Examining the OpenAlex Concepts: A Detailed Case Study of Machine-Derived Classification

Huma Zafar

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master of Information Studies

School of Information Studies
Faculty of Arts
University of Ottawa

Acknowledgements

I would like to extend many thanks to my supervisor, Dr. Stefanie Haustein, for her support, guidance, and expertise, and for numerous discussions that helped refine the direction of this project, which started off what feels like a very long time ago, with very different objectives. I would like to express my thanks as well to my fellow members of the Scholarly Communications Lab for their general support and camaraderie, and to Dr. Jada Watson whose Advanced Research Methods class furnished several inspiring readings and discussions that helped me formulate the analytical framework used in this project. To my dear siblings, Meraj and Abdullah, to my mother, Najmussahar, and to my partner, Brendan, thank you all for never getting bored of listening to me talk out an idea, for your enthusiasm and continuous encouragement, for all the hugs and the nourishment. I am very grateful for the support I have had over the past many months, and I would not have been able to bring this project to completion without it.

Abstract

Machine-learning techniques are becoming increasingly popular in metadata and classification work due to their ability to operate at scale, but insufficient consideration has been given to how effective such techniques truly are against traditional practice. This thesis adopts an approach based on *Data Feminism* (D'Ignazio & Klein, 2020) to analyze the machine-generated OpenAlex concept hierarchy and its associated machine-learning model in comparison to established classification standards and practices. We find that the OpenAlex concepts differ vastly from a traditional classification system, and that this difference inhibits their effectiveness in some respects, while also offering possible ways to address modern criticisms of classification (Olson, 2001; Mai, 2005). We argue that statistical, data-processing approaches to classification cannot replace the human judgment necessary for classification work, and that, to the extent that they continue to be used in this type of work, automatic techniques should be more informed by information theoretical principles and guided by human expertise.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Classification tasks are one of the primary applications of machine learning (ML), and many ML-based systems have been implemented for automatic subject classification of scholarly resources (Golub, 2021). These systems tend to replace manual, time- and labour-intensive processes with faster and more convenient solutions. However, their development may often ignore the information theoretical concerns that drive established and evolving classification practices, as the process of designing and evaluating a machine-learning model differs in many ways from that of applying classification theory towards constructing a classification system.

This project investigates OpenAlex, an open access catalogue of scholarly research with an automatic subject classification model that has the advantage of being open source. OpenAlex is developed and maintained by OurResearch, and is funded

through premium subscriptions as well as a grant from Arcadia, a UK-based charitable foundation that has a mandate to support open-access projects (OurResearch, n.d.; Arcadia, n.d.). OpenAlex contains over two hundred million works and is the successor of Microsoft Academic Graph (MAG), from which it inherits its classification hierarchy. The OpenAlex concepts are organized in a poly-hierarchical structure under 19 disciplines, where each concept may fall under multiple top-level disciplines. Works in OpenAlex are assigned concepts using a machine-learning model that was designed to replicate the performance of the MAG classifier and was originally trained on MAG data (OpenAlex, n.d.). Each concept in OpenAlex is linked to associated entities in Wikidata and Wikipedia.

It should be noted that in December 2023, the OpenAlex team announced their intention to replace their existing classification system with a new one designed in collaboration with the Centre for Science and Technology Studies[1] (Portenoy, 2023). As this project was conceived before that announcement, it focuses on studying the OpenAlex concepts and machine-learning model as they existed before the overhaul.

1.2 Objectives and Research Questions

The objective of this research is to investigate OpenAlex's automatic classification model, not with the direct goal of improving its functionality, but to explore answers to the following questions: What gaps in classificatory power are left by the statistics-driven process of developing machine-learning models? What are the impacts of these gaps on subject classification for scholarly resources? And finally, how

[1] https://www.cwts.nl/

does OpenAlex as a classification system measure up against critical perspectives on current classification practice?

Chapter 2

Theoretical Framework

We have chosen to develop an analytical framework based on the principles of data feminism as described by D'Ignazio and Klein (2020). Data feminism is a collection of strategies for approaching data science from a more intersectional, feminist perspective in order to reduce and redress some of the harms of traditional data science work (D'Ignazio & Klein, 2020). Since some of these harms—such as the disconnect between those who build solutions and those who use them, or the underserving of minoritized viewpoints—are equally relevant when it comes to machine learning, we believe that the critical techniques described by D'Ignazio and Klein (2020) are very useful for studying the OpenAlex automatic classification model.

Our framework uses two principles of data feminism in particular—examining power, and rethinking binaries and hierarchies—to perform first a top-down and then a bottom-up analysis of OpenAlex's classification model. We then draw on a third principle, embracing pluralism, to examine the results of the previous analyses

and imagine how automated classification work might be improved in general.

2.1 Examining power

As per D'Ignazio and Klein (2020), examining power is about identifying structural privilege and oppression, and investigating how it leads to systemic advantages or disadvantages, and for whom. Not only does this involve locating where power exists in a system, but also understanding how the interplay of power within different domains upholds and reinforces existing privilege and oppression (D'Ignazio & Klein, 2020). OpenAlex itself wields power in two ways: the power to determine the set of concepts used by its model, taken from those of its predecessor, MAG; and the power to classify documents according to these hierarchical concepts (OpenAlex, n.d.). To answer questions about where this power comes from, we examine the choices made with respect to the model architecture and with respect to the selection of data used to train the model. We especially seek to understand which outcomes were prioritized with these decisions, and which were given less consideration.

To connect this power analysis within OpenAlex to a broader perspective on the usage of machine learning for classification work, we consider the *matrix of domination* model, first proposed by Collins (2008, as cited in D'Ignazio & Klein, 2020). This model defines four domains within which power can be examined: structural (e.g., laws that encode power); disciplinary (ways that structural power is enforced); hegemonic (cultural norms that enforce power); and interpersonal (the impact of power structures on individuals) (Collins, 2008, as cited in D'Ignazio & Klein, 2020).

(An example of structural power with respect to classification work might be found in the definitions of various classification standards used predominantly in the library and information science field, such as NISO (2010) and Library of Congress (2016), both of which we refer to (and challenge) in our analyses and will return to in our discussion.) We use this model to discuss how the issues that we identify in OpenAlex at a technical level can be seen to arise from more complex power structures at a societal level.

We also use the lens of "small data" as discussed by Welles (2014) to explore some of these choices. Small data refers to those parts of Big Data sets that may be considered statistical outliers due to their relatively smaller representation in the overall dataset (Welles, 2014). However, rather than excluding this data (whether accidentally or intentionally), Welles (2014) argues that "focusing on minority experiences as reference categories, rather than as deviations from the majority reference, enables better, more accurate theory building and data modeling" (p. 2). We apply this perspective towards attempting to find more meaningful interpretations of apparent errors and ambiguities in the output of the OpenAlex model.

2.2 Rethinking binaries and hierarchies

Rethinking binaries and hierarchies involves interrogating the potentially harmful assumptions that get built into classification systems and who these assumptions do and do not serve (D'Ignazio & Klein, 2020). As D'Ignazio and Klein (2020) discuss, binary definitions can enforce an either/or way of thinking that causes everything

that falls outside of the binary to be left uncounted, while hierarchical structure can imply value judgments about the groups that they are used to classify. The cost of classifications that poorly represent the real world can also hinder progressive change, such as by presenting inaccurate bases for policymaking (D'Ignzaio & Klein, 2020). In the world of scholarly communications systems, rigid and inadequate classifications may render certain research invisible and, hence, undervalued by research evaluation protocols.

To this end, we critically examine structural aspects of the OpenAlex concept hierarchy, in terms of both the limitations and the advantages resulting from the non-strictness of its hierarchical structure. We question, as D'Ignazio & Klein suggest, to what degree the issues within the OpenAlex classification are a problem of inadequate categories versus problems with the system itself. We also look at the original motivations for the construction of this classification as part of a platform for scientific research evaluation specifically, and how that informs the shape and coverage of the OpenAlex concepts.

2.3 Embracing pluralism

One of the biggest selling points of OpenAlex's automatic classification model is the enormous scale at which it is able to operate. Classification work at scale is a real challenge for metadata professionals, and machine-learning techniques are becoming an increasingly common replacement for human labour. However, automatic methods risk oversimplifying solutions, as well as disempowering information profession-

als, if they are not designed very carefully and deployed judiciously. To combat the problem of solutions that do not adequately respond to the needs of the communities that they are built for, D'Ignazio and Klein (2020) advocate for pluralistic solutions that synthesize points of view from multiple stakeholders. In particular, they argue for actively involving members of minority communities in design and analysis work. To conclude our discussion of OpenAlex, we consider how our findings bear on this idea of "co-liberation" (D'Ignazio & Klein, 2020, p. 140) in the context of automatic classification work, and explore what kinds of benefits we might obtain from embracing more "messiness and complexity" (D'Ignazio & Klein, 2020, p. 131) in automatic classification models and the processes around them. Ultimately, we hope that this critical analysis of OpenAlex can suggest avenues for conducting classification work at very large scales with greater and more meaningful participation from metadata communities themselves.

Chapter 3

Methodology

3.1 Study Design

We started with a mixed-methods design for this project, with the intention to interleave qualitative and quantitative assessments. As we could not find existing studies in the literature that analyzed classification systems the way that we were interested in, we did not start from any specific guidelines for how to evaluate the quality of a classification system. We chose instead to leave the exact lines of inquiry loosely defined at the outset of this project to give us the freedom to define such guidelines as the project progressed. To do so, we first performed a "close reading" of the MAG and OpenAlex technical literature, as well as of the available documentation and source code for OpenAlex, to gain a sense of where issues related to quality might arise within 1) how the model's outputs (namely its classification decisions and, in the case of MAG, the concept hierarchy itself) were generated; and 2) how

the model(s) were trained, and using what data. We also performed a review of the classification theory literature to identify topics that might provide useful framings for our analyses, such as Beghtol's (2003) discussion of classification for knowledge discovery systems, and Olson's (2001) discussion of the limitations of the Dewey Decimal Classification and the Library of Congress Subject Headings.

The technical readings furnished ideas for how to explore the OpenAlex concepts dataset in more depth; for example, after understanding Shen et al.'s (2018) definition of the subsumption formula used in the construction of the MAG concept hierarchy, we investigated the formula using actual confidence scores from the OpenAlex database, which helped us develop hypotheses about the underlying causes of some of the apparently erroneous relationships in the concept hierarchy; these, we then investigated further through more data exploration, statistical analyses, visualizations, etc. The theoretical literature, on the other hand, gave us ideas for critical examination of both the OpenAlex concepts, as well as the technical literature. That is, we revisited the technical literature with a focus on how, and to what extent, these papers spoke about critical issues in classification theory such as bias, formal structure, warrant, intended usage, etc. This second close reading of the technical literature helped us consider how the specific technical issues we had flagged earlier could be connected to larger, systemic problems around the use of machine learning in classification work.

3.2 Data Collection

We downloaded a local copy of the October 2023 snapshot of the complete OpenAlex data and imported it into a local Postgres database, following the instructions provided in the OpenAlex documentation (OpenAlex, 2024). Due to the enormous size of the OpenAlex dataset, the database itself was created and run on an external drive. The final size of the database after importing all the downloaded data was approximately 1.5 terabytes. There were no modifications needed to the OpenAlex extraction and import scripts, but we added extra indices to some of the tables, in particular the concepts table, to improve response time for large queries. In total, we estimate that all of the steps involved in downloading and setting up the local database took about 10 days.

3.3 Data Analysis

We wrote Python scripts to compile various datasets on the concepts and concept hierarchy, such as the number of ancestors and children for concepts at different levels of the hierarchy, the distribution of top-level disciplines among the ancestors for each concept, path characteristics for the concept graph, etc. We explored these datasets through descriptive statistics, manual inspection, and visualizations. The graph visualizations given in figures 5.1, 5.2, and 5.3 were generated using the free, online version of Cosmograph[1] using CSVs containing graph data and graph metadata produced via our Python scripts. The open source GNU Image Manipulation

[1] `https://cosmograph.app/`

Program (GIMP)[2] was used to add labels to these visualizations, and the open source ImageMagick[3] set of command line tools were used to layout multiple visualizations in a grid. Other graph diagrams, such as in figure 5.4, were generated by converting concept graph data into a dotfile representation[4] using Python, and then running a script to invoke the graphviz[5] `dot` command to convert the dot files into PNG format image files.

The data and scripts supporting the findings of this study are available at Zafar (2024a) and Zafar (2024b).

[2]https://www.gimp.org/
[3]`https://imagemagick.org`
[4]`https://graphviz.org/doc/info/lang.html`
[5]`https://graphviz.org/doc/info/command.html`

Chapter 4

Literature Review

The search for existing literature on this topic was divided among the following areas: technical discussions of the machine-learning models used by Microsoft Academic Graph (MAG) and OpenAlex; analyses of the MAG and OpenAlex metadata; and critical discussions of classification theory in general. We found that there was relatively little published literature on the inner workings of either MAG or OpenAlex, with the only publications being those furnished by Microsoft and OpenAlex themselves. However, we found more literature available that analyzed MAG and, to a lesser extent, OpenAlex from a user perspective and that compared metadata quality to that of other research evaluation platforms. While there is a large body of existing literature on many different aspects of classification theory and practice, we focused our search on papers related to traditional versus non-traditional ideas of classification and to issues of modernizing classification practices, such as those that addressed problems like bias within classifications.

4.1 Microsoft Academic Graph and OpenAlex

4.1.1 Machine-learning models

The Microsoft Academic Graph (MAG) comprised various machine-learning components as part of Microsoft Academic Services (MAS), a platform developed to support bibliometric analyses on science studies in particular (Wang et al., 2020). General overviews of the technical components of MAG are given in Wang et al. (2019) and Wang et al. (2020), but the most thorough discussion available of the design of the MAG machine-learning model is provided by Shen et al. (2018), who have described with some specificity the mechanisms for concept discovery, concept tagging, and concept hierarchy construction. In particular, they have highlighted the technical rather than information-theoretical foundations for the MAG classification, such as in the technique used for determining hierarchical relationships (Shen et al., 2018). (These mechanisms and the details of how the models were trained will be discussed further in chapter 6.) Since OpenAlex's concept hierarchy was taken directly from MAG, a lot of the MAG literature provides useful insight into the origins of OpenAlex's classification system as well. OpenAlex (n.d.) have further provided an overview of the OpenAlex concept tagger as well as the technical design decisions behind its architecture and training process. However, their work contains no further elaboration on the motivations for these decisions beyond the goal of achieving performance similar to the MAG concept tagger.

In their review of the redesign of MAS, Wang et al. (2020) discussed sources of bias that they intended to mitigate in the design of MAG, which notably included

the "confirmation and other cognitive biases that are hard to detect, reconcile, and, most importantly, to explain" (p. 39) that can result from human-developed classifications and human guidance of machine-learning systems. In another paper, Wang et al. (2019) equated taking a "data-driven approach" using machine-learning to "providing consistent data quality", and contrasted this approach with "manual efforts that are often the source of subjective controversies or errors" (p. 2). However, the MAG literature also acknowledged the presence of bias in some of its machine-learning components; for example, Wang et al. (2020) described the use of principal component analysis (PCA) to filter out potentially fraudulent citations from MAG, but noted that one consequence of using PCA was an increased bias against non-English publications because the asymmetry of citation patterns between English and non-English publications was indistinguishable from fraudulent behaviour in the model. Thus, despite the intent to eradicate bias through the use of machine-driven processes, those processes still spawned other forms of bias within MAG.

