Evolution of anatomical concept usage over time: Mining 200 years of biodiversity literature

Prashanti Manda
University of North Carolina at Greensboro
and
Todd J. Vision
University of North Carolina at Chapel Hill

The scientific literature contains an historic record of the changing ways in which we describe the world. Shifts in understanding of scientific concepts are reflected in the introduction of new terms and the changing usage and context of existing ones. We conducted an ontology-based temporal data mining analysis of biodiversity literature from the 1700s to 2000s to quantitatively measure how the context of usage for vertebrate anatomical concepts has changed over time. The corpus of literature was divided into nine non-overlapping time periods with comparable amounts of data and context vectors of anatomical concepts were compared to measure the magnitude of concept drift both between adjacent time periods and cumulatively relative to the initial state. Surprisingly, we found that while anatomical concept drift between adjacent time periods was substantial (55% to 68%), it was of the same magnitude as cumulative concept drift across multiple time periods. Such a process, bound by an overall mean drift, fits the expectations of a mean-reverting process.

Categories and Subject Descriptors: []:

■

1. INTRODUCTION

Scientists have been recording hypotheses and discoveries in books and journals for centuries. Biology and natural history has among the deepest historical records, and observations on morphology, anatomy and other properties of biological taxa have been accumulating since the dawn of scientific publication. We would expect our changing understanding of diversity in the natural world to be manifested in the literature through the introduction of novel anatomical and morphological concepts as well as changes in the usage of existing concepts. With access to much of this literature in digital form, we are in a position to quantitatively explore patterns in the evolution of how biological concepts have changed in meaning over time.

In text analysis, the term *concept drift* is used to indicate a change in the context in which a concept is used [Wang et al. 2010] or a change in the statistical properties of a concept as predicted by a machine learning model [Tsymbal 2004; Gama et al. 2014]. Here, we are interested in the evolution of the textual context in which anatomical concepts are used in the literature, as measured by examining the words surrounding controlled vocabulary terms. *Context vectors* are used to represent the frequency of each word in the neighborhood of a focal term among a collection of term occurrences [Carrillo et al. 2009; Gallant 2000]. Context vectors are amenable to statistical comparisons and are routinely used for Natural Language Processing (NLP) applications such as document retrieval, assessing word similarity, and word sense disambiguation [Gallant 2000; Turney et al. 2010].

In recent years, ontologies [Gruber 1993] have gained popularity as a structured and consistent way to describe biological entities [Bard and Rhee 2004; Blake and Bult 2006]. For example, Uberon [Mungall et al. 2012] is a cross-species ontology that represents anatomical concepts connected to each other via relationships that model traditional anatomical classifications. Anatomy ontologies such as Uberon are the result of careful planning and development with respect to concept naming conventions and modeling anatomical classifications making them a valuable repository of anatomical concepts. Given that these ontologies reflect the state of our current knowledge on anatomical concepts, it is interesting to observe how many of these concepts are represented in literature and track how their usage evolves over time.

To observe how concepts drift over time, we segregated published literature into periods of time and compared the context of Uberon anatomical concepts between successive time periods. If context of concepts is found to drift from one time period to the next, that implies that each time period drifts increasingly from the usage of concepts when they were first introduced. We investigate this by comparing the usage of anatomical concepts in each time period to the original context of the concepts.

In addition to being a snapshot of the current knowledge on anatomical concepts, ontologies also encapsulate important information and semantics in their hierarchical structure. For example, concepts in the Uberon ontology are arranged in 15 levels; the greater the distance from the root, the greater the concept's depth. Ontology concepts closer to the root are more general and abstract while concepts closer to the leaves of the ontology are detailed and specific. It could be hypothesized that concepts experience drift differently based on their depth. General concepts could have greater variability in interpretation and usage due to their abstraction leading to greater concept drift. At the same time, detailed and specific concepts at greater depths would be expected to offer little room for use in different contexts and thus show low concept drift. We test this hypothesis by investigating if concept drift decreases with increase in ontology depth.

