results from social and behavioural sciences research on computer mediated collaborative work, online community building and consensus making (e.g. Grudin 1988;Grudin 1994; Innes and Booher 1999; Kriplean et al. 2007; Krieger et al. 2009; Choi et al. 2010; Halfaker et al. 2014; Janssen et al. 2014; Mason et al. 2016) will be implemented to guide the user interaction design of the above-described prototype platform. We acknowledge and have personally witnessed the fact that user participation in open collaborations is often uneven (Wilkinson 2008), but we will strive to design a system where users with different motivations, skill sets and preferences can be engaged in activities that contribute to the overall goal (Preece and Shneiderman 2009;Lampe et al. 2010; Panciera et al. 2010; Wohn et al. 2012; Morgan et al. 2014). While investigating ways to build a strong core of contributors and leaders (Zhu et al. 2012; Luther et al. 2013), steps and design lessons can be taken to integrate and retain new users (Choi et al. 2010; Halfaker et al. 2011; Halfaker et al. 2014; Steinmacher et al. 2015). This project continues our quest to build low barrier software for biologists based on the existing knowledge of what works to encourage open collaboration and consensus making and also contribute to an understanding of the scientific consensus-making process via the new botanical research we plan to conduct with our tools.

Expected Results

We hypothesise that, with careful design of the user interface that takes into account user-friendliness, efficiency, user motivation and other social and behavioural factors, this approach will increase phenotype data quality, ontology quality and computation efficiency.

1. Data quality: improve the semantic clarity of new phenotype descriptions to dramatically reduce the scope of the subsequent ontologisation effort,

2. Ontology quality: quickly improve the coverage of the phenotype ontology for a particular domain (e.g. a taxonomic group) and

3. Computation efficiency: obtain ontologised matrices and/or EQ statements with higher consistency and hence support a wide range of applications.

Assuming this proof of concept system is successful, this approach can be applied to any other science and engineering domains (e.g. biomedical, geology, astrophysics etc.). This being so, individual domain ontologies can be linked, based on shared concepts and terms, thus building powerful bridges for integration across domains, sciences and beyond.

Conclusion

Readers interested in learning more about our project and eventually evaluating our software prototypes can obtain further information from our github project page (https://biosemantics.github.io/author-driven-production) or contact authors. In summary, the goal of this project is to investigate the feasibility of transforming phenotype authors' writing practice to produce computable phenotype data at the time of publication, with increased speed, scale, quality and consistency, while collectively curating phenotype ontologies, making them reflect a community consensus. Through thorough user experience research, we will also identify ways to reduce the entry barrier and promote user adoption of the new practice. When publishers adopt this new idea, we believe the ultimate goal of producing massive high-quality phenotype data for the entire scientific community can be achieved. We seek readers input or critique of the proposed approaches to help achieve community-based computable phenotype data production in the near future.

Grant title

ABI innovation: Authors in the Driver's Seat: Fast, Consistent, Computable Phenotype Data and Ontology Production