4.1.2 Metadata quality

There have been several studies analyzing MAG and OpenAlex metadata coverage, mostly focused on accuracy of author and publication information compared to other databases. Herrmannova and Knoth (2016) compared MAG publication coverage to that of other public datasets, such as CORE and Mendeley, and concluded that, overall, MAG metadata is comprehensive and of comparable accuracy to other datasets. Hug and Brändle (2017) analyzed publication recall in MAG using a publication list from the University of Zurich, comparing against the recall for the same list in Web

of Science and Scopus as benchmarks. They found that all three sources had similar coverage, though MAG and Scopus generally outperformed Web of Science, and MAG had the most exclusive coverage (Hug and Brändle, 2017). They also had very similar findings to Herrmannova and Knoth (2016) with respect to the quality of publication metadata (such as publication years) in MAG (Hug and Brändle, 2017). Culbert et al. (2024) compared OpenAlex reference and metadata coverage to that of Web of Science and Scopus, and found that OpenAlex was generally slightly less reliable than the other two platforms.

While none of the studies that we found analyzed the quality of subject coverage in a systematic way, in their assessment of the Microsoft Academic platform as a bibliometrics tool, Hug et al. (2017) observed that the fields of study provided in MAG were "dynamic, too specific" and that "field hierarchies are incoherent", and, as a result, cautioned against the use of MAG fields of study for field-normalization in research evaluation studies (p. 377). They noted specific issues, such as the unusual subordination of *Social sciences* to *Sociology* in the first two levels of the MAG hierarchy (when, canonically, the former would be considered the parent field of the latter) but did not conduct a systematic study of the MAG fields of study as a whole (Hug et al., 2017). Scheidsteger and Haunschild (2023) studied the extent of discrepancies between MAG and OpenAlex metadata, evaluating how many papers were re-classified into different concepts under OpenAlex from their concept assignments in MAG, but mainly to quantify the difference between the two systems. When they did make qualitative assessments of the concept assignments in OpenAlex, it was to compare whether the same papers had similarly suitable or unsuitable con-

cepts assigned in MAG, rather than to judge the quality of the concepts themselves (Scheidsteger & Haunschild, 2023).

4.2 Classification Systems and Theory

4.2.1 Classifications for knowledge discovery

In a paper discussing knowledge discovery classification systems, which she termed "naive" classifications and distinguished from classifications for information retrieval, Beghtol (2003) articulated the importance of understanding the relationships between different purposes of classification. While classifications for information retrieval focus on providing access to "knowledge that we already have and that has already been stored in documents", classifications for the purpose of discovering new knowledge aim to "enhance domain knowledge for the pursuit of scholarly activity and research" (Beghtol, 2003, p. 65). In other words, the principle goal of knowledge discovery classification design is to organize collections in ways that allow new insights to be made from or about those collections, though the design steps involved may be similar to those used for information retrieval classifications (Beghtol, 2003).

Beghtol (2003) discussed structural elements that are present in many classifications, but that may be used in different ways for knowledge discovery than they are for information retrieval. For example, she noted that hierarchies in knowledge discovery classifications tend to be shallow and have nondescript class names (such as 'Type I', 'Type II', etc.), with detailed scope notes that explain the meaning of each class instead (Beghtol, 2003). While this approach would not be suitable for

information retrieval tasks, it may work well for a classification focused on identifying and studying patterns of interest among resources. Beghtol (2003) also argued that a knowledge discovery classification may be conducive to more flexible information environments because it can enable users "to discover a way to ask new questions of available, often newly revealed, evidence" (p. 71), making it, as Olson (2004) has put it, "an active agent as well as a reflective one" (p. 3). On the other hand, information retrieval classifications tend to hew closer to established terminology and therefore present more fixed ways of exploring collections, that are revised at most as often as new knowledge is accepted into a field (Beghtol, 2003).

Though they did not use the framing of knowledge discovery classification in particular, Glänzel and Schubert (2003) highlighted the shortcomings of classifications designed for information retrieval when applied to research evaluation tasks, and proposed their own two-level hierarchical classification designed to work well with various scientometric calculations. There have been many other studies of research evaluation taxonomies focusing on techniques for generating classifications based strongly on patterns within the collection itself, which arguably corresponds well to Beghtol's (2003) idea of knowledge discovery classification. Many of these studies have explored the use of different types of citation analyses to cluster research papers and determine a classification scheme from the resulting clusters, relying primarily on automatic techniques (Klavans and Boyack, 2017; Waltman and van Eck, 2012; Shibata et al., 2009). Other studies have focused on clustering based on bibliographic metadata (Boyack et al., 2011; Boyack et al., 2013), and some have analyzed the effectiveness of combining automatic techniques with guidance from experts and

incorporation of existing classifications (Glänzel and Schubert, 2003; Archambault et al., 2013; Rafols and Leydesdorff, 2009), though Rafols and Leydesdorff (2009) also echoed the sentiment of Wang et al. (2019) regarding the value of automatic techniques over human-indexing, stating, "However puzzling the results of the algorithmically generated classifications may be, these structures provide us with the *explanandum* of... a sociology of science" (p. 1833, emphasis in original). A common theme among these studies was the challenge presented by interdisciplinarity; more rigorous automatic methods were unable to capture multiple disciplines for works (Waltman and van Eck, 2012), or required manual intervention to do so (Rafols and Leydesdorff, 2009). The complexity of multi-disciplinary classification was also considered generally undesirable, as genuinely interdisciplinary work was perceived to be uncommon (Pena-Rocha et al., 2024).

4.2.2 Warrant

The role of warrant in classification design decisions has often been discussed in the literature (Beghtol, 1986; Lee, 2017; Barité and Rauch, 2020; Barité, 2018; 2019). Bullard (2017) has argued that warrant embodies meaningful choices that bridge the gap between "daily practice" and "longstanding theoretical concerns in classification research" (p. 76), and can therefore illustrate how conflicts in classification design are navigated in practice, sometimes through the combination of contradictory warrants. Bullard (2017) has given an overview of four types of warrant, including the two most commonly applied warrants in classification design, literary and scientific warrant. The application of literary warrant, in which "classification decisions are

derived from the scholarship being organized" may appear to be free of bias stemming from individual indexers' interpretations of texts, as terminology can be lifted from the literature itself (Bullard, 2017, p. 78). However, in practice, all indexing requires the use of some contextual knowledge to interpret the subject of a work, to place it within or across disciplines, and to understand what user needs it may fulfill (Mai, 2005). Furthermore, the terms that are most prevalent in the literature may encode larger power structures within the discipline, such as a majority consensus that prefers one set of terminologies over another, which can constitute another source of bias perpetuated by the use of literary warrant (Barité, 2018). Scientific warrant takes into account the knowledge landscape from a broad perspective, rather than as extracted from a particular collection, and can involve relying on subject matter experts to outline the structure of a classification and provide guidance for interpreting works within a field (Bullard, 2017). Hjørland (2002) has described a variety of approaches for performing domain analysis to help classification designers understand the underlying structure of knowledge within a domain, but these strategies still tend to privilege majority consensus and, as we discuss below, do not sufficiently circumvent problems of universality and bias (Feinberg, 2007).

4.2.3 Bias in classification

Mai (2011) has argued that, while classification theory has historically favoured a realist approach, where the goal is to extract the "actual meaning of documents" (p. 721), meaning arises from the interaction between users and documents within a particular context, and therefore, the application and use of a classification sys-

tem always requires interpretation and understanding of local context. Rather than attempting to describe the world as it is, undertaking classification "from a semiotic distance" (Mai, 2011, p. 721) to the world requires acknowledging that different perspectives can produce different, but equally valid, descriptions. Such plurality of meaning makes both universality and so-called neutrality in classification impossible, as such a stance would imply "that one's view is a view from nowhere, that one somehow holds a view that is superior to other's views" (Mai, 2011, p. 724).

Hjørland and Albrechtsen (1995) have also advocated for approaching classification from a social perspective, contextualized to the needs of a particular domain, and recognizing the plurality of "worldviews, individual knowledge structures, biases, subjective relevance criteria, particular cognitive styles, etc." (p. 409) that exist within different domains. However, this domain analytic approach nonetheless relies on a single definition of (the boundaries of) each domain, privileging the point of view of appointed experts above others in determining what is included within a domain (Feinberg, 2007). Feinberg (2007) has applied the concept of "situated knowledge" (Haraway, 1998, as cited in Feinberg, 2007) to advocate for developing classification systems with more explicit and deliberate acknowledgement of the choices and assumptions made therein, rather than pursuing the elimination of bias altogether.

Olson (2001) has analyzed how traditional classification systems perpetuate specific social and cultural biases under the guise of universal perspectives, such as how the lack of symmetry in the language and structure used for topics about women compared to topics about men in the Library of Congress Subject Headings presents

women as an "exception" and men as the "norm" (p. 646). Olson (2001) has noted that attempting to make standards more inclusive "is useful to a point" (p. 661) and has proposed instead that information professionals find ways to make the limits of information systems more "permeable", such as by allowing multiple authorized forms of a heading to accommodate different terminology, and by developing technical interfaces to mediate between different forms of headings more easily.

The prevailing notions regarding bias in the reviewed literature were in strong opposition to the idea that classification should strive for universal, neutral descriptions, and a desire for more plurality within classifications as a means of embracing different points of view. Attempting to eliminate bias in classification was seen as an ineffective approach to mitigating its harms, as all classification requires some judgment, and thus, bias will always be an inherent part of classification work. Recognizing and acknowledging bias, and in doing so, limiting the scope of a classification and making room for alternative descriptions, was repeatedly presented in the literature as a better strategy for improving classification systems.

Chapter 5

An Overview of the OpenAlex Concepts

5.1 Introduction

In this chapter, we will perform a top-down analysis of the OpenAlex concept hierarchy, discussing various characteristics of its structure and coverage, and reviewing its usage among researchers. This chapter will only consider those aspects of the OpenAlex concepts that can be observed from the classification itself. The internal workings of the MAG and OpenAlex classification models will be discussed in the next chapter.

5.2 Structure and Coverage

5.2.1 Poly-hierarchy

The OpenAlex concepts are organized within 19 top-level (L0) disciplines and distributed over six levels of hierarchy (L0-L5). They form a poly-hierarchical structure, where concepts can have multiple parents and may descend from multiple L0 disciplines. Figure 5.1 depicts the clustering of the OpenAlex concepts based on the L0 discipline(s) to which they belong. Each node in this graph represents a concept, and the edges of the graph connect each child concept directly to each of its L0 discipline(s). (Both the nodes and the edges are too numerous to be distinctly visible.) The colouring of the graph gives an intuitive look at how and where the OpenAlex disciplines overlap; for example, some of the yellow-green clusters near the bottom of the *Biology* and *Medicine* (green) clusters depict the groups of concepts that intersect *Sociology* (deep yellow) with those other disciplines.

The platform for which MAG was developed was originally focused specifically on scientific research (Wang et al., 2019), and there is still a heavy imbalance in favour of science disciplines observable within the OpenAlex concepts. Figure 5.2 shows copies of the same graph as in Figure 5.1, but with only one L0 discipline highlighted at a time, from which the dominance of scientific disciplines in OpenAlex can be seen clearly. In particular, we can see that *Biology*, *Chemistry*, and *Medicine* make up a substantial portion of the classification, while *Art* and *History* are noticeably much smaller.

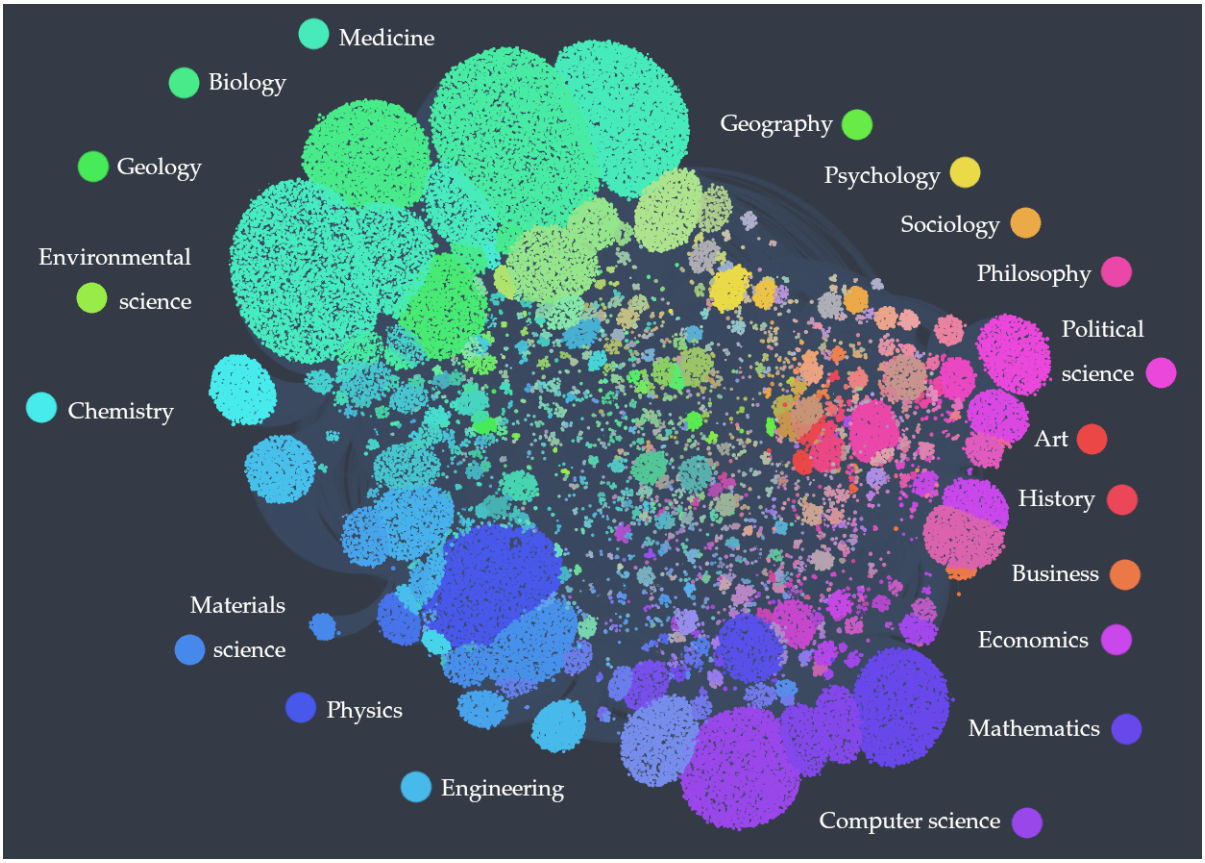


Figure 5.1: A colour-coded representation of OpenAlex L0 disciplines and their overlaps. Each of the 19 L0 disciplines has been assigned a unique colour, and the concept graph has been coloured in by blending all of the colours of each concept's ancestors.