The Biodiversity Heritage Library (BHL), a consortium of libraries that digitizes biodiversity literature from before the 17th century to the 21st century [Gwinn and Rinaldo 2009] is an excellent source of vertebrate biodiversity literature. Literature in BHL has been used in several studies for taxonomic name recognition [Akella et al. 2012; Wei et al. 2010; Page 2013] and other data mining applications [Thessen et al. 2012]. Parallel to our goal of estimating how anatomical concepts change over time, Page demonstrated the use of BHL content to track changes in scientific names used for the sperm whale over time [Page 2011].

In summary, here, we present an ontology-based temporal data mining analysis on vertebrate biodiversity literature to investigate if the context of anatomical concepts drifts over time and to esti-

mate any concept drift. We conduct ontology-based concept recognition and annotation of literature with concepts from the Uberon ontology and then compute concept drift for the annotated concepts across periods of publication time periods. Specifically, we investigate the following questions 1) How much context drift is there over time? 2) Is context drift cumulative over long time periods? 3) Is the magnitude of concept drift associated the depth of the concept in the ontology?

2. METHODS

2.1 Creating a corpora of literature

A scientific corpora was created for analysis of anatomical concepts by identifying publications that describe vertebrate anatomy. A publication was deemed relevant to vertebrates if 1) the title of the item (book, journal, paper etc.) contained the word "vertebrate(s)", or 2) the item's subject title contained the word "vertebrate(s)". OCR text for the selected literature was downloaded from the BHL database.

2.2 OCR cleaning and filtering

OCR translation sometimes introduces spelling and other formatting errors in the resulting OCR text. We conducted the following cleaning and filtering steps on the downloaded OCR text to account for OCR errors. First, we used an open source software called OCR normalizer [Underwood 2013] to correct common errors such as s/f substitutions, words divided across a line break, spelling normalizing to British or American English, unpacking syncope in eighteenth-century verbs like "remember'd etc. Stop words such as "the, and, etc.", that are not informative to the context of anatomical concepts were also removed. Words in each publication were compared to a dictionary built from English words, abbreviations, scientific families, genera [Best 2013], and concept names from three anatomy ontologies (Uberon [Mungall et al. 2012], Vertebrate Spatial Anatomy Ontology [Dahdul et al. 2012], Teleost Anatomy Ontology [Dahdul et al. 2010]). We then used the percentage of recognized words in the document as an OCR quality score (Q, Equation 1) to quantify the quality of a publication's cleaned OCR text [Tanner et al. 2009].

$$Q = \frac{W_R}{W} \times 100 \tag{1}$$

W_R is the number of recognized words in the document

W is the total number of words in the document

OCR quality scores were computed for each publication and those with $Q \leq 60\%$ were removed from the corpora. The remaining high quality publication corpora were grouped by year to create high quality yearly corpora.

2.3 Concept Recognition

Concept recognition and annotation of anatomical concepts from the Uberon ontology were conducted on the high quality yearly corpora from above. When a piece of text in the corpora matched an Uberon concept name or its exact synonym perfectly, the text piece was annotated to the Uberon concept. For example, the word "snout" was annotated to *UBERON:0006333 (snout)*. If a piece of text and a substring of the text both match different concepts, both annotations were recorded. For example, "head sensillum" results in two annotations – *UBERON:0000963 (head sensillum)* and *UBERON:0000033 (head)*.

Ontology depth was measured relative to the root of the ontology. All direct children of the root have a depth of one, and the depth of a concept is computed as the depth of its parent plus one. If a concept has multiple parents, the depth of the concept is computed as the depth of the deepest parent plus one (*i.e.* the longer path takes precedence).

2.4 Grouping yearly corpora into time periods

A concept was deemed to have sufficient annotations for analysis if the number of annotations was greater than five times the number of time periods. The yearly corpora were further clustered into nine larger time periods such that each time period contained at least 10% of the total annotations. If the limit of annotations in a time period was reached within an yearly corpus, the rest of the yearly corpus was still added to the same time period to avoid splitting yearly corpora into different time periods. This might cause a slight difference in the actual percentage of annotations present in each larger time period.

2.5 Context vector generation

Context vectors represent the context of a word's usage in a corpus by the frequency of the words immediately surrounding it [Carrillo et al. 2009]. The idea is that similarity in neighboring word usage is indicative of semantic similarity.