Hosting institution

University of Arizona

Author contributions

Cui, Macklin and Sachs contributed the initial idea of the project. All authors edited and reviewed the manuscript.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Berez-Kroeker A, Holton G, Kung S, Pulsifer P (2017) Developing standards for data citation and attribution for reproducible research in linguistics: project summary and next steps. Presentations from the Linguistic Society of America symposium and poster session on Data Citation and Attribution in Linguistics. Austin TX
- Brase J, Lautenschlager M, Sens I (2015) The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite. D-Lib Magazine 21 https://doi.org/10.1045/january2015-brase
- Choi B, Alexander K, Kraut RE, Levine JM (2010) Socialization tactics in Wikipedia and their effects. Proceedings of the ACM Conference on Computer Supported Cooperative Work. Association for Computing Machinery, New York, 107-116 pp.
- Cook R, Vannan SS, McMurry B, Wright D, Wei Y, Boyer A, Kidder JH (2016) Implementation of data citations and persistent identifiers at the ORNL DAAC. Ecological Informatics 33: 10-16. https://doi.org/10.1016/j.ecoinf.2016.03.003
- Cui H, Mande P, Dahdul W, Dececchi A, Ibrahim N, Mabee P, Balhoff J, Gopalakrishnan H (2015) CharaPaser+EQ: Performance Evaluation Without Gold Standard. In: Grove A (Ed.) Proceedings of the 78th Annual Meeting of Association for Information Science and Technology. St. Louis, Missouri
- Cui H, Xu D, Chong S, Ramirez M, Rodenhausen T, Macklin JA, Ludäscher B, Morris RA, Soto EM, Koch NM (2016) Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics 17: 471. https://doi.org/10.1186/s12859-016-1352-7
- Dahdul W, Dececchi TA, Ibrahim N, Lapp H, Mabee P (2015) Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. Database 2015 https://doi.org/10.1093/database/bav040
- Edmunds R, Su B, Balhoff J, Eames BF, Dahdul W, Lapp H, Lundberg J, Vision T, Dunham R, Mabee P, Westerfield M (2015) Phenoscape: Identifying candidate genes for evolutionary phenotypes. Molecular Biology and Evolution 33 (1): 13-24. https://doi.org/10.1093/molbev/msv223
- Egaña M, Rector A, Stevens R, Antezana E (2008) Applying ontology design patterns in bio-ontologies. In: Gangemi A, Euzenat J (Eds) Knowledge Engineering: Practice and Patterns. Springer Berlin, Heidelberg, 7-16 pp.
- Franz N, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B (2015) Reasoning over taxonomic change: Exploring alignments for the *Perelleschus* use case. PLOS ONE 10 (2): e0118247. https://doi.org/10.1371/journal.pone.0118247
- Grudin J (1988) Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. Proceedings of Computer-Supported Cooperative Work. Association for Computing Machinery, New York, 85-93 pp.
- Grudin J (1994) Groupware and social dynamics: Eight challenges for developers. Communications of the ACM 37 (1): 93-104.
- Gupta S, Zabarovskaya C, Romine B, Vianello DA, Hudson Vitale C, McIntosh LD (2017) Incorporating Data Citation in a Biomedical Repository: An Implementation Use Case. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2017: 131-138.

- Halfaker A, Kittur A, Riedl J (2011) Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. Proceedings of the 7th International Symposium on Wikis and Open Collaboration. Association for Computing Machinery, New York, 163-172 pp.
- Halfaker A, Geiger RS, Terveen LG (2014) Snuggle: designing for efficient socialization and ideological critique. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14, 311-320 pp. https://doi.org/10.1145/2556288.2557313
- Huang F, Macklin JA, Cui H, Cole HA, Endara L (2015) OTO: Ontology Term Organizer. BMC Bioinformatics 16: 47. https://doi.org/10.1186/s12859-015-0488-1
- Innes J, Booher D (1999) Consensus Building and Complex Adaptive Systems. Journal of the American Planning Association 65 (4): 412-423. https://doi.org/10.1080/01944369908976071
- Janssen M, der Voort Hv, van Veenstra AF (2014) Failure of large transformation projects from the viewpoint of complex adaptive systems: Management principles for dealing with project dynamics. Information Systems Frontiers 17 (1): 15-29. https://doi.org/10.1007/s10796-014-9511-8
- Krieger M, Stark EM, Klemmer SR (2009) Coordinating tasks on the commons: Designing for personal goals, expertise and serendipity. Proceedings of 27th International Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, 1485-1494 pp.
- Kriplean T, Beschastnikh I, McDonald DW, Golder S (2007) Community, consensus, coercion, control: CS*W or how policy mediates mass participation. Proceedings of GROUP: ACM Conference on Supporting Group Work. Association for Computing Machinery, New York, 167-176 pp.
- Lampe C, Wash R, Velasquez A, Ozkaya E (2010) Motivations to participate in online communities. Proceedings of the ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, 1927-1936 pp.
- Luther K, Fiesler C, Bruckman A (2013) Redistributing leadership in online creative collaboration. Proceedings of the ACM Conference on Computer-Supported Cooperative Work. 1007-1022 pp.
- Mabee P, Ashburner M, Cronk Q, Gkoutos G, Haendel M, Segerdell E, Mungall C, Westerfield M (2007) Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution 22 (7): 345-350. https://doi.org/10.1016/j.tree.2007.03.013
- Manda P, Dahdul W, Cui H, et al. (in press) Annotation of phenotypes using ontologies: a Gold Standard for the training and evaluation of natural language processing systems. Database.
- Mason S, Holley D, Wells A, Jain A, Wuerzer T, Joshi A (2016) An experiment-based methodology to understand the dynamics of group decision making. Socio-Economic Planning Sciences 56: 14-26. https://doi.org/10.1016/j.seps.2016.06.001
- Morgan JT, Gilbert M, McDonald DW, Zachry M (2014) Editing beyond articles: diversity & dynamics of teamwork in open collaborations. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14 https://doi.org/10.1145/2531602.2531654
- Panciera K, Priedhorsky R, Erickson T, Terveen L (2010) Lurking? Cyclopaths? A quantitative lifecycle analysis of user behavior in a geowiki. Proceedings of the ACM

Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, 1917-1926 pp.

- Pavlech L (2016) Data Citation Index. Journal of the Medical Library Association : JMLA 104 (1): 88-90. https://doi.org/10.3163/1536-5050.104.1.020
- Pender J (2016) Climatic niche estimation, trait evolution and species richness in North American *Carex* (Cyperaceae) (M.Sc. Thesis). University of Ottawa
- Piwowar HA, Vision TJ (2013) Data reuse and the open data citation advantage. PeerJ 1 (175): . URL: https://doi.org/10.7717/peerj.175
- Preece J, Shneiderman B (2009) The reader-to-leader framework: Motivating technology mediated social participation. AIS Transactions on Human-Computer Interaction 1 (1): 13-32. https://doi.org/10.17705/1thci.00005
- Presutti V, Blomqvist E, Daga E, Gangemi A (2012) Pattern-Based Ontology Design. In: Suárez-Figueroa MC, Gómez-Pérez A, Motta E, Gangemi A (Eds) Ontology Engineering in a Networked World. Springer, New York, 35-64 pp.
- Robinson-Garcia N, Jimenez-Contreras E, Torres-Salinas D (2016) Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology 67 (12): 2964-2975. https://doi.org/10.1002/asi.23529
- Steinmacher I, Conte T, Gerosa MA, Redmiles D (2015) Social barriers faced by newcomers placing their first contribution in open source software projects. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15 https://doi.org/10.1145/2675133.2675215
- Ware M, Mabe M (2015) The STM Report: An Overview of Scientific and Scholarly Journal Publishing. 4th Edition. International Association of STM Publishers, The Hague, 180 pp.
- Wilkinson DM (2008) Strong regularities in online peer production. Proceedings of the 9th ACM Conference on Electronic Commerce. Association for Computing Machinery, New York, 302-309 pp.
- Wohn D, Velasquez A, Bjornrud T, Lampe C (2012) Habit as an explanation of participation in an online peer-production community. Proceedings of the 2012 ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, 2905-2914 pp.
- Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, Beaulieu JM (2013) Three keys to the radiation of angiosperms into freezing environments. Nature 506 (7486): 89-92. https://doi.org/10.1038/nature12872
- Zhu H, Kraut R, Kittur A (2012) Effectiveness of shared leadership in online communities. Proceedings of the ACM Conference on Computer Supported Cooperative Work. Association for Computing Machinery, New York, 407-416 pp.