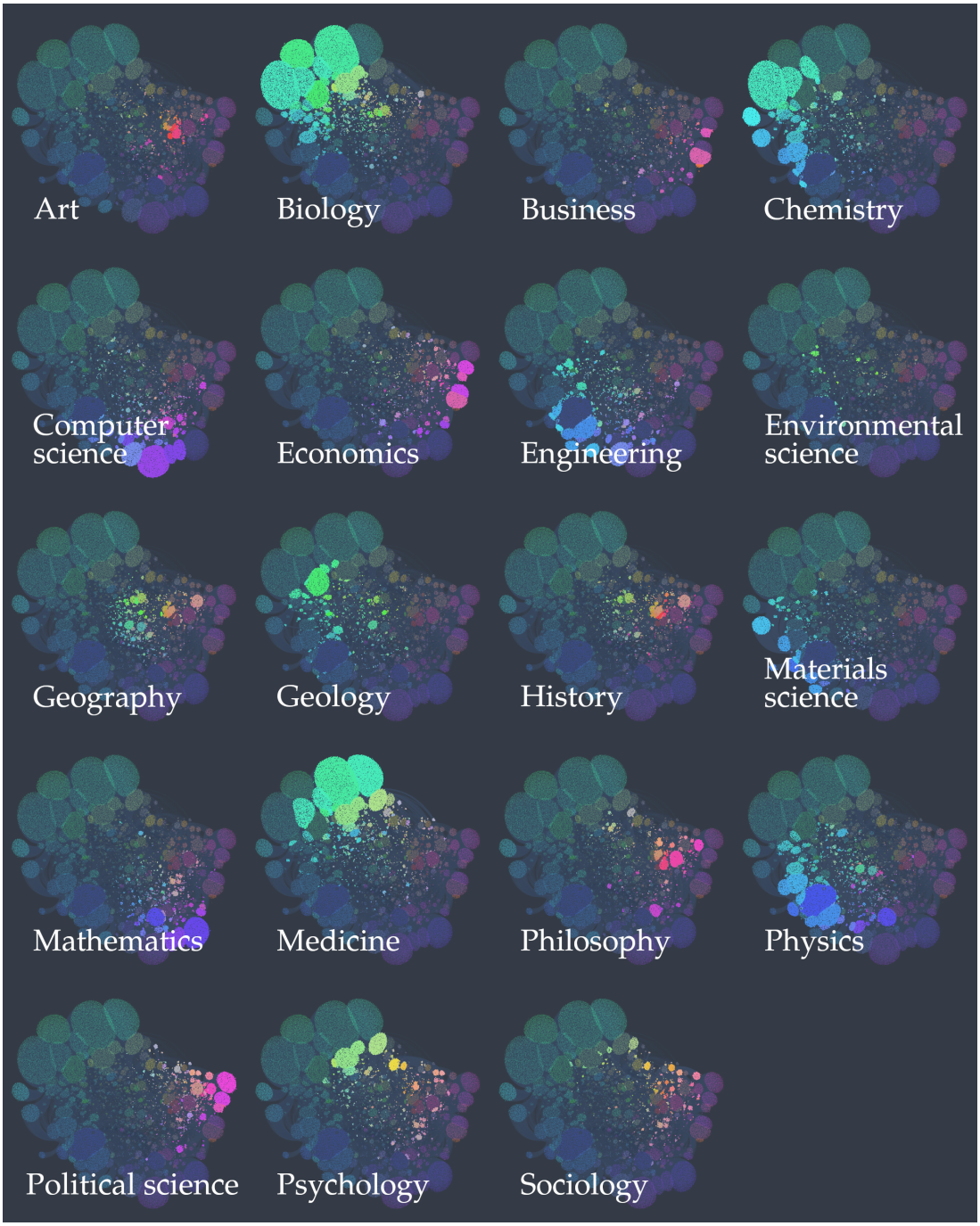


Figure 5.2: Highlight of each L0 discipline's coverage in OpenAlex.

Every concept in the OpenAlex hierarchy can be reached from some set of L0 disciplines, which form its ancestor set. The number of concepts that have a given number of L0 ancestors is presented in table 5.1. Only a little over a quarter of the OpenAlex concepts fall under a single discipline, and over 4500 concepts fall under four or more.

Size of ancestor set, n	Number of concepts with n L0 ancestors
1	18693
2	28151
3	13540
4	3590
5	853
6	167
7	42
8	10
9	8
≥ 10	0

Table 5.1: Number of OpenAlex concepts that have a given number of L0 ancestors.

A sample illustration of the poly-hierarchical nature of the OpenAlex concepts is given in figure 5.3, which depicts a clustering of concepts that have exactly five L0 ancestors. We chose to look at concepts with five ancestors for two reasons: first, such a large number of distinct ancestors might suggest that these are highly specialized, multidisciplinary concepts, which would be interesting to explore; and, second, the total number of concepts under this size could fit comfortably into a single diagram. The coloured nodes in the graph represent each child concept, and the central grey nodes of each cluster represent the group of five L0 concepts that form the ancestor set for that cluster. (Only select nodes have been labelled to preserve readability.)

We can see that a few of the L0 groups have numerous children, and at a glance, these clusters appear to have ancestor sets made up of somewhat closely related disciplines. However, most of the groups appear to have only a handful of children, or, in some cases, only a single child concept. If we understand these groups as representing multidisciplinary concepts, it is worth asking what utility is provided to users of this classification by so many unique combinations of different disciplines that contain only one or a few concepts.

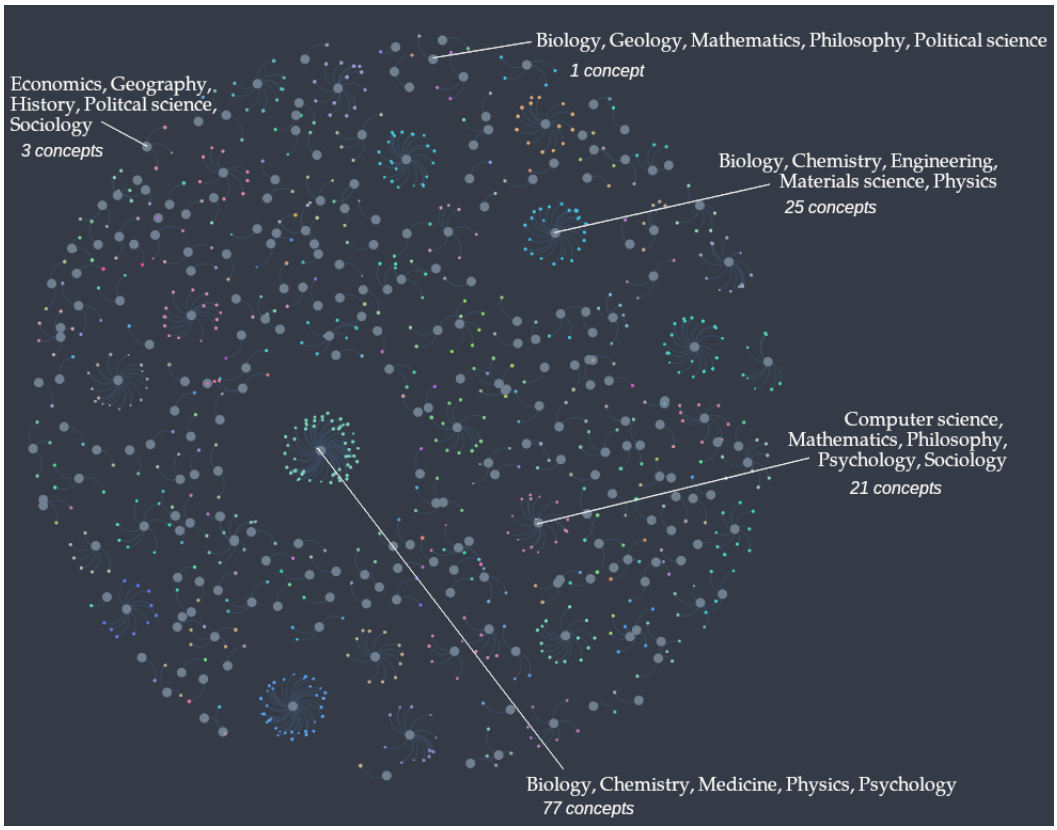


Figure 5.3: Clusters of concepts that have exactly five L0 ancestors.

The details for a few randomly selected clusters with two or fewer child concepts

are given in table 5.2. The child concepts in each of these cases do not appear to be multidisciplinary, or at least not to the extent of the given ancestors. For example, the concept of *Egg incubation* would seem to fit wholly within *Biology*, and, while *Urban design* might be discussed in the context of interdisciplinary research (for example, in a paper that discusses the history of urban design in a particular geographical region), we might not necessarily consider the concept itself to belong under the disciplines of *History*, *Geography*, and *Biology*.

L0 Ancestor Set	**Child Concept**	**Concept Description**
Art, Computer science, Mathematics, Philosophy, Sociology	Literal translation (`https://openalex.org/C2777761643`)	Word-by-word translation of a text
Biology, Geography, History, Mathematics, Medicine	Stover (`https://openalex.org/C2780500427`)	Leaves and stalks of field crops
Art, Biology, Chemistry, Philosophy, Psychology	Egg incubation(`https://openalex.org/C73152493`)	The process by which certain oviparous (egg-laying) animals hatch their eggs
Biology, Chemistry, Computer science, Materials science, Mathematics	Filament winding (`https://openalex.org/C2780008023`)	Fabrication technique using strength fibres in a binding matrix
	Wood-plastic composite (`https://openalex.org/C2781441102`)	Wood-plastic composite
Art, Biology, Engineering, Geography, History	Placemaking (`https://openalex.org/C162861558`)	Approach to public space design
	Urban design (`https://openalex.org/C205300905`)	Process of designing and shaping cities, towns and villages

Table 5.2: Examples of concepts with five L0 ancestors.

The preceding tables and figures suggest that there is a large amount of disciplinary overlap in the OpenAlex poly-hierarchy, and raise questions about what that overlap signifies. For a concept to be in hierarchical relationship with many different disciplines, does it need to represent the intersection of those disciplines? Is it a combination of (all of) those disciplines, and if so, how?

Two different approaches to hierarchy

NISO (2010) differentiates between hierarchical relationships and merely associative relationships, the latter of which are defined as "associations between terms that are neither equivalent nor hierarchical" (p. 51). The guidelines given in NISO (2010) provide only three possible interpretations for hierarchical relationships in a controlled vocabulary: *generic* relationships, where the child is a member of the class represented by the parent; *instance* relationships, where the child is an instance of the category represented by the parent; and *whole-part* relationships, where the child forms a part of the parent concept. These constraints still apply in the case of poly-hierarchy; each individual relationship within a poly-hierarchical structure must fit one of the three aforementioned categories, though the same concept may have different types of relationships with different parents (NISO, 2010). Importantly, even if a term does have different types of relationships among multiple parents, the term itself is still understood to represent "a single concept" as mandated by NISO (2010, p. 23). We can see this with the example given in NISO (2010) of the term *skull* and its parents, *bones* and *head*, where $bones \longrightarrow skull$ is a generic hierarchical rela-

tionship (*skull* is a member of the class *bones*), while *head*$\longrightarrow$*skull* is a whole-part relationship (NISO, 2010). However, the conceptual referent of *skull* is the same in both relationships; the interpretation of "skull" does not change, for example, to "brain" (as in the idiom, "thick-skulled") in the *head*$\longrightarrow$*skull* relationship.

Wang et al. (2019) reveal a somewhat different approach to hierarchical relationships, in discussing the (poly-)hierarchy of the MAG fields of study:

> Furthermore, concepts are hierarchical in nature. For example, "machine learning" is a concept frequently associated with "artificial intelligence" that, in turn, is a branch of "computer science" but often intersects with "cognitive science" in "psychology." Accordingly, a taxonomy must allow a concept to have multiple parents and organize all concepts into a directed acyclic graph (DAG). (p. 5)

Here, they describe relationships that are variously hierarchical (artificial intelligence "is a branch of" computer science) and associative (machine learning is "associated with" artificial intelligence in an unspecified way; artificial intelligence "intersects with" cognitive science), but conclude that they should all be treated as parent-child (i.e., hierarchical) relationships within the taxonomy. For MAG, the taxonomic structure appears intended to capture many different sorts of connections between concepts, and, in contrast to NISO (2010), we might think of the MAG relationships as a means of *contextualizing* one concept in terms of another rather than of *defining* a concept as a member, instance, or part of its parent. This also implies that the interpretation of a concept may change in the context of different relationships; for example, if *cognitive science* has parents *computer science* and *psychology*, then

cognitive science contextualized through the *computer science*$\longrightarrow$*cognitive science* relationship may describe a different point of view within the field than might be indicated by the *psychology*$\longrightarrow$*cognitive science* relationship.

Loosening the rules of hierarchical relationships in this way introduces ambiguity within the classification that reflects the type of real-world ambiguity discussed by Mai (2005) regarding different usages of specific terms in the context of different disciplines. For example, studies of a concept like "poverty" may have different priorities and purposes within economics than they might within sociology or women's studies (Klein, 1996b, as cited in Mai, 2005). According to Mai (2005), strict, formal definitions of disciplines are insufficient to properly classify work related to boundary-crossing concepts; a more thorough representation of the "activities, collaboration, and common goals" (p. 606) of researchers working across those disciplines is also necessary. However, such representations do not figure into the types of relationships standardized in NISO (2010).

Relatedly, Olson (2001) has discussed how the context established by hierarchical structures within classifications can impose limitations on the meanings of terms therein. According to Olson (2001), the structure of relationships amongst terms encodes specific perspectives, and the more limited the choices within the structure, the more problematically universal the perspective can appear to be. Olson (2001) gives an example from the Library of Congress Subject Headings (LCSH) of the term *Women* whose placement under the parent heading *Females* conceptually restricts it to a biological context. If we examine this heading further, however, we can see that many of the narrower terms under *Women*, such as *Minority women*, *Muslim women*,

Video games for women, *Advertising and women*, etc., relate more to specific social and cultural experiences of womanhood, rather than to biological aspects of women. But this socio-cultural dimension of the heading *Women* is invisible from within the LCSH *structure*, although it is, as Olson (2001) has noted, hinted at in the authority record for this subject heading, which lists the Library of Congress Classification range HQ1101-HQ2030.9 (Sociology) in the 053 (LC Classification Number) MARC field.

Interpreting hierarchical relationships in OpenAlex

Attempting to understand the structure of OpenAlex concepts as a traditional hierarchy (such as defined by NISO (2010)) brings about confusion regarding the meaning of apparently "multidisciplinary" concepts. We will instead revisit the examples we saw in table 5.2 using the ideas that the meaning of a term can be allowed to be plural, and that different relationships can lend different contexts to the interpretation of a term.

Figure 5.4 shows the full structure surrounding the concept *Literal translation*, which has L0 ancestors *Art*, *Computer science*, *Mathematics*, *Philosophy*, and *Sociology*. We see that *Literal translation* itself has two contexts (i.e., direct parents): *Source text*, and *Perspective (graphical)*. These contexts reveal the different meanings of *Literal translation* within the hierarchy: the translation of text from one language to another, and the translation of objects from one position or perspective to another. Exploring further up the hierarchy, we see other nuances to these interpretations; textual translation can have a sociological dimension (focused on the

translators themselves), it can be treated as a linguistic problem, and it can be a process carried out by machine—each of which contains a different construction of the concept of translation. Furthermore, object translation might be concerned with artistic techniques (which, again, may involve machined-assisted processes), or may instead apply to abstract, mathematical objects such as vectors translated in a plane. All of these distinct usages of *Literal translation* within OpenAlex, obfuscated under a single term, make themselves known when we trace through the contexts that exist around this concept within the hierarchy.

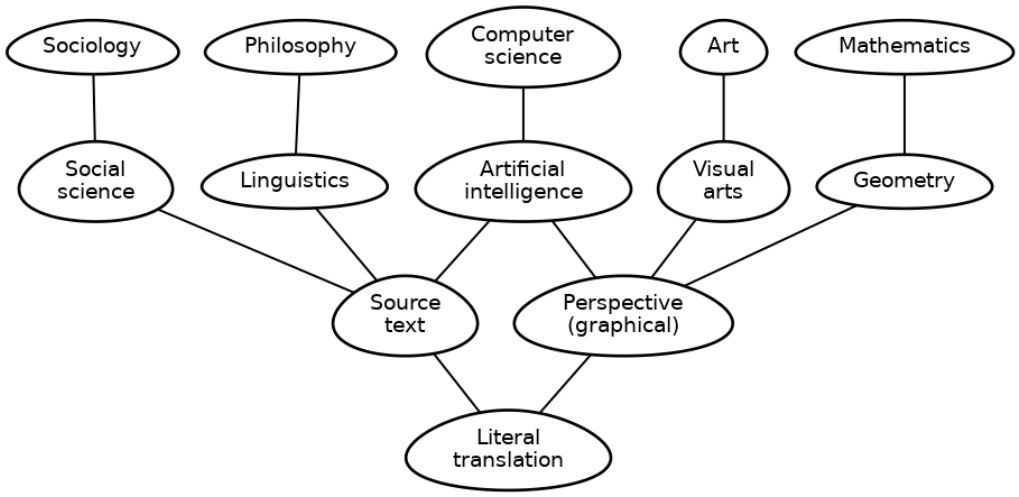


Figure 5.4: A graph depicting all paths through the OpenAlex hierarchy that lead to *Literal translation*.