For a given target word, W, let W_C be the sorted set of surrounding words within a defined window size on either side of W across all instances of W in document D. The context vector ($\vec{C_D}$) for W in D contains the number of times each word occurs in W_C.

$$\vec{C_D} = (O_1, O_2, \cdots, O_N) \tag{2}$$

D: document from which context is computed,

N: number of words in W_C

O_i: occurrence frequency of the ith word in W_C

For each Uberon concept that had at least 45 occurrences summed across all nine corpora, a context vector was calculated for that concept in each time period using a window size of five.

2.6 Computing concept drift of anatomical concepts

We measure concept drift as a change in the context vector for a given ontology concept between time periods, as indicated by the complement of vector cosine similarity [Tan et al. 2006]. Mathematically, the cosine similarity (S) of two context vectors, $\vec{C_1}$ and $\vec{C_2}$, over W_C words is defined as :

$$S(\vec{C_1}, \vec{C_2}) = \frac{\sum_{i=1}^{W_C} O_{i,1} O_{i,2}}{\sqrt{\sum_{i=1}^{W_C} O_{i,1}^2} \sqrt{\sum_{i=1}^{W_C} O_{i,2}^2}} \tag{3}$$

$O_{i,1}$is the occurrence frequency of the ith word in vector $\vec{C_1}$

$O_{i,2}$ is the occurrence frequency of the ith word in vector $\vec{C_2}$

In order to track the evolution of concept usage with time, we compared concept context vectors between each successive time

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2747v1 | CC BY 4.0 Open Access | rec: 25 Jan 2017, publ: 25 Jan 2017

period pair. We also compared concept context vectors between early and later time periods to determine whether concept drift was cumulative.

(1) Comparison between successive time periods: Context vectors of Uberon concepts from one time period were compared with the context vectors from the successive time period to estimate drift between every successive time period pair. In a time period comparison between T_j and T_{j+1}, if a concept present in T_{j+1} is absent in T_j, then the concept's context is compared to its context from the latest time period before T_j where the concept was last observed.

(2) Comparison between time periods and the original context of concepts: Every concept is first observed in some time period; the concept context vector from this time period is called the original context of the concept. Context vectors from each time period are compared to the original context of concepts to estimate how far each time period has drifted from the original context of each concept.

3. RESULTS

3.1 Compiling the corpus, data cleaning, and concept recognition

413 publications relevant to vertebrates with a combined word count of about 40 million were downloaded from the BHL database. The data cleaning steps described in section 2.2 were applied to each publication to correct for OCR translation errors reducing the word count in downloaded corpora by 63.45% (Figure 1a).

OCR quality scores were computed for each cleaned publication corpus and 47 publications with $Q \leq 60\%$ were removed from the analysis. The resulting high quality yearly corpora had quality scores ranging approximately from 61% to 95% (Figure 1) with a mean score of 80.96%. These high quality corpora were annotated with Uberon concepts in the concept recognition step (section 2.3) resulting in a total of 942,923 annotations (Figure 1a).

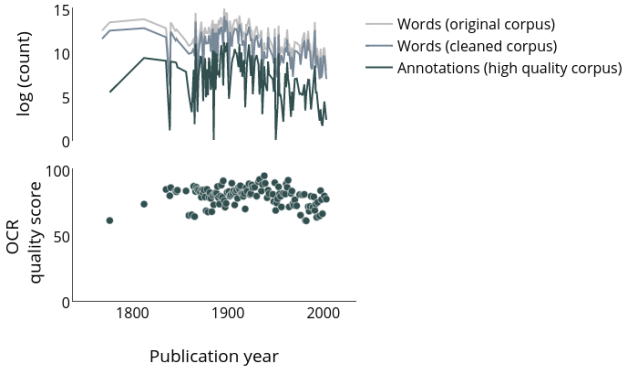


Fig. 1: (a) Comparison of word counts from the original, cleaned corpora and annotation counts from high quality corpora. (b) OCR scores for high quality corpora after filtering corpora with below threshold scores.

