Phenotero: Annotate as you write

Daniela Hombach1,2 * | Jana M. Schwarz1 * | Ellen Knierim | Markus Schuelke1 | Dominik Seelow1,2,3 | Sebastian Köhler1,2,3,4

1NeuroCure Clinical Research Center, Charité Universitätsmedizin, Berlin, Germany
2Charité-BIH Centrum für Therapieforschung, Charité Universitätsklinikum, Berlin, Germany
3Berlin Institute of Health (BIH), Berlin, Germany
4Einstein Center Digital Future, Berlin, Germany

Correspondence
Sebastian Köhler, Charité-BIH Centrum für Therapieforschung CC 4, Charité Universitätsmedizin, Charitéplatz 1 10117 Berlin, Germany.
Email: sebastian.koehler@charite.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01GM1608; H2020 European Research Council, Grant/Award Number: 779257; NeuroCure Clinical Research Center, Grant/Award Number: Exc 257

In clinical genetics, the Human Phenotype Ontology as well as disease ontologies are often used for deep phenotyping of patients and coding of clinical diagnoses. However, assigning ontology classes to patient descriptions is often disconnected from writing patient reports or manuscripts in word processing software. This additional workload and the requirement to install dedicated software may discourage usage of ontologies for parts of the target audience. Here we present Phenotero, a freely available and simple solution to annotate patient phenotypes and diseases at the time of writing clinical reports or manuscripts. We adopt Zotero, a citation management software to create a tool which allows to reference classes from ontologies within text at the time of writing. We expect this approach to decrease the additional workload to a minimum while ensuring high quality associations with ontology classes. Standardized collection of phenotypic information at the time of describing the patient allows for streamlining the clinic workflow and efficient data entry. It will subsequently promote clinical and molecular diagnosis with the ultimate goal of better understanding genetic diseases. Thus, we believe that Phenotero eases the usage of ontologies and controlled vocabularies in the field of clinical genetics.

KEYWORDS
clinical text, controlled vocabularies, deep phenotyping, human phenotype ontology, ontology

1 | INTRODUCTION

Information driven approaches are becoming ever more important in biological and medical science. Consequently, biomedical ontologies are used in several research projects and applications.1–4 Ontologies are often defined as a formal specification of a shared conceptualization.5 In research and in the clinic, these ontologies provide a standardized, machine-interpretable way to describe complex biomedical concepts such as human clinical features, animal phenotypes, or gene properties. The application of ontologies is essential for robust information retrieval but also for conducting downstream analyses in the study of diseases, for example, for differential diagnostics and exploration of pathomechanisms.6–8

The Human Phenotype Ontology (HPO),9,10 for instance, is widely used to collect and characterize the symptoms and clinical features of a patient by referencing standardized ontology terms (HPO classes) for individual phenotypic features. To encode diagnoses, other ontologies like Medical Subject Headings (MeSH, www.nlm.nih.gov/mesh/), Disease Ontology (DO),11 or the Monarch Disease Ontology (MONDO, www.obofoundry.org/ontology/mondo.html) are often used.

While the HPO has become the standard to describe patients’ phenotypes in genetic databases and to exchange phenotypic data, medical reports still suffer from a lack of standardization. The journal Cold Spring Harbour Molecular Case Studies has hence already started to require HPO-based descriptions in manuscripts, thus making patients’ descriptions easily recordable, exchangeable, and comparable with other publications, studies, and database contents.

The association between an ontology class (eg, short stature) to an object (eg, a patient) is often termed annotation and requires expert curation. In medicine, physicians have to translate their plain medical terms into standardized ontology entries. This is a highly specialized task, which rises and falls with the physician’s expertise.
both regarding clinical features and the use of ontologies. Current methods used for medical text are dedicated annotation software solutions such as PhenoTips\textsuperscript{12} and Patient Archive (patientarchive.org). Another approach is to retrospectively recognize ontology classes in written text using text mining software, for instance, NCBO annotator\textsuperscript{13} or BioLarK.\textsuperscript{14}

PhenoTips can be used to generate a local patient database and to annotate patients with terms from the HPO. However, such software adds another layer of complexity as they are disconnected from the writing process and require stepping away from word processing tools in order to identify the best matching ontology classes. In addition, they often require nontrivial software installation: PhenoTips, for instance, is optimized to run on a dedicated Linux server, creating a nearly insurmountable obstacle in many clinical work environments. The test version of PhenoTips is available for Windows computers, but is not recommended for production use. Moreover, these tools are aimed at much more complex tasks like the representation of growth charts or pedigree information and are not ideally suited for the mere assignment of ontology-based terms to human phenotypes in a scientific manuscript or clinical report.