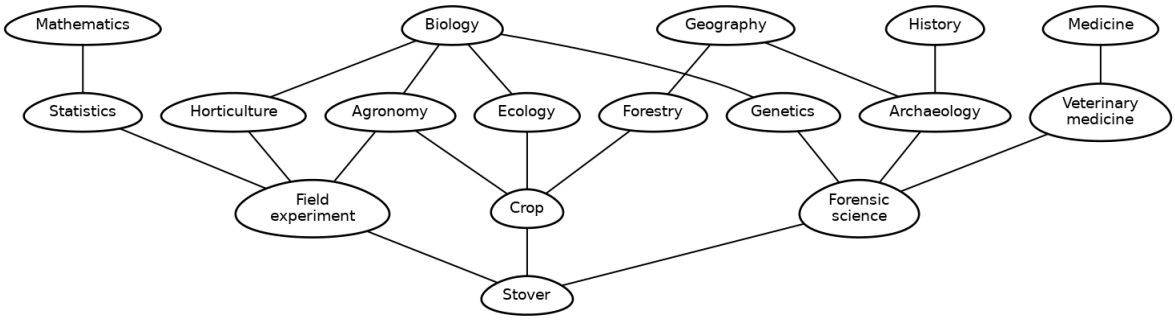


Figure 5.5: A graph depicting all paths through the OpenAlex hierarchy that lead to *Stover*.

Figure 5.5 depicts the contexts for another example from table 5.2, the concept *Stover* with L0 ancestors *Biology*, *Geography*, *History*, *Mathematics*, and *Medicine*. According to the description in OpenAlex, the term "stover" refers to "[l]eaves and stalks of field crops", and hence would appear to pertain (perhaps, mainly) to biological studies. Again, by focusing on one context at a time within the structure surrounding *Stover*, we can arrive at a better understanding of what interpretations the different relationships provide for this concept, the details of which exercise we leave with the reader. The interesting piece in this example is the concept *Field experiment* which may literally refer to experiments conducted in fields (aspects of which would be related to concepts such as *Horticulture* and *Agronomy*), but the phrase "field experiment" may also refer to the research method in which studies are conducted in real-world settings rather than in a controlled lab. This latter definition suggests an explanation for the relationship between *Field experiment* and *Statistics*, as statistical methods may often be applied towards analyzing the results of this type of field experiment (e.g., Chen et al., 2010; Sutter et al., 2013). The

Statistics⟶*Field experiment* relationship, which is the reason for *Mathematics* to be included in the ancestor set of *Stover*, is not even an associative relationship by the NISO (2010) definition cited earlier; it is merely a loose connection between these concepts that can be observed in research publications. This analysis of the contexts for *Field experiment* shows that *Mathematics* is arguably incorrectly included in the ancestor set for *Stover* via the *Statistics*⟶*Field experiment* relationship. Furthermore, as in the previous example of *Literal translation*, the different contexts for *Field experiment* give sufficiently different interpretations of the term to cause problematic ambiguity within the classification.

We will return to the topic of ambiguity in section 5.2.2, but we can see another example of how the different contexts for a concept can identify its ambiguous nature. In figure 5.6, we find that two of the contexts for *Egg incubation* are *Hatching* and *Incubation*, which on the surface appear easy to explain. However, *Hatching* itself has a total of three contexts, two of which—*Ecology* and *Animal science*—provide a biological interpretation, and one —*Visual arts*—in which it refers to the artistic technique of drawing close parallel lines. Given this understanding, *Hatching* should only be part of the concept hierarchy for *Egg incubation* as far the biological meanings of that term, and its artistic contexts should be disambiguated from that usage. Similar issues occur for the concept *Incubation*: in a psychotherapy context, incubation refers to a type of unconscious mental process, while in a biochemistry context, it refers to the maintenance of a controlled environment (such as for organism growth). For many applications, these differences in usage and meaning would merit splitting *Incubation* into different, disambiguated concepts.

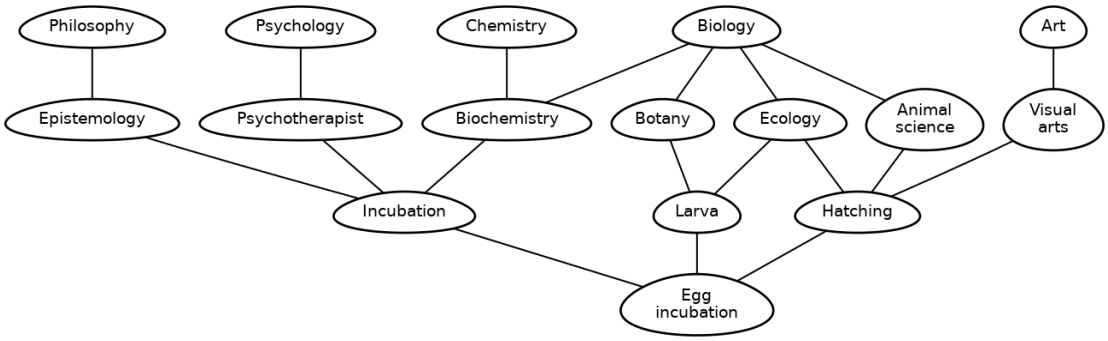


Figure 5.6: A graph depicting all paths through the OpenAlex hierarchy that lead to *Egg incubation*.

The final poly-hierarchical example we will discuss in this section is given in figure 5.7, and shows relationships that are not only ambiguous, but arguably completely erroneous. In this example, both of the parent-child relationships, *Composite material* $\longrightarrow$ *Composite number*, and, *Composite number* $\longrightarrow$ *Wood-plastic composite*, appear to be based mainly, if not solely, on the presence of the word "composite" in each of these terms. There is no semantic connection between *Composite number* and the two other terms to justify these relationships in any other way. We would therefore conclude that the inclusion of *Composite number* in the parent contexts for *Wood-plastic composite* is in error, and by extension, neither *Computer science* nor *Mathematics* should be present in the ancestor set of *Wood-plastic composite* at all.

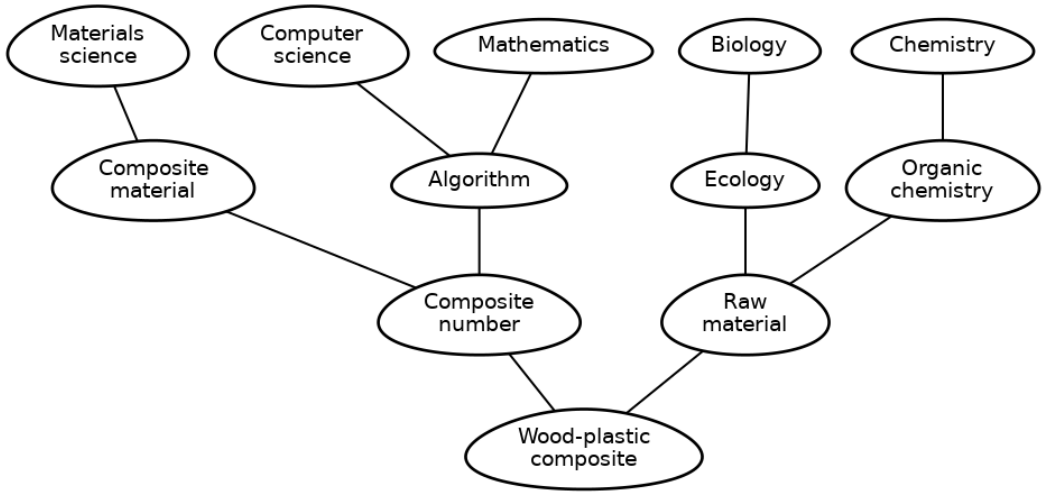


Figure 5.7: A graph depicting all paths through the OpenAlex hierarchy that lead to *Wood-plastic composite*.

5.2.2 Ambiguity of concept terms

Ambiguity arises in a controlled vocabulary when a homophonic or polysemic term is not accompanied by any other qualifying information to specify its meaning (NISO, 2010). We saw in section 5.2.1 some examples of ambiguity in OpenAlex, where the same term, such as *Hatching*, referred to different concepts within different disciplines. From the discussion of poly-hierarchical relationships presented in that section, it follows that any time there are multiple parents for a concept (which signify different contexts for that concept) there is a risk that ambiguity may be present within that concept term. In the previous examples, this risk appeared to be higher when the contexts stemmed from disciplines that were less closely related, such as *Biology* and *Art*, than when they were more closely related subjects, such as *Agronomy* and *Horticulture*.

Ambiguity markers in OpenAlex

The use of parenthetical qualifiers in term names is a standard approach to marking different usages of an ambiguous term (NISO, 2010), and while such qualifiers are present in some OpenAlex concepts, such as *Absorption (acoustics)*, these appear to not have been applied by MAG or OpenAlex. As we will discuss in chapter 6, the MAG concept names were drawn from Wikipedia, and, as a result, any qualifiers in the concept names appear to originate from the corresponding Wikipedia article title. Thus, the disambiguation is effective only in the context of Wikipedia and does not affect the usage of the term in OpenAlex, which is often not limited to the context given in the qualification. NISO (2010) has also mandated that qualifiers "***should*** be added to each homograph" (p. 21, emphasis in original), but for each qualified term in OpenAlex, there is no corresponding term with the same homograph (with or without a different qualifier), which makes the qualifier unnecessary. (For example, there is no other concept in OpenAlex with a name like "Absorption (X)".) On the other hand, there are a small number of concepts (approximately 850) whose descriptions state that these are titles of disambiguation pages in Wikipedia. These indicate concepts that possibly *could* be split into multiple, qualified terms, as per the different usages listed on their disambiguation pages.

The discussions in preceding sections suggest that it is worthwhile to distinguish between two different types of ambiguity present within the OpenAlex hierarchy: ambiguity arising through polysemy or homophony, and semantic ambiguity that may contribute to (machine-inference of) incorrect relationships between concepts. The former type of ambiguity, where the same term has multiple meanings, can occur

in different ways. The first is where a word or phrase might be used alternatively as the proper name of something else, such as in the OpenAlex concepts *Subtext*[1], *FAUST*[2], and *SALSA*[3]. The first two, apart from their literary denotations, are also the names of programming languages, and the last one has many distinct referents in OpenAlex, including the well-known Latin dance, but also a type of herb (Li et al., 2005), a text corpus (Burchardt et al., 2006), a system for log analysis (Tan et al., 2008), etc.

The second way polysemic ambiguity occurs in OpenAlex is when a term can have different interpretations in different contexts, but not completely different meanings, as with the concept *Improvisation*[4] which falls under the L0 disciplines of *Art* and *Physics*. OpenAlex describes this concept as a "process of devising a solution to a requirement in an ad hoc fashion". In artistic contexts, however, improvisation refers to various spontaneous forms of artistic creation and expression rather than spontaneous problem-solving. Another example is the concept *Narrative*[5], which is described as an "account of a series of related events or experiences". Its L0 ancestors are *Art* and *Philosophy*, and it has two parent contexts, *Literature* and *Linguistics*. While the senses of "narrative" under both of these contexts generally agree with each other and with the given description, a literary study of narrative would likely exhibit a different analytical point of view than would one focused on narrative structure from a linguistic perspective.

[1] `https://openalex.org/C2779703954`
[2] `https://openalex.org/C2776056051`
[3] `https://openalex.org/C2776317494`
[4] `https://openalex.org/C125468537`
[5] `https://openalex.org/C199033989`

Finally, the third category of polysemic ambiguity in OpenAlex, which may overlap with the other two, contains terms that are applied according to a usage that is *not* captured within any of their contexts. An example is the concept *Local color*[6], described as the "natural color of an object unmodified by adding light and shadow or any other distortion; best seen on a matte surface, due to it not being reflected, and therefore distorted". The description in OpenAlex matches the artistic context for this term, but the expression "local colour" can also refer to the unique customs or characteristics of a particular place, which is apparent from the titles of some of the works tagged with this concept, such as, "The Significance of Charles W. Chesnutt's 'Conjure Stories'"[7]; "Folk for Whom? Tourist Guidebooks, Local Color, and the Spiritual Churches of New Orleans[8]"; "Local-Color Literature and Modernity: The Example of Jewett"[9]; and, "New Trend of Local-color Novel Writing Since 1990s"[10]. At the same time, other works under this concept clearly refer to artistic usage: "Color Histogram Image Retrieval Based on Spatial and Neighboring Information"[11]; "Consistent Segmentation Based Color Correction for Coarsely Registered Images"[12]; "A Proposal of Improvement Method of Local Color Correction"[13]; etc. While these examples show the conflation of a technical artistic term with an unrelated idiomatic expression, the idiomatic usage is not suggested at all by the contexts for *Local color*. As shown in figure 5.8, two of the three paths for this concept indicate

[6] `https://openalex.org/C2779255053`
[7] `https://openalex.org/W205042996`
[8] `https://openalex.org/W4248601133`
[9] `https://openalex.org/W2254058509`
[10] `https://openalex.org/W2358676189`
[11] `https://openalex.org/W2351678691`
[12] `https://openalex.org/W2532542564`
[13] `https://openalex.org/W2967591619`

usage related to visual arts or computer graphics, while the third path, *Computer science*$\longrightarrow$*Artificial intelligence*$\longrightarrow$*Local color*, is difficult to find any interpretation for.

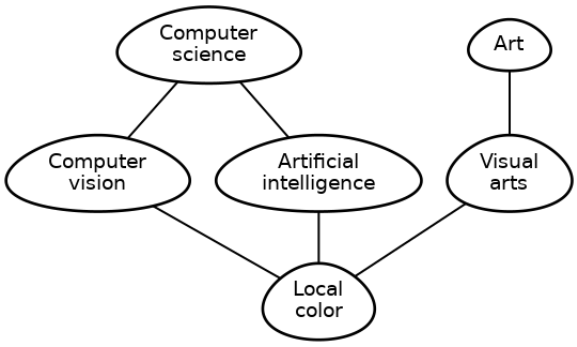


Figure 5.8: A graph depicting all paths through the OpenAlex hierarchy that lead to *Local color*.

Aside from polysemic ambiguity, we also find ambiguity within OpenAlex resulting from semantic confusion of terms that overlap but refer to distinct concepts. An example previously discussed involved the concepts *Composite material*, *Composite number*, and *Wood-plastic composite*, which appeared to have been connected hierarchically with each other likely due to the occurrence of "composite" in each of these terms. Another example is of several technical concepts about "architecture" in the computer science or engineering sense, such as *Software architecture*, *Systems architecture*, *Service-oriented architecture*, etc., being given the parent context *Architecture*, in the visual arts and archaeological sense, again, most likely due to the repetition of the term "architecture" (see figure 5.9).