2,620 unique Uberon concepts from different depths in the ontology were represented in the 942,923 annotations covering approximately 18% of Uberon concepts (as of 2015-05-26). We found no significant differences in the distributions of concepts from different depths between the Uberon ontology and BHL annotations (two-sample Kolmogorov-Smirnov test, $p = 0.67$, using Bonferroni correction for multiple tests)

3.2 Grouping yearly corpora into larger time periods

Nine time periods spanning from 1776 to 2003 were created by grouping yearly corpora into larger time periods that contain at least 10% of the total annotations (Table I). 799 Uberon concepts were found to have sufficient annotations across these time periods.

Table I. : Time periods and years covered in time period from grouping yearly corpora

Time Period	Years in Time Period
T_1	1776 - 1866
T_2	1867 - 1884
T_3	1885 - 1893
T_4	1894 - 1897
T_5	1898 - 1899
T_6	1901 - 1909
T_7	1910 - 1915
T_8	1916 - 1925
T_9	1926 - 2003

In each time period we observe: 1) a certain number of unique ontology concepts; 2) the introduction of new concepts that have not been observed in previous time periods and; 3) the obsoletion of some concepts (Figure 2). A concept is said to be obsoleted in a time period if it is last observed in that time period. As expected, the percent of new term introductions decreases with time. The fact that new concepts only account for up to 10% of concepts in any time period indicates that a large number of concepts are common between successive time periods. Interestingly, the first four time periods see no obsoleting of concepts. We see a gradual increase in concepts becoming obsolete from T_5 to T_8.

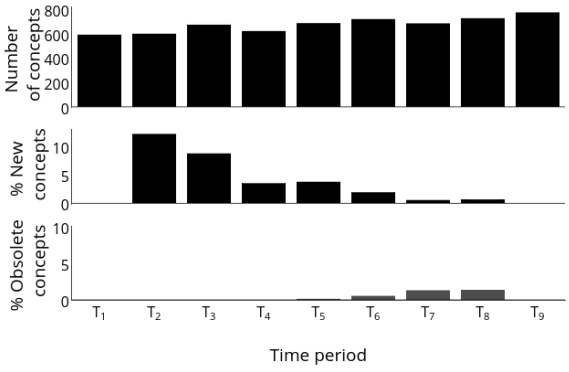


Fig. 2: **Characteristics of time periods.** Number of terms 1) in a given time period, 2) introduced in a time period, and 3) obsoleted in a time period

3.3 Estimation of concept drift

Context vectors were computed for the 799 Uberon concepts with sufficient annotations using a window size of 5. For each concept, context vector comparisons were conducted between successive time period pairs and between each time period and the original concept contexts. Note that these comparisons could be made only for time periods where a given concept is present.

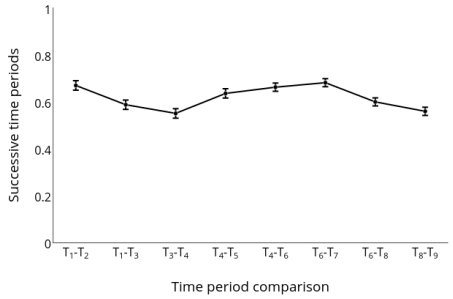
3.3.1 *Drift between successive time periods.* When comparing context vectors for successive time period pairs, we found concept drift in the approximate range of 55% to 68% (Figure 3a). There was a statistically significant difference between the magnitude of concept drift between different successive time period comparisons as determined by one-way ANOVA ($F_{7,5275} = 29.42$, $p = 3.95 \times 10^{-40}$).

400 of the 799 concepts were present in all time periods enabling context vector comparisons across all successive time period pairs. The distribution of mean concept drift for this subset of concepts across all successive time periods was extremely similar to Figure 3a. These concept drift scores were analyzed using a one-way repeated measures ANOVA for greater power. The test also showed a statistically significant difference in concept drift between different successive time periods ($F_{7,393} = 65.52$, $p = 2.22 \times 10^{-16}$).