Text mining software, on the other hand, is usually run retrospectively and hence also disconnected from the writing process. Moreover, NCBO Annotator or BioLarK require the text to be sent via the internet to an external server, which can cause data security issues, especially when dealing with confidential patient information. This poses a severe problem when it comes to medical reports which routinely include personal and identifying information as names, addresses, and diagnoses. Further, automatic solutions often try to assign ontology classes to all terms or phrases that vaguely resemble a human phenotype, leading to faulty and confusing annotations. Moreover, these tools frequently fail in distinguishing actual patient-related phenotypes from terms related to family history. Tools such as NCBO annotator are also problematic when it comes to negated mentioning of concepts. This makes the resulting data noisy and usually requires subsequent manual curation.

Our goal is to close the gap between writing and annotating to enable physicians and clinical and biomedical researchers to seamlessly associate standardized biological or clinical terms as they write: Our free tool Phenotero allows referencing ontology classes from HPO and MONDO directly within word processing software such as Microsoft Word, LibreOffice, or Google Docs (annotate-as-you-write). Hence, authors of clinical reports or scientific manuscripts can link patient symptoms or diseases to an ontological entry, for example, from the HPO, while they are composing their text; much like adding a reference to a scientific paper. This tremendously eases later usage of information buried in the text because it saves downstream users from having to conduct the association from symptoms to ontology classes over and over again. Moreover, it enables the person who knows the patient best—the physician or researcher familiar with their case—to annotate phenotype terms without delay, leading to high-quality associations.

Phenotero is based on the free, open-source citation software Zotero and thus builds on existing features: a word-processor plugin, standardized and extendible citation styles, a search function, and active community support.

2 | MATERIALS AND METHODS

2.1 | Conversion of HPO and MONDO to libraries

Our primary goal was to keep the user within the Zotero application without the necessity for further external tools. In order to import the HPO and MONDO ontologies into Zotero we provide two strategies.
The first and simpler method is to join the phenotypes and diseases group on the Zotero website (see www.zotero.org/groups/2168222/phenotypes and www.zotero.org/groups/2168493/diseases). Once the user has joined these groups, they are automatically being notified about updates in the database file, which can then be imported to their Zotero working environment with just one click (Zotero sync function) (Figure 1).

A second option is the import via JSON files which we make available for download (phenotero.github.io/data_json/). For this, we transferred the OBO versions of the ontologies into JSON-format (HPO: releases/2018-03-08, MONDO: releases/2018-03-16). Within Zotero, users can then import these files into the library of their choice. The obo2json converter is written in Java and can be obtained from GitHub (https://github.com/phenotero/obo2json). In a digital bibliography, entries must have values for specific fields such as title, author and type. For Phenotero, each ontology class is defined as an "entry-dictionary" type, to better distinguish it from articles, books, etc. This reference-type is rarely used in biomedical articles and should hence not interfere with other Zotero capabilities. The ontology IDs are stored in the container-title field and the synonyms under author, as this field allows having multiple entries per item. The exact mapping between OBO fields and JSON-format for Zotero is described in Table 1. Note that we converted the ontologies to JSON-format in order to create the group libraries described before.

We decided to use MONDO, as this ontology integrates a major fraction of the most important disease ontologies such as MeSH, DO, ICD9, ICD10, NCI Thesaurus, OMIM, and Orphanet.

### 2.2 Definition of Phenotero citation styles

The style of the actual citations and bibliographies or reference lists in the final document is determined by style sheets written in the citation style language (CSL). For most scientific journals a citation style in CSL format is available, which can be easily imported into Zotero and CSL format is available, which can be easily imported into Zotero. Zotero runs on Windows, Mac, and Linux operating systems and the installation is easy and quick (www.zotero.org).

To provide convenient ways of incorporating Phenotero citations and reference lists into word processing software such as Microsoft Office Word, we defined various citation styles: Two stand-alone files, "Phenotero—Phenotypes" (phenotype.csl) and "Phenotero—Phenotypes-IDlist" (phenotype_idlist.csl) are applicable for users who only wish to add phenotype annotations without generating a common bibliography. The style "Phenotero—Phenotypes" lists the ontology class IDs together with their primary label, for example, "[HP:0000767] Pectus excavatum". "Phenotero—Phenotypes-IDlist," which is intended for further downstream use in external tools, only shows the IDs themselves, so that they can be easily copied into other tools (Figure 1).

In addition, we provide a number of style sheets which allow incorporation of phenotype annotations into commonly used journal citation styles. These styles put all articles and other library objects into a standard reference list, but append the references of phenotype and disease ontology classes at the end. In the future, further journal citation styles could be adapted in a similar fashion. All Phenotero csl-files are available for download at phenotero.github.io/data_csl.