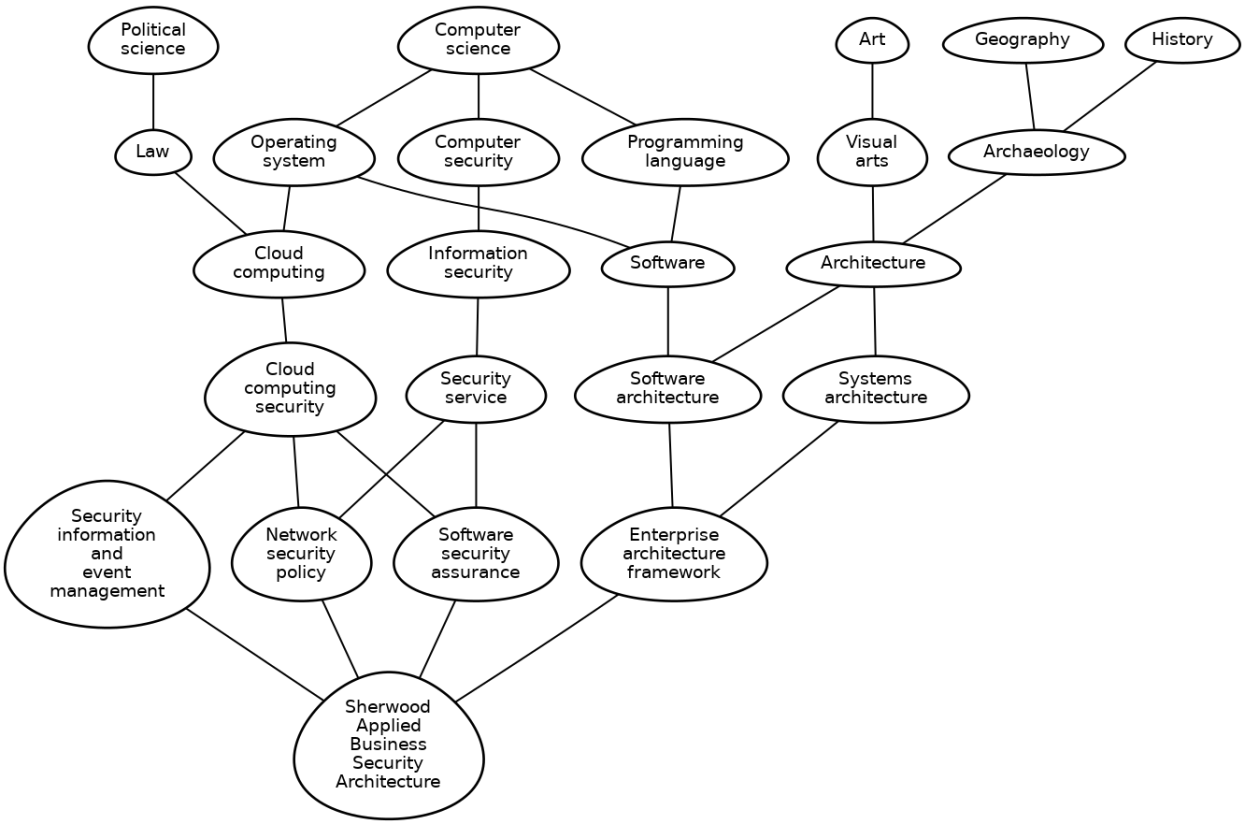


Figure 5.9: A graph showing conflated meanings of various concepts related to "architecture".

Furthermore, such ambiguity does not always occur across dissimilar disciplines. One example can be seen in the relationship *Alice (programming language)*$\longrightarrow$*Alice and Bob*. These are both terms that may be found in computer science literature, though unlikely to co-occur in the same publications, and wholly unrelated to each other. The concept *Alice and Bob* refers to fictional characters often used to name the different parties involved in cryptographic algorithms, while the Alice programming language is an educational language that takes its name from the title character

of *Alice in Wonderland*. Again, these appear to have been put into hierarchical relationship with each other primarily because of the common term "Alice" in both concepts.

Systematic exploration of ambiguity

To identify the potential scope of ambiguous concepts in OpenAlex, we started by querying the OpenAlex database for the set of L0 ancestors for each concept that had at least two distinct ancestors. We computed the full set of paths through the concept hierarchy for each of these concepts, and generated images visualizing each concept's path graph using the graphviz[14] command-line application. In total, there were 46,361 image files generated after this initial phase, of which we analyzed a subset by random inspection. We then generated a version of this paths dataset in JSON Lines format and explored it heuristically using Python. We searched for structural patterns that would indicate the degree of similarity among all the paths for a concept, with the hypothesis that a concept whose paths were more similar to each other might be less likely to be ambiguous. The two main features of interest that we found were the presence of convergence amongst a concept's paths, and the distribution of a concept's paths among its L0 ancestors.

[14]https://graphviz.org/doc/info/command.html

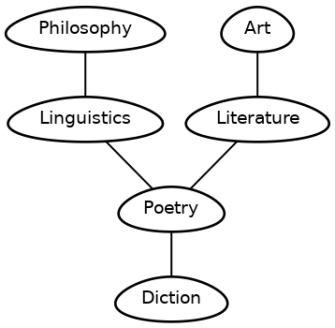


Figure 5.10: Concept path graph for *Diction*, showing the convergence of paths from both L0 ancestors, *Art* and *Philosophy*. This concept has depth of convergence 2.

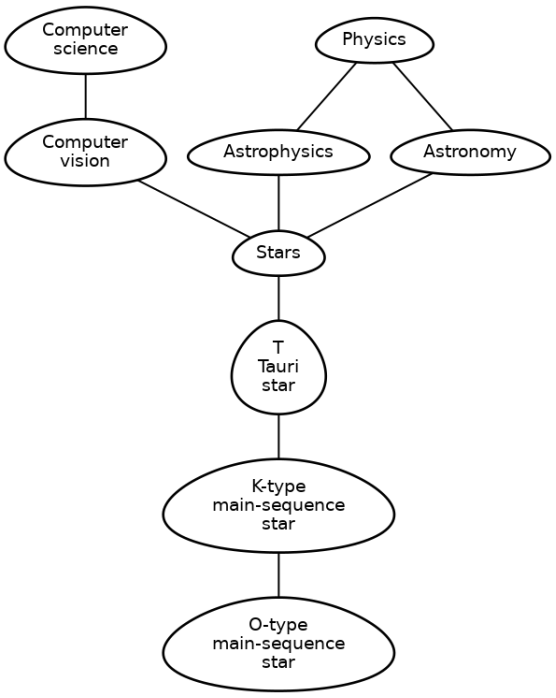


Figure 5.11: Concept path graph for *O-type main sequence star*, showing the convergence of paths from both L0 ancestors, *Computer science* and *Physics*. This concept has depth of convergence 4.

The *depth of convergence* for a concept was identified by enumerating all the

paths for the concept and computing the number of nodes in the largest common "tail" among all paths. Figures 5.10 and 5.11 show two examples of converging paths, in which the longest common tails are, respectively, *Poetry*$\longrightarrow$*Diction*, and *Stars*$\longrightarrow$*T Tauri star*$\longrightarrow$*K-type main sequence star*$\longrightarrow$*O-type main sequence star*.

We found that concepts with higher depths of convergence tended to have L0 ancestors that were in closely related disciplines, or that belonged to a specific intersection of two L0 disciplines, such as *Mathematical economics*. To quantify this observation, we arranged the 19 L0 concepts into two groups as follows:

- Group 1: *Biology*, *Chemistry*, *Computer science*, *Engineering*, *Environmental science*, *Geography*, *Geology*, *Mathematics*, *Materials science*, *Medicine*, *Physics*, *Psychology*
- Group 2: *Art*, *Business*, *Economics*, *History*, *Philosophy*, *Political science*, *Psychology*, *Sociology*

This division of concepts was not intended to be definitive in any way, but only to broadly separate scientific and humanities concepts, so that if a concept had ancestors that belonged to both of these groups, it could be flagged as having significantly diverse ancestors to warrant further inspection. We thus included *Psychology* in both groups, due to its close association with both medical/biological and sociological disciplines. Using these groupings, we tabulated how many concepts at a given depth of convergence had ancestors that all belonged to only one of the groups, summarized in table 5.3.

Depth of convergence	Total number of concepts	% concepts with related ancestors
1	27254	80%
2	12578	85%
3	4741	92%
4	1776	92%
5	12	100%

Table 5.3: The number of concepts with a given depth of convergence. This data was only computed for concepts with more than one L0 ancestor.

We can see that crossover between these groups was generally not common, and even at a depth of convergence of 2, the likelihood that a concept had related ancestors was high. Through iterative random inspection of the results, we found that concepts with high convergence *and* related ancestors were generally not polysemic or homophonic, while those with high convergence and more varied ancestors generally fell into two categories: concepts with an interdisciplinary nature that existed within a specific intersection of disciplines (such as *Good agricultural practice* under both *Biology* and *History*); and concepts where there appeared to be some erroneous relationships at the L1/L2 level which then propagated to lower levels. An example of the latter is the inclusion of *Political science* among the ancestors of *Security testing* due to the chain of relationships, *Political science* $\longrightarrow$ ***Law*** $\longrightarrow$ ***Cloud computing***, as a result of which, all descendants of *Cloud computing* (which includes *Security testing*) were included under *Political science*. We refer to this category of errors as *transitivity errors*, and discuss in more detail the reasons behind their occurrence in chapter 7.

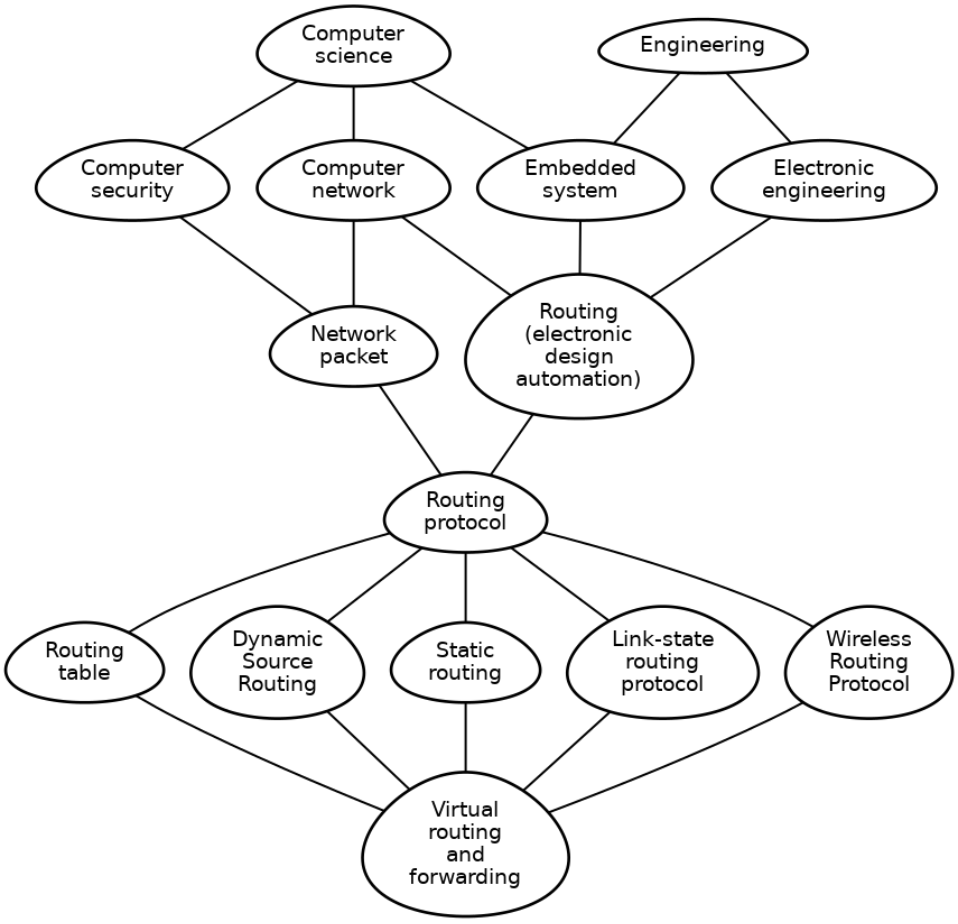


Figure 5.12: A graph depicting all paths through the OpenAlex hierarchy that lead to *Virtual routing and forwarding*.

We analyzed the *path distribution* for a concept by tabulating the total number of paths originating from each of its ancestors and then computing the population standard deviation, σ, of this data. For example, the concept *Virtual routing and forwarding* (figure 5.12) has 20 total paths originating from *Computer science* and 10 from *Engineering*, and therefore has $\sigma = 5.0$. We used the population standard deviation to characterize path distribution as follows: *imbalanced* if $\sigma > 1$; *slightly*

imbalanced if $0.5 < \sigma \leq 1$; and *balanced* if $\sigma \leq 0.5$. This characterization allowed us to identify concepts with outliers among their ancestors, which could be an indication of either ambiguity or errors within the relationships for that concept.

When concepts had unrelated ancestors, we found that both ambiguity and transitivity errors seemed to correspond with a balanced path distribution together with high path convergence. Concepts with balanced path distribution among related ancestors tended to have fewer instances of errors, though we still found examples of ambiguity among such concepts. Concepts with slightly imbalanced path distribution and related ancestors often had a smaller number of ancestors in total, with one "odd ancestor out" that was included in the paths either through a transitivity error (as in, ***Physics*** $\longrightarrow$ ***Acoustics*** $\longrightarrow$ *Music education* $\longrightarrow$ *Musical composition*), or through an interdisciplinary concept (such as *Biochemistry*). Concepts with imbalanced path distributions and related ancestors exhibited very few erroneous relationships, transitivity errors, or ambiguity issues, while those with unrelated ancestors overwhelmingly appeared to usually have at least one incorrect relationship. This last category (imbalanced path distribution with unrelated ancestors) was the strongest heuristic we uncovered for identifying (a subset of) concepts that would be likely to have incorrect or ambiguous hierarchical relationships among their paths.

5.3 Usage Patterns

5.3.1 Utility of concept hierarchy for research evaluation

There have been many research evaluation studies conducted using MAG or OpenAlex, but very few of them make substantial use of the concept hierarchy itself. For instance, in their analysis of transdisciplinary publications, Huang et al. (2022) compared papers across disciplines using only the categorizations provided by the first level (L0) of the hierarchy (from which they also removed seven concepts due to insufficient disciplinarity). Liu et al. (2022) compared reference indicators across different disciplines using the L1 rather than L0 concepts in MAG, again mostly ignoring the hierarchical relationships between those concepts and the rest of the hierarchy. Similarly, Xu et al. (2024) chose three of the L0 concepts (*Computer science*, *Business*, and *Sociology*) for their analysis, leaving the rest of the hierarchy alone; Zafar et al. (2023) used both L0 and L1 concepts, but ultimately focused their study on six L1 disciplines in isolation, chosen for unspecified reasons; and Bu et al. (2021) used only the 19 L0 concepts to select papers from each discipline, in order to implement their own topic generation.

These studies suggest that the hierarchical structure of the concepts has not been particularly useful for research evaluation studies so far. This may be due in part, as Hug et al. (2017) have observed, to the highly specific nature of concepts at lower levels, and to the apparent incoherence of relationships even within the first two levels. While, as we have seen, the hierarchical structure of the OpenAlex concepts is able to convey a lot of information about interconnections between disciplines, it is

also possible that extracting this information cleanly from the hierarchy is not worth the effort for most research evaluation studies.

5.3.2 OpenAlex API and catalogue

Drabinski (2013) has noted that the public catalogue is where "the ideology of the knowledge organization structure is apparent, and therefore where the contingency of classification and subject description are most obvious" (p. 106). A comprehensive overview of the OpenAlex concept hierarchy should thus include an examination of the technical means provided to users for accessing, understanding, and using that hierarchy. The OpenAlex concepts are mediated to users through OpenAlex's API and catalogue, both of which allow users to retrieve information about concepts and to perform searches for works that are tagged with particular concepts. Details provided through the API include the display name, hierarchy level, and description of a concept; the URLs of corresponding concepts in Wikidata and Wikipedia; the number of works that are tagged with the concept; a list of related concepts; and a list of the concept's ancestors. While these details are very useful for understanding the meaning and usage of a given concept, it is difficult for users to gain a broader view of the concept hierarchy from the API responses, as they do not include the list of children for each concept, which means that navigating the hierarchy cannot be done only through API calls.