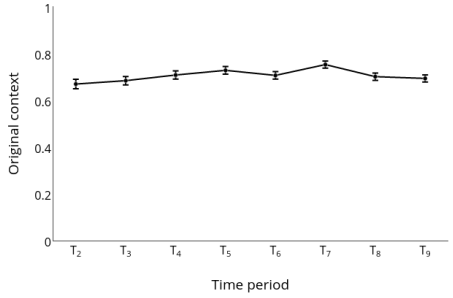
3.3.2 *Drift from original context.* Concept drift relative to original context varied from 75% to 67% (Figure 3b). If concept drift was cumulative, we would have expected to see an increase relative to the original context over time. However, the observed drift from original context does not fit this pattern.

The result is suggestive of a mean reverting process, such as an Ornstein-Uhlenbeck process [Blackwell 1998], in which the variable of interest returns to a long-term mean over time ([Huggins and Schaller 2013; Lee and Lee 2006]). Such processes have been studied intensively in finance, particularly in analysis of stock prices [Poterba and Summers 1988; Kim et al. 1991], and a common test used to verify the presence of mean reversion is the Variance Ratio Test ([Poterba and Summers 1988]), which is based on the idea that the variance of a non-stationary series increases over time. Therefore, if the variance of a series is k in the first period, it is expected to be nk in the nth period. However if the ratio of variances from subsequent time periods relative to the initial time period is consistently less then one, it indicates the presence of mean reversion. Taking the drift from T_1 to T_2 as the baseline, we found that the variance of T_3 - T_9 relative to T_1 was consistently lower than the baseline (0.45, 0.28, 0.19, 0.16, 0.11, 0.11 and 0.09, respectively), supporting the hypothesis that some factor is limiting long-term concept drift.

3.3.3 *Effect of ontology depth on concept drift.* The concept drift distributions for terms at different depths in the ontology for both successive and original context comparisons are shown in Figures 4a and 4b. In both cases, there was a statistically significant difference between concept drifts at different depths as determined by one-way ANOVA for successive time period comparisons ($F_{11,5271} = 9.47$, $p = 3.87 \times 10^{-17}$) and comparisons to original context ($F_{11,5271} = 17.11$, $p = 8.73 \times 10^{-34}$). The significant trend appears to have been due to a sharp decrease in concept drift at a depth of 14 in the ontology, which is the second-most specific ontology level represented in the dataset. This level includes eight Uberon concepts: metatarsal bone, fibula, mandible, ulna, metacarpal bone, femur, tibia, and humerus.



(a) Drift between successive time periods.



(b) Drift relative to T_1.

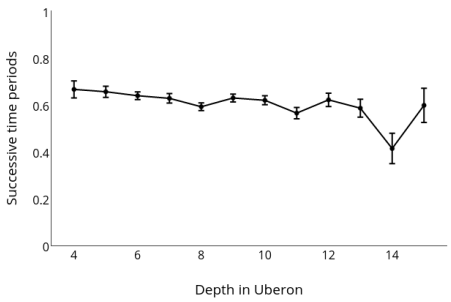
Fig. 3: Concept drift, measured as $1 - S$. Two standard errors are shown.

4. DISCUSSION

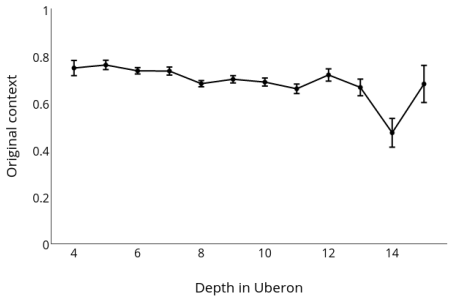
Progression of time and advancements in research could change the meaning, or at least the usage, of even seemingly objective scientific concepts such as parts of anatomy. Here, we conducted a temporal data mining analysis of vertebrate biodiversity literature to observe how and if the context of anatomical concepts changes over time. Our results indicate that concept vectors do drift by 55-75% on a decadal time scale, as measured by cosine similarity. This may reflect evolution in the meaning of the terms themselves, or evolution in the context in which the terms are applied, or both.

We were interested to see if the granularity of anatomical concepts as measured by depth in the Uberon ontology would effect the magnitude of concept drift. We found little pattern except for an unusually low level of concept drift at a depth of 14, out of a total number of 15 levels. While we did hypothesize that more specific ontology terms would have reduced concept drift, we are puzzled by the lack of a similar decrease in concept drift at Level 15. Ontology terms were examined manually for factors that might explain and anomalous level of concept drift, such as a lower number of annotation instances or a lower rate of spelling errors. However, we did not observe any substantial differences with respect to these factors between concepts at levels 14 and 15.