### 3 RESULTS

Phenotero is a simple tool for referencing phenotypes and diseases in written text and allows physicians and clinical researchers to readily stay within word processing software. This enables a novel workflow for direct in-text annotations of ontology classes, which we call "annotate as you write."

#### 3.1 Setting up Phenotero

Phenotero requires a running installation of Zotero and the associated Microsoft Office Word or LibreOffice Writer plugin. These plugins are bundled with Zotero from version 4.0 upward and add a Zotero toolbar to Word or Writer. Zotero runs on Windows, Mac, and Linux operating systems and the installation is easy and quick (www.zotero.org).

Following this, users should join the Phenotero Zotero Groups (www.zotero.org/groups/2168222/phenotypes and www.zotero.org/groups/2168493/diseases). Here, they can obtain the latest MONDO and HPO libraries. Zotero will automatically start to synchronize contents of the groups to the local machine. The current Phenotero MONDO library contains 20,503 diseases and it takes around 40 minutes to import it into Zotero. We omitted MONDO text definitions from this library in order to reduce time required for import. The current HPO library contains 13,342 phenotypes and is imported within approximately 20 minutes.

The citation style files in CSL format are available at phenotero.github.io/data_csl and can be imported within seconds. We currently offer seven citation styles, but will make more available upon user request. More information about Phenotero’s installation and usage can be found in Appendix S1 as well as online on the website at phenotero.github.io.

#### 3.2 Use cases

Three simple case examples shall help to highlight the benefit of Phenotero’s usage.
In the first scenario, the phenotype information from a clinical report or scientific manuscript can be reliably extracted after the text has been annotated with Phenotero. This is especially useful for data exchange or further analysis with HPO-based tools, such as PhenGen, eXtasy, PhenIX, or MutationDistiller (http://mutationdistiller.org). The latter is very handy in case of software that lack auto completion to suggest suitable ontology classes matching the entered text. In this scenario, users need to select the Phenotero citation style “Phenotero–Phenotypes-IDlist” and copy-paste the list of HPO-IDs from the bottom of the text into one of the aforementioned tools (see Figure 2 centre).

In the second scenario, a physician or researcher is writing a report about a genetic counselling of a patient. Beyond describing the patient's disease and clinical symptoms, the corresponding sentences are then annotated with HPO references (see Figure 2 left). A recent study has presented that patients appreciate very much a proper description of their medical condition at eye level. Having links to ontologies such as HPO or MONDO helps the patient to find textual definitions, lay person descriptions, and further information about the mentioned phenotypes and diseases.

The third example is based on the assumption that an increasing number of scientific journals are seeking ways to publish manuscripts with references to controlled vocabularies. We assume that for this purpose, HPO or MONDO references shall be integrated into the reference section. This is a prototype implementation, but we expect this to be useful for several biomedical journals, especially those that already request HPO encoded phenotypes to be submitted such as Cold Spring Harbor Molecular Case Studies or Orphanet Journal of Rare Diseases. We provide citation styles for multiple journals (eg, CSH-Laboratory-Press, AJHG, Orphanet Journal of Rare Disease, and Nature), which will append the referenced ontology classes to the reference section (see Figure 2 right).

In all scenarios, the Phenotero approach makes the phenotype and disease information amenable to automated computational analyses of written reports, as the referenced ontology classes can be easily extracted from the text documents. Thus, the framework can also be used to gradually build a patient database consisting of text documents. Text documents become easily indexable, searchable, and amenable to sophisticated clustering and similarity calculation algorithms. For this purpose, we created an executable jar file available (see Phenotero website at https://phenotero.github.io) that processes all docx- and odt-files in a directory and its subdirectories and generates a tsv-file with all extracted ontology identifiers per document. This file can then be used for further analyses.
4 | DISCUSSION

Phenotero provides a quick and easy solution for referencing ontology-based phenotypes and diseases by direct in-text annotations of classes of ontologies such as the HPO or MONDO. It is an extension of the Zotero citation management open source software, which runs on all common operating systems and is easily integrated into multiple word processors. Users are confronted with a minimum of computational complexity and can readily stay within their usual working environment, for example, Microsoft Office Word, LibreOffice Writer, and since autumn 2018 also Google Docs.

The use of Phenotero is very simple and almost identical to writing a manuscript using Zotero. Users can find thorough and user-friendly documentation online at www.zotero.org as well as step-by-step instruction videos on platforms, for example, YouTube. The Phenotero website (https://phenotero.github.io) additionally provides in-depth, Phenotero-specific documentation and short screencasts.