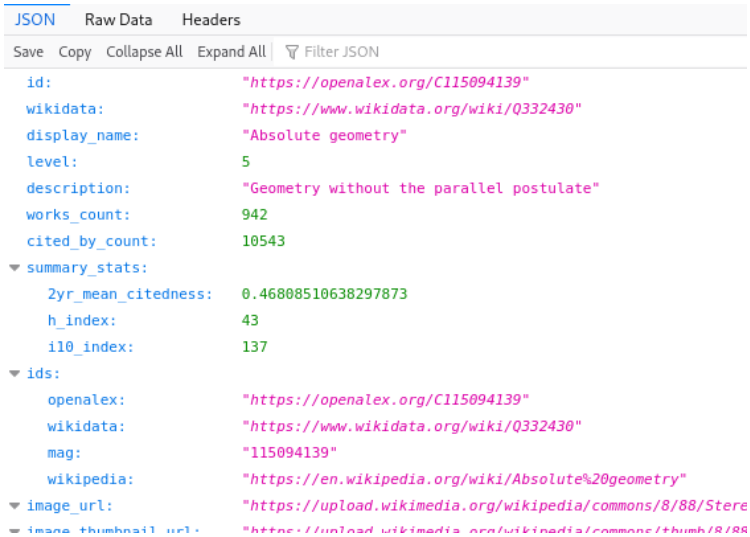


Figure 5.13: Example OpenAlex API response for the concept *Absolute geometry*. `https://api.openalex.org/concepts/C115094139`

The OpenAlex catalogue contains a page for each concept, listing some statistics related to the concept's prevalence in the catalogue, and containing links to the more detailed API response and to a search interface where users can find works that are tagged with that concept. Whereas the API supports both Boolean 'and' and 'or' searches for works tagged with multiple concepts, the catalogue interface only supports Boolean 'or'. Concept search also appears to no longer be supported through the catalogue, nor is there any navigational information provided to assist users in browsing concepts or exploring the concept hierarchy. OpenAlex also makes available a spreadsheet[15] of the concepts, from which the hierarchical structure can be inferred, though, based on the date given in its title, this sheet may not be actively maintained.

[15] `https://docs.google.com/spreadsheets/d/1LBFHjPt4rj_9r0t0TTAlT68NwOtNH8Z211BMsJDMoZg/edit?gid=575855905#gid=575855905`

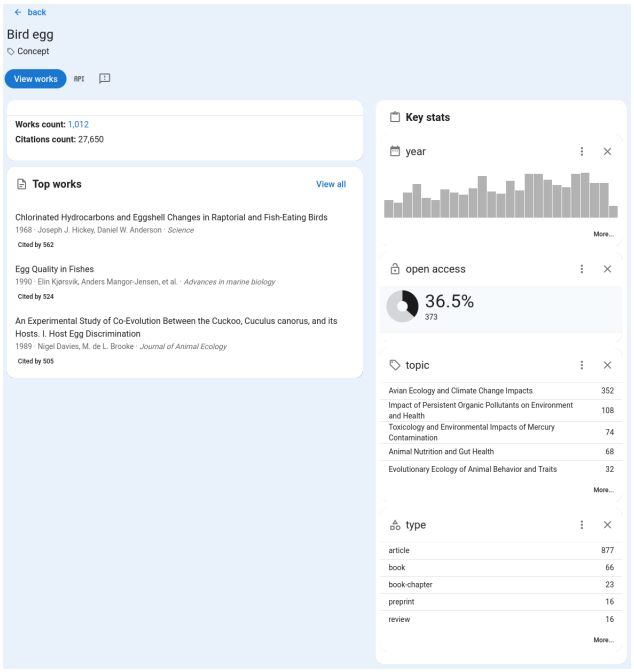
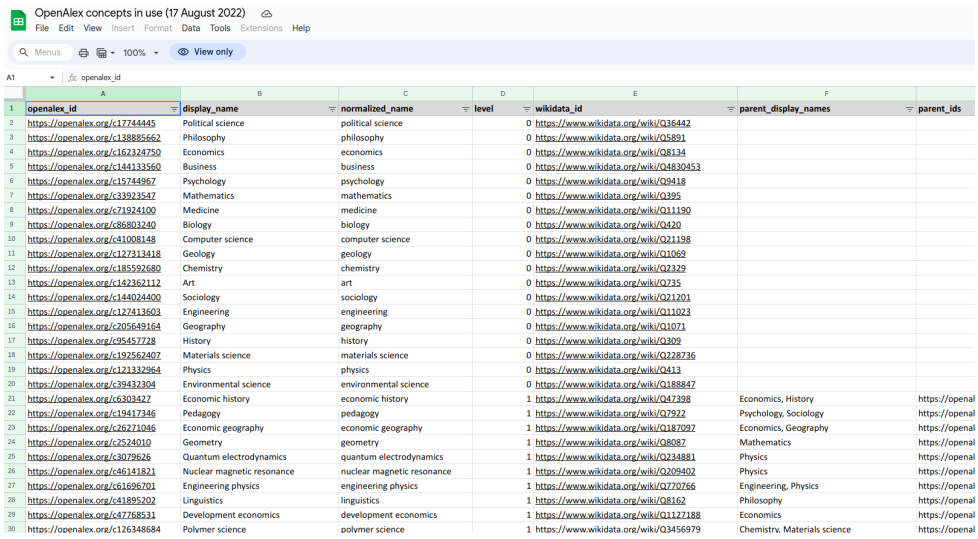


Figure 5.14: Screenshot of the OpenAlex catalogue page for the concept *Bird egg*.



	openalex_id	display_name	normalized_name	level	wikidata_id	parent_display_names	parent_ids
2	https://openalex.org/c17744445	Political science	political science	0	https://www.wikidata.org/wiki/Q36442		
3	https://openalex.org/c138885662	Philosophy	philosophy	0	https://www.wikidata.org/wiki/Q5891		
4	https://openalex.org/c162324750	Economics	economics	0	https://www.wikidata.org/wiki/Q8134		
5	https://openalex.org/c144133560	Business	business	0	https://www.wikidata.org/wiki/Q4830453		
6	https://openalex.org/c15744967	Psychology	psychology	0	https://www.wikidata.org/wiki/Q9418		
7	https://openalex.org/c33923547	Mathematics	mathematics	0	https://www.wikidata.org/wiki/Q395		
8	https://openalex.org/c71924100	Medicine	medicine	0	https://www.wikidata.org/wiki/Q11190		
9	https://openalex.org/c86803240	Biology	biology	0	https://www.wikidata.org/wiki/Q420		
10	https://openalex.org/c41008148	Computer science	computer science	0	https://www.wikidata.org/wiki/Q21198		
11	https://openalex.org/c127313418	Geology	geology	0	https://www.wikidata.org/wiki/Q1069		
12	https://openalex.org/c185592680	Chemistry	chemistry	0	https://www.wikidata.org/wiki/Q2329		
13	https://openalex.org/c142362112	Art	art	0	https://www.wikidata.org/wiki/Q735		
14	https://openalex.org/c144024400	Sociology	sociology	0	https://www.wikidata.org/wiki/Q21201		
15	https://openalex.org/c127413603	Engineering	engineering	0	https://www.wikidata.org/wiki/Q11023		
16	https://openalex.org/c205649164	Geography	geography	0	https://www.wikidata.org/wiki/Q1071		
17	https://openalex.org/c95457728	History	history	0	https://www.wikidata.org/wiki/Q309		
18	https://openalex.org/c192562407	Materials science	materials science	0	https://www.wikidata.org/wiki/Q228736		
19	https://openalex.org/c121332964	Physics	physics	0	https://www.wikidata.org/wiki/Q413		
20	https://openalex.org/c39432304	Environmental science	environmental science	0	https://www.wikidata.org/wiki/Q188847		
21	https://openalex.org/c6303427	Economic history	economic history	1	https://www.wikidata.org/wiki/Q47398	Economics, History	https://openal
22	https://openalex.org/c19417346	Pedagogy	pedagogy	1	https://www.wikidata.org/wiki/Q7922	Psychology, Sociology	https://openal
23	https://openalex.org/c26271046	Economic geography	economic geography	1	https://www.wikidata.org/wiki/Q187097	Economics, Geography	https://openal
24	https://openalex.org/c2524010	Geometry	geometry	1	https://www.wikidata.org/wiki/Q8087	Mathematics	https://openal
25	https://openalex.org/c3079626	Quantum electrodynamics	quantum electrodynamics	1	https://www.wikidata.org/wiki/Q234881	Physics	https://openal
26	https://openalex.org/c46141821	Nuclear magnetic resonance	nuclear magnetic resonance	1	https://www.wikidata.org/wiki/Q209402	Physics	https://openal
27	https://openalex.org/c61696701	Engineering physics	engineering physics	1	https://www.wikidata.org/wiki/Q770766	Engineering, Physics	https://openal
28	https://openalex.org/c41895202	Linguistics	linguistics	1	https://www.wikidata.org/wiki/Q8162	Philosophy	https://openal
29	https://openalex.org/c47768531	Development economics	development economics	1	https://www.wikidata.org/wiki/Q1127188	Economics	https://openal
30	https://openalex.org/c126348684	Polymer science	polymer science	1	https://www.wikidata.org/wiki/Q3456979	Chemistry, Materials science	https://openal

Figure 5.15: Screenshot of the OpenAlex concepts spreadsheet.

The designs of both the API and catalogue interface make grappling with the hierarchical structure of the OpenAlex concepts inconvenient. In fact, the catalogue gives no sign to users that there is an underlying hierarchy at all. This obscurity might be justified by the observation that most studies tend not to engage meaningfully with the hierarchical structure of the concepts, or it might actually reinforce that behaviour. Without the ability to enagage more deeply with the concept hierarchy, users cannot easily interrogate how well it serves their purposes, what assumptions exist with regards to its correctness and completeness, and how those assumptions bear on any studies conducted using these concepts.

5.4 Conclusion

In this chapter, we analyzed the non-standard, poly-hierarchical structure of the OpenAlex concepts, and explored some strengths of that structure with respect to representing more complex relationships between concepts that cross disciplines. We also discussed some of the errors and problems with ambiguity within the OpenAlex classification. Finally, we discussed the usage patterns of researchers working with the MAG and OpenAlex concepts, and the degree to which the concepts and their structure are made legible to users.

Chapter 6

Examining the Construction of the OpenAlex Concept Hierarchy

6.1 Introduction

To understand the inner workings of the OpenAlex concept hierarchy, we need to first examine the models and algorithms used to construct the MAG fields of study on which the OpenAlex concepts are heavily based. As per Shen et al. (2018), the construction of the MAG classification comprises three stages, in order: term selection, indexing, and hierarchy construction. This sequence is unusual in that the structuring of the classification is not independent from the indexing step but follows from it. In this chapter, we will discuss the technical approaches taken by MAG and OpenAlex in implementing each of the three stages above, and compare those implementations with other perspectives from classification theory and practice.

6.2 Term Selection

6.2.1 NISO guidelines for term selection

We will start by comparing the term selection process for the MAG concept hierarchy with the approach recommended by NISO (2010). NISO (2010) prescribe several principles for choosing controlled vocabulary terms so that they are well suited for indexing purposes, summarized as follows:

- Homographs should be avoided, or if necessary, disambiguated using parentheticals.
- Scope notes should be used to clarify the meanings of terms.
- There should be an awareness of the general types of the entities present in the controlled vocabulary (places, objects, disciplines, etc.) during vocabulary compilation, as these types can have a bearing on, e.g., how the vocabulary is structured.
- Use of highly specific terms should be limited due to the cognitive burden of managing the increased vocabulary size and associated relationships, but for complex concepts that have *very high* usage, pre-coordinated terms can simplify indexing.

The specific directives of these principles change a little when the indexing happens primarily through a machine-learning process—as in the case of MAG and OpenAlex—but the underlying ideas are still applicable to a machine-derived classification. We discuss the machine learning perspective on the above principles below:

- Disambiguation: It is crucial that the representations of distinct concepts are distinguishable *within the machine-learning model*, but this may not only involve qualifying concept names. Differentiation among concepts may also be achieved by selecting and extracting concept features that capture important differences in meaning and usage. If there is insufficient disambiguation in the representation of entities, then there is a risk that the model will misclassify items because it has not learned how to distinguish different usages of the same term. (This issue will be explored further when discussing the indexing process in section 6.3.)

- Scope notes: As they are primarily intended to ensure consistent application of the classification, scope notes in a machine-learning context should include not just (human-readable) definitions of terms, but also explanations of what features were extracted by the model for each term, and what source data was used in the representation of each term, in the spirit of datasheets for datasets (Gebru et al., 2021). Such documentation would allow users to have some insight into the "definition" of each term from the point of view of the model, and how and why the classifier might have applied each term to different works.

- Entity types: The classifier model should have a sufficiently rich understanding of the semantics of the entities that it assigns, as this information may help improve its reliability in distinguishing different usages of similar terms. The importance of entity types to other aspects of the classification construction may depend on the intended structure of the vocabulary, as we will see when discussing the hierarchy construction in section 6.4.

- Term specificity: Limiting the use of pre-coordinated terms in order to keep the vocabulary more manageable may be less of a concern for a machine-based classifier that can handle tens of thousands of terms more easily than could human indexers. However, it should still be considered whether users of the classification would actually benefit from a classification of enormous size.

What we have found from a review of the literature is that the technical processes and models used by MAG for term selection, as given in Shen et al. (2018), did not appear to deliberately afford much consideration to the above principles. Since concept representation for MAG included features derived from Wikipedia article text (Shen et al., 2018), we expect that there was a degree of incidental disambiguation within the model between terms that may have had similar headings but referred to distinct concepts (with distinct article text); however, due to the more simplistic treatment of concepts in the OpenAlex model (as we will discuss further in section 6.3.1), it is unclear whether such disambiguation was available to the OpenAlex concept tagger as well.

Terms were selected for MAG by manually curating an initial set of fields of study from Wikipedia entities, and then exploring the Wikipedia link graph outwards from those entities according to the idea that "if the majority of an entity's nearest neighbours are [fields of study], then it is highly likely [a field of study] as well" (Shen et al., 2018, p.3). Thus, entity type was nominally taken into account in this process by eliminating candidate concepts whose entity type was not an appropriate field of study. However, the definition of "field of study" for this entity type analysis included any type of entity that could be the subject of a paper, rather than entities

that represented whole fields or disciplines. For example, Shen et al. (2018) state that "person" would not be a valid entity type for the MAG classification (as in, an entity that represents a specific person), but "protein" would be. It does not appear that a more in-depth representation of entity types beyond "field of study" was used during the indexing or hierarchy construction phases of the MAG model (Shen et al., 2018).

Finally, it does not appear that MAG took any steps to limit the specificity of concepts when selecting term labels from Wikipedia, and thus the specificity of concepts in MAG and OpenAlex corresponds exactly to the specificity of Wikipedia article titles (Shen et al., 2018). The consequence of this choice is a proliferation of highly specific concepts (such as the names of individual proteins, genes, chemical compounds, etc.), with the resulting hierarchical relationships often being convoluted, as illustrated by the example in figure 6.1.