If concept drift is cumulative, then each subsequent time period should drift further from the original context. However, we found the magnitude of drift from original context remained roughly constant, and application of the Variance Ratio Test confirmed that concept drift was behaving as if it were a mean reverting process.



(a) Drift between successive time periods for concepts at different depths in the Uberon ontology.



(b) Drift from original context for concepts at different depths in the Uberon ontology.

Fig. 4: Concept drift ($1 - S$) for terms at different depths in Uberon. Error bars show two standard errors.

We conclude from this that concept drift is not a random walk, but rather that there is some conservative force acting to constrain the context in which a term is used over a time scale of centuries. This may be due to the long memory of the biodiversity literature. It is encoded in the rules of taxonomy how the original description of a new taxon by the earliest author to use a taxonomic name will constrain subsequent treatments of that taxon in all future literature. Such conservative practices may directly or indirectly act as a mean-reverting force, thereby explaining the non-cumulative concept drift for anatomy concepts, as well.

Despite the large corpus of vertebrate literature used for the analysis, we recognized only 18% of the anatomical concepts listed by Uberon. Despite the low coverage, the observed concepts resulted in a substantial dataset of over 900,000 annotations. One reason for not observing the majority of Uberon concepts in BHL literature might be that concepts in Uberon were unknown during the time span covered in BHL. Alternatively, the concepts may have been known but expressed differently than in Uberon's controlled vocabulary. A concept recognition algorithm more sophisticated than our naive exact matching approach might be able to construct a more comprehensive dataset from this same literature.

The concept drift reported in this study might be biased to certain extent by factors such as spelling errors in OCR text or Uberon concepts which might have other meanings distinct from scientific usage (e.g. UBERON:0002544 "digit"). For example, Wei *et al.* (2010) reported that approximately 35% of a subset of taxonomic names from the BHL OCR text had spelling errors introduced during translation [Wei et al. 2010]. This issue of spelling errors is relevant to our analysis since concept drift could be artificially inflated due to the presence of misspelled words in the context vectors of anatomical concepts. The corpus used in our analysis has an average spelling error rate of 21% and the average spelling error rate in the context vectors is 20.2%. This means that the actual drift might be lower than the drift observed in this analysis if we had a perfect corpus of literature with no OCR translation errors. We identified 57 concepts from the Uberon ontology which might have meanings apart from their scientific ones, such as UBERON:0000974 "neck", UBERON:0000978 "leg" and UBERON:0001021 "nerve". These may inflate concept drift since they can be used in varying contexts. Forty of these ambiguous concepts were present in our context vector analysis. We repeated the analysis after excluding these concepts but found no substantial changes in the results.

The purpose of ontologies is to provide a shared and consistent vocabulary for describing biological entities. It would be interesting for future work to compare concept drift in the earlier literature with drift in contemporary literature published after the adoption of biological ontologies.

5. DATA AVAILABILITY

Data and results associated with this work are available at http://dx.doi.org/10.5281/zenodo.259505.

ACKNOWLEDGMENTS

This work was conceived at a Encyclopedia of Life and Biodiversity Heritage Library research sprint supported by the Richard Lounsbery Foundation and hosted at the National Evolutionary Synthesis Center (NESCent) in Durham, NC. The authors thank Dr. Cyndy Parr from Encyclopedia of Life for organizing the sprint and NESCent for hosting it. The authors also thank the BHL team on site during the sprint, especially William Ulate for providing valuable guidance on navigating the BHL API and database. This work was also funded by the National Science Foundation (DBI-1062542).

REFERENCES

Lakshmi M Akella, Catherine N Norton, and Holly Miller. 2012. NetiNeti: discovery of scientific names from text using machine learning methods. *BMC Bioinformatics* 13, 1 (2012), 211.

Jonathan BL Bard and Seung Y Rhee. 2004. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics* 5, 3 (2004), 213–222.