Usage of ontology-based phenotypes in data that is shared with others, such as manuscripts and even physicians’ reports, has many advantages. First of all, a fully described patient’s phenotype leaves less room for uncertainties than full-text descriptions often do. In addition, the phenotypic data becomes computer-readable and can easily be processed in further steps, for example, added to databases or analysed for similarities. Another issue is privacy: The use of ontology-based symptoms allows easy separation of phenotypic from personal information in medical records. This is very important when data is exchanged—complete medical records contain many items to make a patient easily identifiable. A mere list of ontology terms is obviously stripped from all these data and can hence be shared with others without revealing the aforementioned data. This new way of referencing ontologies in written reports does not only provide an easy-to-use tool but also a solution to the important issue of data privacy as no sensitive data has to be uploaded to the internet, for example, to apply text mining.

A potential drawback is that Zotero allows for only one bibliography per document, which may cause problems if users aim for separate bibliographies for publications and ontology references. For this purpose, we found it useful to offer the option of sorted bibliographies which separate references from ontology annotations (see Figure 2 right).

Although Phenotero is currently designed for phenotype and disease ontologies, it can in principle work with any ontology that can be transformed into a digital bibliography. Extensions that enable various annotations from the field of anatomy, cell or gene ontology are principally possible. The Phenotero approach is almost language independent, that is, the clinical reports and manuscripts can be written in any language and still contain references of HPO phenotypes and MonDO disease ontology classes. However, searching the referenced ontology classes requires English, but may be simplified once the ontologies have been completely translated into the target language (see, eg. crowdin.com/project/hpo-translation). Finally, we note that the presented approach is not limited to Zotero but can also work with Endnote (alias: Phendnote), Mendeley (alias: Phendeley), since the libraries, can be exported from Zotero and imported to other reference management software as well. However, we chose Zotero for its ease of use and open-source/free-use policy.

With Phenotero, we hope to get one step closer toward ubiquitous deep phenotyping in a standardized format. We deem this goal beneficial for various downstream tasks on the side of patients as well as of physicians and researchers.

ACKNOWLEDGEMENTS

This work was supported by the E-RARE project Hipbi-RD [01GM1608], the H2020 research project Solve-RD [779257], the “Studienstiftung des deutschen Volkes”, and the NeuroCure Clinical Research Center (Exc 257).

CONFLICTS OF INTEREST

Nothing to declare.

ORCID

Daniela Hombach https://orcid.org/0000-0003-3382-122X
Jana M. Schwarz https://orcid.org/0000-0002-7048-208X
Markus Schuelke https://orcid.org/0000-0003-2824-3891
Dominik Seelow https://orcid.org/0000-0002-9746-4412
Sebastian Köhler https://orcid.org/0000-0002-5316-1399

REFERENCES

1. Bastarache L, Hughey JJ, Hebbing S, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science. 2018;359:1233-1239.
2. Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6(252):252ra123.
3. Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. Semin Perinatol. 2015;39:623-631.
4. Thompson R, Johnston L, Taruscio D, et al. RD-connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014;29(suppl 3):S780-S787.
5. Bost WN. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. The Netherlands: University of Twente; 1997.
6. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
7. Mungall CJ, McMurry JA, Kohler S, et al. The monarch initiative: an integrative platform for analyzing biomedical data. Nucleic Acids Res. 2017;45:D712-D722.
8. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457-464.
9. Köhler S, Doelken SC, Mungall CJ, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42:D966-D974.
10. Köhler S, Vasilevsky NA, Engelstad M, et al. The human phenotype ontology in 2017. Nucleic Acids Res. 2017;45:D865-D876.
11. Kibbe WA, Arce C, Felix V, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43(Database issue):D1071-D1078.
12. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat*. 2013;34(8):1057-1065.

13. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. *Summit Transl Bioinform*. 2009;2009:56-60.

14. Groza T, Köhler S, Doelken S, et al. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database (Oxford)*. 2015;2015:bav005.

15. Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods*. 2014;11(9):935-937.

16. Sifrim A, Popovic D, Tranchevent L-C, et al. eXtasy: variant prioritization by genomic data fusion. *Nat Methods*. 2013;10(11):1083-1084.

17. Vasilevsky NA, Foster ED, Engelstad ME, et al. Plain-language medical vocabulary for precision diagnosis. *Nat Genet*. 2018;50(4):474-476.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Hombach D, Schwarz JM, Knierim E, Schuelke M, Seelow D, Köhler S. Phenetero: Annotate as you write. *Clin Genet*. 2019;95:287–292. [https://doi.org/10.1111/cge.13471](https://doi.org/10.1111/cge.13471)