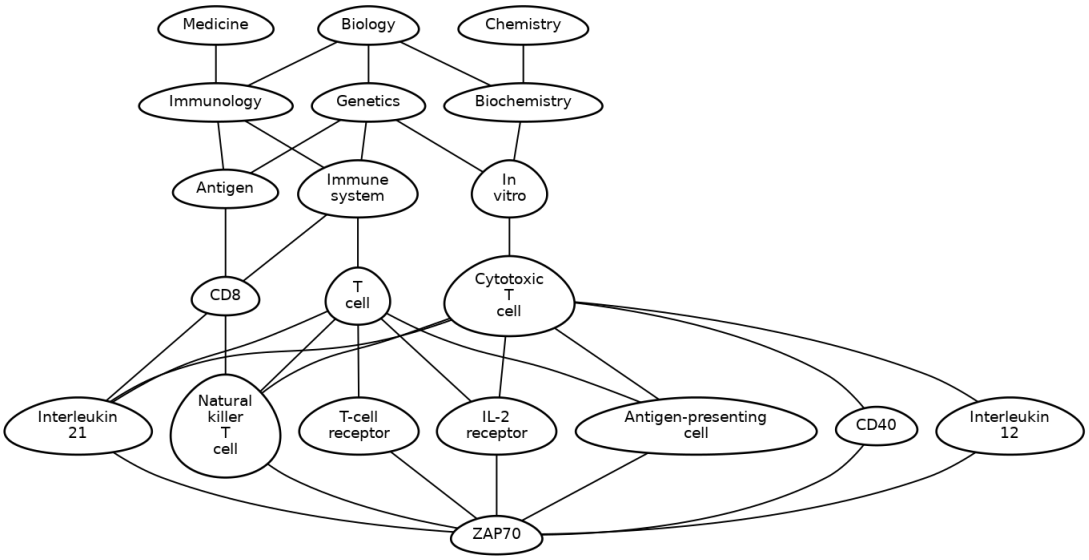


Figure 6.1: A graph showing the interconnected relationships among highly specific, protein-related concepts.

6.2.2 Literary and scientific warrant

As Bullard (2017) has noted, "[w]arrant is an element of all classification design, regardless of whether it is named as such and regardless of the particular technological basis of the system" (p. 76). While warrant is not explicitly mentioned in the MAG or OpenAlex literature, Wang et al. (2019) have explained that their inability to find agreement among existing taxonomies led them to develop the MAG taxonomy "solely from the document collection" (p. 5). The wording of this statement strongly suggests literary warrant as the basis for the MAG classification, even though, in an earlier paper, Shen et al. (2018) stated that MAG leveraged "Wikipedia articles as the source of concept discovery" (p. 2). As a possible reconciliation of these "contradictory positions" (Bullard, 2017, p. 77), we argue that, in the context of

machine-derived classifications, it is important to consider not just where the names of terms are taken from, but also how the representations of those terms are constructed within the machine-learning model; and, therefore, that questions of warrant should be extended to term representations as well as to their human-readable labels. Wang et al. (2020) have noted that, for MAG, "[i]nternally, each concept is a collection of semantic representations of textual passages", and that "an attempt is made to find, for each concept, a tutorial article on the web that is closed in semantics and uses the title of the article as the human-readable label for the concept" (p. 408). Thus, when Wang et al. (2019) mention using solely the document collection to build the classification, they are referring to building the *internal representations* of the concepts, while Shen et al. (2018), in mentioning Wikipedia as the source of concept discovery are referring to sourcing the *external labels* of the concepts.

While the internal representations of the concepts in MAG do comprise text taken from publications in the collection, they also include text from outside of the collection (Shen et al., 2018), and thus, are not based exclusively on literary warrant. In fact, the usage of Wikipedia indicates a combination of literary and scientific warrant in the construction of the MAG concepts. The reasoning given by MAG for extracting term labels from Wikipedia rather than from the collection is that it allowed them to "generate high-quality *concepts* with comprehensive coverage" (Shen et al., 2018, p. 2, emphasis in original). The stated goals of "high-quality" (in terms of human-readability) and "comprehensive coverage" suggest the use of scientific warrant, which puts emphasis on singular ways to refer to concepts and a single correct structure in which those concepts can be organized (Bullard, 2017).

Constructing concept representations using literary warrant can arguably be a good approach for supporting research evaluation activities, a primary use-case for both MAG and OpenAlex, since the proximity these representations have to the research collections themselves allows the classification models to more easily uncover patterns within the collection. Such patterns may help, for example, in recognizing emergent trends within and across research disciplines. However, there is also a need for some consistency in the definitions and coverage of different disciplines, in order to make interoperation possible, such as for comparing performance indicators across institutions or within a given field. While the use of scientific warrant can help serve this need, we did not find any sufficient justification for the decision to use Wikipedia in particular as a universal and authoritative correct structure for research classification in any of the MAG or OpenAlex literature.

6.3 Indexing

The process of indexing generally consists of two fundamental steps: analysis of a document to determine its subject(s), followed by translation of that subject matter into the terms of the classification (Mai, 2005). For the MAG and OpenAlex classification models, the analysis of documents was accomplished through feature extraction, and the translation to index terms, through machine-learning classification. Features are attributes derived from items that, when chosen effectively, can serve as more useful input to a machine-learning model than the raw data of the items themselves. MAG and OpenAlex differed notably on both of the above

steps; the features extracted by MAG incorporated more network information and were based on more pieces of metadata than those extracted by OpenAlex, while the main differences between the machine-learning models used by each system lay in the amount of context made available to each model.

6.3.1 Entity representation and feature extraction

The first step in constructing the machine-learning models for both MAG and OpenAlex required creating internal representations of publications, publication venues, and concepts, from which features could be extracted (Shen et al., 2018; OpenAlex, n.d.). These internal representations comprised various metadata related to each publication (summarized in table 6.1). The representation of publications in the MAG model combined publication metadata (such as titles and abstracts) with metadata from the publication's references, from its citing publications, and from the representation of its publication venue. Publication venues in turn were represented using their name and aggregated metadata from a sampling of their publications. The representation of concepts in MAG differed depending on whether the concept belonged to the top two levels of the hierarchy (which were manually constructed before concept tagging) or not. Concepts on the top two levels of the hierarchy were manually mapped to publication venues, and the representations of those venues (which, recall, included text from publication metadata) were incorporated into the representations of the concepts themselves. Since it was infeasible to do this for all concepts, the representations for all other concepts seem to have been only based on text from the first paragraph of their Wikipedia article. Thus, for some concepts, their representing

text overlapped to an extent with the text used for representing the publications themselves, while for others, these representations were disjoint.

Entity	**Own Metadata**	**Network Metadata**	**Non-corpus data**
Microsoft Academic Graph			
Publication	title, abstract, other (unspecified)	metadata from references, citing works, and publication venue; sampled metadata from other publications under the same venue	N/A
Venue	name	sampled metadata from publications	N/A
L0, L1 Concept	label (Wikipedia entity name)	sampled metadata from publications assigned to concept	first paragraph of Wikipedia article
L2-L5 Concept		N/A	
OpenAlex			
Publication	title, abstract, document type	publication venue name	N/A
Concept	label (from MAG)	N/A	N/A

Table 6.1: Summary of entity representations in MAG and OpenAlex classification models.

A potential benefit of including network metadata in a publication's representation, rather than only metadata from the publication itself, is that it draws in more contextual information for that publication. Several studies (e.g. Klavans and Boyack, 2017; Waltman and van Eck, 2012) have shown that using citation relationships

improves clustering accuracy, but indirectly using such information, by concatenating metadata from references to the publication's representation, may not necessarily be as effective. Zhang et al. (2023) investigated the degree to which different types of metadata have an impact on the accuracy of models for subject classification, and found that metadata indicating publication venue was particularly helpful for both broad and specific subject tagging, but that metadata from a publication's references seemed to decrease the performance of the model.

The representation of publications in the OpenAlex model did not incorporate data from references or from citing works, but was based on only the publication title, document type, abstract, and journal title. Since concepts were treated only as labels to be targeted by the classifier, they were not given as rich a representation as they were in MAG.

For the MAG model, four types of feature vectors were extracted from the representations, namely bag-of-words, bag-of-entities, word embeddings, and entity embeddings (Shen et al., 2018). For OpenAlex, only bag-of-words was extracted from the paper metadata (OpenAlex, n.d.). The difference between these choices is that bag-of-words only signals the words that occur in a text and therefore misses any semantics that might arise from the order of the words, while embeddings capture much more semantic information about a word and its relation to the rest of the vocabulary (Rudkowsky et al., 2013). Therefore the model trained on only bag-of-words would be expected to have less contextual information available than one which combined bag-of-words with embeddings.

However, embeddings themselves must be trained, which raises the question of

possible bias resulting from the choice of training data used for an embedding. Shen et al. (2018) state that the word embeddings used in MAG were pre-trained with "13B words based on 130M titles and 80M abstracts from English scientific publications" (p. 4), and that this model performed better than ones using word embeddings trained from general English text or from Wikipedia. However, they do not specify the breakdown of the model's accuracy within different disciplines. Using only scientific publications to train word embeddings could conceivably lead to poor performance on non-scientific works, as well as skewed semantic interpretations within the model, such as a bias towards matching ambiguous/polysemic expressions with scientific concepts. We can observe this latter effect in polysemic concepts within OpenAlex, which often indicate only the scientific meaning of the concept in their descriptions, but are applied in non-scientific contexts as well. One such example is the concept *Fugue (hash function)*, which is applied to not only mathematic and cryptographic publications in the OpenAlex catalogue, but also music and cognitive science publications.

6.3.2 MAG and OpenAlex classifier architectures

Once feature vectors were extracted from the MAG representations, they were concatenated into a single vector representation for each entity, and the cosine similarities were computed between concept and publication vectors to obtain a confidence score measuring a concept's relevance to a given publication (Shen et al., 2018). Those concept-publication pairs whose scores exceeded a set threshold (unspecified in the literature) were produced as the output of the MAG concept tagger.

The OpenAlex classifier, by contrast, was built using a fully connected deep learning neural network to learn concept-publication pairs based on the historical MAG output (OpenAlex, n.d.). The use of fully connected layers in this type of model may allow a deep-learning network to learn complex patterns with higher accuracy, because no a priori constraints are placed on its learning; such models are also well suited to sparse and unstructured input data, like the bag-of-words representations used by OpenAlex (Kocsis et al., 2024). Whereas MAG constructed semantically rich entity representations to guide the similarity computations in its classifier, the deep-learning approach used by OpenAlex put more of the onus on the model itself to discover semantics and patterns among their simpler input data. While this approach responded to the relatively sparser data available to OpenAlex, it could also mean that the model behaviour would be more difficult to reason about (such as to identify potential causes of bias or errors) and more difficult to adjust or correct.

6.3.3 Usage of concept hierarchy in indexing

While in MAG, the construction of the concept hierarchy *followed from* the results of concept tagging and was therefore not available for use during indexing, OpenAlex used the pre-existing MAG hierarchy to supplement the output of its concept tagger. Traditionally, hierarchical structures in classification systems enable the selection of the most specific, yet broadest, concepts applicable to a resource (Library of Congress, 2016). The Library of Congress (2016, pp. 2-3) provides the following guidelines when using hierarchical classifications for indexing:

4. Specificity. [. . .] Follow the hierarchical reference structure built into the subject authority file (cf. H 370) to find as close a match as possible between the topic of the work and the headings that exist to express that topic in the Library of Congress subject heading system. [. . .]

5. Depth of indexing. A given heading, depending upon its place in a hierarchy, may subsume several subtopics that are also represented by headings in the subject authority file. Assign to a work only the headings that most closely correspond to the overall coverage of the work. Do not assign headings that represent the subtopics normally considered to be included in an assigned heading's coverage. *Example:*

Title: Beginning gymnastics.

```
650 #0 $a Gymnastics.
```

[Do not assign separate headings for parallel bars, balance beam, horizontal beam, vaulting horse, tumbling, etc., instead of, or in addition to, **Gymnastics**.]

OpenAlex's usage of its concept hierarchy in indexing follows the opposite of the approach described above: starting from precise (lower-level) concepts assigned to a publication by the concept tagger, it finds all paths from those concepts up to the L0 concepts, chooses the path along which the sum of the confidence scores for the concepts is the highest, and adds all the concepts on this path, including the L0 ancestor, to the set of concepts for the publication (OpenAlex, 2023). The contradiction between this approach and the Library of Congress guidelines might be justified by the fact, as discussed in section 5.2.1, that the OpenAlex concept hierarchy is not structured as a traditional (poly-)hierarchy. The parent-child structure of the hierarchy

often encapsulates more general associative relationships, not (strictly) hierarchical ones, so that following the structure from parent to child concepts does not always entail going from a general concept to a specific sub-concept. Thus, adding all the concepts along a path, instead of just the most specific one, might not necessarily create the redundancy that the Library of Congress guidelines attempt to mitigate.

However, there are other issues with this use of the concept hierarchy for indexing. According to the OpenAlex documentation, this step was added to the indexing process to ensure that at least one path from the concepts assigned to a publication would lead to the top level of the concept hierarchy (OpenAlex, 2023). Not only does this step risk adding erroneous concepts to a publication if the original concept assigned by the classifier is inaccurate, or if there are errors or ambiguity along the given part of the concept hierarchy, but this treatment of the hierarchy also seems to conflate the act of browsing the hierarchy itself with the purpose of indexing. We discussed in section 5.3.2 the lack of affordances within the OpenAlex API and catalogue for easily navigating the concept hierarchy, and encoding some of those navigational paths into the concept assignment itself might be an attempt to make the hierarchical structure more visible to users in a different way.

6.4 Hierarchy Construction

6.4.1 Relative weighted coverage

The MAG concept hierarchy was constructed after concept tagging, using an approach that Shen et al. (2018) refer to as a "subsumption-based model" (p. 5). In

this approach, the **confidence score** for a concept X and publication p, which we denote $w_X(p)$, is summed over a given set of publications, P, to get the **weighted coverage**, $W_X(P)$, of X over P. For example, in the set of publications given in table 6.2, the weighted coverage for *Cube (algebra)* would be 0.78+0.76+0.76+0.75 = 3.05, and for *Combinatorics* would be $0.64 + 0.52 = 1.16$.

Publication (p)	**Concept** (X)	**Confidence Score** ($w_X(p)$)
An Application of Dumont's Statistic	Combinatorics	0.64
Pseudospectra of elements of reduced Banach algebras II	Combinatorics	0.52
Of Cubes, and of the Extraction of Cube Roots	Cube (algebra)	0.78
How to create your own desktop library cube	Cube (algebra)	0.76
Nonpositively curved cube complexes	Cube (algebra)	0.76
The cube of the sum formula	Cube (algebra)	0.75

Table 6.2: Example weights of OpenAlex publication-concept pairs.

The **relative weighted coverage** with respect to a concept X and a set of publications P, which we will denote $R(X, P)$ is then defined by Shen et al. (2018) as the weighted coverage for X on P divided by the sum of all weights for X across the corpus. Then, concept X is determined to be a child of concept Y when the following inequality holds

$$R(X, O) - R(Y, O) = \frac{W_X(O)}{\sum_{p \in C} w_X(p)} - \frac{W_Y(O)}{\sum_{p \in C} w_Y(p)} > \epsilon$$

where O is the set of publications tagged with both X and Y, C is the corpus of all publications, and e is a given positive threshold (which the authors note is empirically usually between 0.2 and 0.5) (Shen et al., 2018). Intuitively, we can understand this calculation as saying that, if X appears to be more strongly associated to a paper than Y is to that same paper, then X is probably a more specific concept than Y; and if this holds over a large enough set of publications relative to the whole corpus, then X should be considered a child of Y in the hierarchy.

Issues with relative weighted coverage

While there are some issues with using the relative weighted coverage calculation as the basis for determining hierarchical relationships, it is possibly most reliable when the set of publications tagged with the candidate child concept, X, is a subset (or near subset) of O, the set of all publications tagged with both X and Y, *and* the set of publications tagged with Y is much larger than O. That is to say, X appears (almost) exclusively in the presence of Y, but not the other way. In this case, the term $R(X,O) = \frac{W_X(O)}{\sum_{p \in C} w_X(p)}$ will be close to 1 but $R(Y,O) = \frac{W_Y(O)}{\sum_{p \in C} w_Y(p)}$ will not be, so the inequality will hold for a well-chosen value of ϵ.

On the other hand, if the confidence scores for the candidate parent concept, Y, are low for all or most publications that are tagged with both concepts, then the term $R(Y,O)$ in the above inequality will be close to 0, and the overall inequality will be mainly determined by the confidence scores for X, irrespective of Y. This could lead to concepts being put in hierarchical relation with each other even when one is very weakly associated to the same publications as the other. Furthermore,

there is no guarantee that, in any case, X and Y have a meaningful relationship to each other simply because they both occur in a number of the same publications, nor does this calculation take into account any other semantic information about X or Y that could help guarantee such a relationship.