Jason Best. 2013. Darwin Score. (2013). https://github.com/idigbio-citsci-hackathon/darwin-score [Accessed: Feb 2014].

Paul G Blackwell. 1998. Ornstein–Uhlenbeck process. *Encyclopedia of Biostatistics* (1998).

Judith A Blake and Carol J Bult. 2006. Beyond the data deluge: data integration and bio-ontologies. *Journal of biomedical informatics* 39, 3 (2006), 314–320.

Maya Carrillo, Esaú Villatoro-Tello, Aurelio López-López, Chris Eliasmith, Manuel Montes-y Gómez, and Luis Villasenor-Pineda. 2009. Representing context information for document retrieval. In *International Conference on Flexible Query Answering Systems*. Springer, 239–250.

Wasila M Dahdul, James P Balhoff, David C Blackburn, Alexander D Diehl, Melissa A Haendel, Brian K Hall, Hilmar Lapp,

John G Lundberg, Christopher J Mungall, Martin Ringwald, and others. 2012. A unified anatomy ontology of the vertebrate skeletal system. *PloS One* 7, 12 (2012), e51070.

Wasila M Dahdul, John G Lundberg, Peter E Midford, James P Balhoff, Hilmar Lapp, Todd J Vision, Melissa A Haendel, Monte Westerfield, and Paula M Mabee. 2010. The teleost anatomy ontology: anatomical representation for the genomics age. *Systematic Biology* 59, 4 (2010), 369–383.

Stephen I. Gallant. 2000. Context Vectors: A Step Toward a "Grand Unified Representation". In *Hybrid Neural Systems*, Stefan Wermter and Ron Sun (Eds.). Lecture Notes in Computer Science, Vol. 1778. Springer Berlin Heidelberg, 204–210.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 46, 4 (2014), 44.

Thomas R Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220.

Nancy E Gwinn and Constance Rinaldo. 2009. The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA journal* 35, 1 (2009), 25–34.

Doug Huggins and Christian Schaller. 2013. *Fixed Income Relative Value Analysis: A Practitioners Guide to the Theory, Tools, and Trades*. John Wiley & Sons.

Myung Jig Kim, Charles R Nelson, and Richard Startz. 1991. Mean reversion in stock prices? A reappraisal of the empirical evidence. *The Review of Economic Studies* 58, 3 (1991), 515–528.

Cheng-Few Lee and Alice C Lee. 2006. *Encyclopedia of Finance*. Springer Science & Business Media.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 13, 1 (2012), R5.

Roderic DM Page. 2011. Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. *BMC Bioinformatics* 12, 1 (2011), 187.

Roderic DM Page. 2013. BioNames: linking taxonomy, texts, and trees. *PeerJ* 1 (2013), e190.

James M Poterba and Lawrence H Summers. 1988. Mean reversion in stock prices: Evidence and implications. *Journal of financial economics* 22, 1 (1988), 27–59.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Addison Wesley.

Simon Tanner, Trevor Muñoz, and Pich Hemy Ros. 2009. Measuring mass text digitization quality and usefulness. *D-Lib Magazine* 15, 7/8 (2009), 1082–9873.

Anne E. Thessen, Hong Cui, and Dmitry Mozzherin. 2012. Applications of Natural Language Processing in Biodiversity Science. *Adv. Bioinformatics* 2012 (2012), 391574:1–391574:17.

Alexey Tsymbal. 2004. *The problem of concept drift: definitions and related work*. Technical Report TCD-CS-2004-15. The University of Dublin, Trinity College, Department of Computer Science, Dublin, Ireland.

Peter D Turney, Patrick Pantel, and others. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37, 1 (2010), 141–188.

Ted Underwood. 2013. OCR Normalizer. (2013). https://github.com/tedunderwood/DataMunging/tree/master/OCRnormalizer [Accessed Feb 2014].

Shenghui Wang, Stefan Schlobach, and Michel Klein. 2010. What is concept drift and how to measure it? In *Knowledge Engineering and Management by the Masses*. Springer, 241–256.

Qin Wei, P. Bryan Heidorn, and Chris Freeland. 2010. Name matters: taxonomic name recognition (TNR) in biodiversity heritage library (BHL). In *iConference 2010 Proceedings*. 284–288.