6.5 Conclusion

In this chapter, we looked at each of the technical steps used to construct the MAG and OpenAlex concept hierarchies, and examined the design of these steps with respect to standard classification practices. We found that none of the processes involved were aligned, whether deliberately or accidentally, with information theoretical principles, and that this misalignment might contribute to significant problems in the classifications.

Chapter 7

Discussion

In the preceding chapters, we have studied the OpenAlex concepts top-down, analyzing their structure and content, as well as bottom-up, reviewing the technical design and implementation of the associated machine-learning models. While at first glance, the poly-hierarchy of the OpenAlex concepts appears to contradict classification standards, its flexibility also allows room for representations that are difficult to do with traditional classificatory structures. We found that relaxing strict definitions of hierarchical relationships in favour of interpreting those relationships contextually allowed the structure of the classification itself to better capture different facets of interdisciplinary concepts. For a platform aimed at supporting research evaluation, this ability of the OpenAlex concepts to capture multidisciplinarity could be an advantage for identifying emerging research trends and shifting disciplinary boundaries. However, that lack of strictness was also accompanied by ambiguity and errors that reduced the usefulness of the classification, especially when the hierarchical relation-

ships connecting specific concepts to broader disciplines were shown to be unreliable.

Through both our top-down and bottom-up analyses, we observed that the OpenAlex concepts do not conform to the requirements of a hierarchical classification system in structure and in design. The technical choices made in their construction do not line up with various classification standards, and their structure is more similar to a tag list with associative relationships than a hierarchical classification. The relationships within OpenAlex seem to work best between only two levels at a time and are not usable for the same navigational and indexing purposes as traditional classification hierarchies.

We argue that these issues demonstrate that the use of automatic techniques cannot compensate for information theoretical principles and standard practices. In chapter 5, we discussed the hierarchy construction process, which involved using a subsumption-based model (Shen et al., 2018) to determine if pairs of concepts should be placed in hierarchical relationship with each other. To summarize, if concept A is applied to a subset of the same publications as concept B, then (given that the confidence scores of these concepts are high enough) A is considered to be a more specific topic than B, and B is therefore taken as a parent concept for A in the hierarchy. However, this calculation measures only the strength of the co-occurrence between A and B in the collection, and any further inferences about how these concepts are related (such as that one is a sub-topic of the other, etc.) are not justified by the subsumption principle alone.

It makes sense then that the definitions of hierarchical relationships given in NISO (2010) are not satisfied by the OpenAlex hierarchy, as there is no process within the

hierarchy construction algorithm that takes into account information that would be relevant to those definitions. This fact also explains the lack of transitivity within the OpenAlex hierarchical relationships, as the co-occurrence of words in natural language is, by itself, generally not a transitive relation. If we characterize the hierarchical relationships in OpenAlex as providing a context for concepts rather than as restricting their definition, then it does not necessarily follow that such contexts should carry forward to all descendants of a concept, as the immediate relevance of the context diminishes as its distance from the concept increases. To revisit an example from section 5.2.2, while *Law* may be a reasonable context for topics in *Cloud computing* (such as compliance laws for cloud platforms, for example), and *Political science* may be for topics in *Law*, *Political science* itself does not necessarily serve as a very useful context for subtopics of cloud computing. Interestingly, Shen et al. (2018) note the lack of transitivity in the MAG hierarchy as a "limitation of subsumption-based models" (p. 5), but do not provide further interpretation of this structural issue.

If transitivity does not follow from the relationships created by the MAG subsumption algorithm, what do the paths through the hierarchy signify? Technically, they encode pairwise co-occurrences, meaning that a chain of relationships such as *Medicine* $\longrightarrow$ *Anesthesia* $\longrightarrow$ *Airway* represents the fact that *Medicine* and *Anesthesia* co-occur in a significant subset of publications, as do *Anesthesia* and *Airway*, in a possibly overlapping, or possibly disjoint, subset. As discussed in section 6.4.1, the subsumption formula is most likely to correspond to a meaningful relationship between two concepts when the child concept occurs almost entirely within the con-

text of the parent concept; this roughly corresponds to the observation in section 5.2.2 that high convergence among the different paths leading to a concept sometimes signaled fewer errors or ambiguity issues for that concept. We also observed that having unrelated ancestors was often correlated with errors or ambiguity, and we can see from the hierarchy construction that this would result from the same term co-occurring with different concepts in different sets of publications, with the algorithm unable to distinguish among each of its different usages.

In section 5.2.2, we identified instances of polysemic ambiguity within terms, as well as semantic ambiguity among terms that resulted in erroneous relationships within the hierarchical structure. One possible cause of ambiguity where the same concept (such as *Subtext*) names multiple different things could have been that the classification itself was missing terms that would denote the other meanings of the concept (for example, "Subtext (literature)", "Subtext (programming language)", etc.), and hence the classifier could not target different labels in different contexts. This could have resulted from the concept discovery process used in MAG to generate the initial set of concepts; as that process relied on Wikipedia entities (Shen et al., 2018), any concepts not differentiated in Wikipedia article titles would likely also not have been differentiated in MAG, and hence, in OpenAlex. Even though the MAG model produced an enormous number of specific concepts, due to its reliance on Wikipedia entities, it may still not have been able to provide sufficient coverage to avoid ambiguity within the classification. Another possibility is that retaining only the top 9% of concepts from the MAG classification in OpenAlex (OpenAlex, n.d.) could have dropped some (lesser-used) concepts that did distinguish different

uses of the same term in MAG, leading to the conflation of those different usages in the OpenAlex classification model. Relying only on the statistical popularity of a concept to determine its importance could have exacerbated some of the issues with errors and ambiguity for more marginal concepts within OpenAlex.

Given the issues discussed at length above, the subsumption formula is arguably a poorly justified choice as the basis of the concept hierarchy construction, yet its use nonetheless points us towards some of the power structures that appear to have goverened the development of MAG and OpenAlex. Recall the matrix of domination model that we discussed in section 2.1, which identifies four domains in which power operates: the structural, the disciplinary, the hegemonic, and the interpersonal (Collins, 2008, as cited in D'Ignazio & Klein, 2020). If we consider this model with respect to machine-learning, some sources of structural power might be the organizations which control access to the computing resources needed to train and deploy machine-learning models, as well as those which control funding for this expensive work. Power in the hegemonic domain might be described in terms of cultural attitudes towards the perceived value of machine learning, such as the (flawed) idea that machine-derived solutions are free of bias. As we have seen in section 4.2.3, the prevailing view of bias within the information studies field is that it is an inevitable part of any classification system and should be deliberately accounted for in a system's design (Mai, 2011; Feinberg, 2007; Olson, 2001), and one consequence of this argument is that, if accepted, it inherently limits the power of any technical solution. Power in the disciplinary domain might manifest in the form of policies that determine how machine-learning projects are funded, which may increase the influence

of cost and resource constraints on projects, making solutions that cost more time or effort less desirable. This in turn may reinforce both the power of organizations providing costly computational resources, as well as the cultural beliefs about the value of machine-learning over other types of solutions. Finally, the effect of these manifestations of power in the interpersonal domain might then be a devaluing of the labour of subject matter experts who are replaced by machine-learning models, and a feeling of futility towards the idea of trying to change models that do not work properly or meet needs well.

As no one is empowered to advocate for expertise developed outside of technical frameworks, such as the guidelines for classification system design given in NISO (2010), or reflections on the purpose of classification in, e.g., Beghtol (2003), less effective solutions such as the subsumption formula are adopted despite their shortcomings. These solutions then become entrenched and their outputs replicated, as we have seen in the case of OpenAlex's replication of the MAG hierarchy, with the end result that the power structures which worked against the possibility of developing a better solution in the first place are reinforced.

While some of the model behaviours that led to erroneous outcomes could potentially be changed by using different training data, choosing different model architectures, etc., there remains a fundamental issue that classification work cannot be reduced to data processing alone, but requires deliberate consideration of the naming and selection of terms and of the structure of relationships among those terms, all of which should be driven by the specific purpose of the classification and the goals of its users. The techniques used by MAG and OpenAlex may have been successful at

finding statistical pattens within collections (such as co-occurrence relationships) but even the most advanced machine-learning models still cannot be expected to reason about these patterns meaningfully (Mirzadeh et al., 2024; Zhang et al., 2022; Xie et al., 2024), and some human judgment is needed to intentionally shape "puzzling" results, as Rafols and Leydesdorff (2009, p. 1833) have put it, into useful, coherent classification systems. For example, some research evaluations may actually benefit from information about interdisciplinarity to different degrees, and neither should polyhierarchical structure be denounced outright nor should a single implementation be algorithmically imposed; rather, determining how and whether a classification represents multiple disciplines for works should be based on an assessment of the utility of the patterns and structures uncovered through automatic techniques for particular purposes.

As Mai (2011) has argued, classification goes beyond simply describing the contents of resources, and requires, in addition to subject analysis, a context-dependent understanding of the needs of users. Such work, therefore, necessarily requires making conscious decisions that may exhibit bias in the sense of favouring some needs over others, and, as Feinberg (2007) proposes, should acknowledge and explain those decisions rather than naively trying to eliminate (appearance of) bias by replacing human labour with machine labour. Automatic techniques can help explore new avenues in classification if we treat their outputs as one of many inputs to the process of developing a classification. Visualization tools such as those used in this project could be the basis for interfaces that would allow classification designers to probe large and potentially complicated datasets produced by machine-learning models.

Specializations of classifications for different purposes could be accomplished from the same model output, that is to say, different *explanans* could be furnished for the same, fundamental *explanandum*. Honing machine output through the subjective and deliberate intervention of subject-matter experts and information professionals would not destroy the validity of the resulting classification, but could, in fact, greatly improve its usability and effectiveness.

Chapter 8

Conclusion

We have seen that the processes involved in creating the machine-derived OpenAlex concept hierarchy and its predecessor, the MAG fields of study, diverged from established classification theory and practice, with the resulting concept hierarchy having limited utility as a hierarchical classification. However, the unusual structure of the concept hierarchy offers a genuinely interesting opportunity to question and reinterpret standard practices, such as how the hierarchical relationships in OpenAlex provide a possible response to the issue of universal perspective within (strict) hierarchical structures raised by Olson (2001). The less formal approach to relationships within MAG and OpenAlex might also have some advantages for a scholarly research classification system, where the boundaries of disciplines are not strictly defined, but often continuously evolving. Finally, the automatic processes used by MAG and OpenAlex are more effective than manual approaches with respect to operating at the scale of a classification intended to cover the enormous volume of scholarly re-

search (Wang et al., 2019).

These advantages of machine-derived classifications, however, should not completely obscure the problems that can arise when insufficient care is taken towards their design. We found numerous ambiguity issues and erroneous or misleading relationships in the OpenAlex concept hierarchy, which stemmed from the technical design decisions made by MAG and OpenAlex. Our investigations have led us to believe that, while machine-based techniques can be a useful tool for classification design and implementation, that work should still be guided by human judgement and information-theoretical foundations in order to achieve the best possible results.

This project contributes an in-depth case study of a machine-derived classification, conducted through close readings of technical papers, classification guidelines and information studies literature; statistical and heuristic explorations of OpenAlex data; and inspection of OpenAlex code where available. We generated several datasets from the OpenAlex concepts, but due to practical limitations, were only able to explore random samples of these datasets for our analyses. Future work could study these datasets more wholistically and using participants from different backgrounds, including information professionals, students, and scholarly communications researchers, in order to compare assessments of OpenAlex concept quality across different user groups. Further studies could also apply a similar methodology as presented in this work towards studying the new OpenAlex topics classification system.

References

Arcadia. (n.d.). Homepage. Retrieved from https://arcadiafund.org.uk/

Archambault, É., Beauchesne, O. H., & Caruso, J. (2013). Towards a multilingual, comprehensive and open scientific journal ontology.. Retrieved from https://api.semanticscholar.org/CorpusID:85557623

Barité, M. (2019). Towards a general conception of warrants: First notes. *KNOWLEDGE ORGANIZATION*. Retrieved from https://api.semanticscholar.org/CorpusID:213566439

Barité, M., & Rauch, M. (2020). Cultural warrant: Old and new sights from knowledge organization. *Knowledge Organization at the Interface*. Retrieved from https://api.semanticscholar.org/CorpusID:229658553

Barité, M. (2018). Literary Warrant. *Knowledge organization*, *45*(6), 517–536. doi: 10.5771/0943-7444-2018-6-517

Beghtol, C. (1986). Semantic validity: concepts of warrant in bibliographic classification systems. *Library Resources & Technical Services*, *30*, 109-125. Retrieved from https://api.semanticscholar.org/CorpusID:63192530

Beghtol, C. (2003). Classification for information retrieval and classification for

knowledge discovery: Relationships between professional and naive classifications. *Knowledge Organization*, *30*, 64-73. Retrieved from https://api.semanticscholar.org/CorpusID:59632099

Boyack, K., Small, H., & Klavans, R. (2013, September). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*, 1759-1767. doi: 10.1002/asi.22896

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., ... Börner, K. (2011, March). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, *6*(3), 1-11. Retrieved from https://doi.org/10.1371/journal.pone.0018029 doi: 10.1371/journal.pone.0018029

Bu, Y., Li, M., Gu, W., & Huang, W.-b. (2021). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology*, *72*(5), 523-539. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24433 doi: https://doi.org/10.1002/asi.24433

Bullard, J. (2017). Warrant as a means to study classification system design. *Journal of Documentation*, *73*(1), 75-90. Retrieved from https://doi.org/10.1108/JD-06-2016-0074

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2006, May). The SALSA corpus: a German corpus resource for lexical semantics. In N. Calzolari et al. (Eds.), *Proceedings of the fifth international conference on language resources and evaluation (LREC'06).* Genoa, Italy: European

Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/339_pdf.pdf

Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010, September). Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, *100*(4), 1358–98. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.100.4.1358 doi: 10.1257/aer.100.4.1358

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus.*

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism.* The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/11805.001.0001 doi: 10.7551/mitpress/11805.001.0001

Drabinski, E. (2013). Queering the catalog: Queer theory and the politics of correction. *The Library Quarterly: Information, Community, Policy*, *83*(2), 94–111. Retrieved 2024-09-24, from http://www.jstor.org/stable/10.1086/669547

Feinberg, M. (2007). Hidden bias to responsible bias: an approach to information systems based on Haraway's situated knowledges. *Inf. Res.*, *12*. Retrieved from https://api.semanticscholar.org/CorpusID:41326675

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021, November). Datasheets for datasets. *Commun. ACM*, *64*(12), 86–92. Retrieved from https://doi.org/10.1145/3458723 doi: 10

.1145/3458723

Glänzel, W., & Schubert, A. (2003, March 01). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357-367. Retrieved from https://doi.org/10.1023/A:1022378804087 doi: 10.1023/A:1022378804087

Golub, K. (2021, December). Automated subject indexing: An overview. *Cataloging Classification Quarterly*, *59*, 1-18. doi: 10.1080/01639374.2021.2012311

Herrmannova, D., & Knoth, P. (2016). An Analysis of the Microsoft Academic Graph. *D-Lib Magazine*, *22*(September/October 2016). Retrieved from http://www.dlib.org/dlib/september16/herrmannova/09herrmannova.html doi: 10.1045/september2016-herrmannov

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches - traditional as well as innovative. *Journal of Documentation*, *58*(4), 422-462. Retrieved from https://doi.org/10.1108/00220410210431136 doi: 10.1108/00220410210431136

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, *46*(6), 400-425. doi: https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Huang, Y., Lu, W., Liu, J., Cheng, Q., & Bu, Y. (2022). Towards transdisciplinary impact of scientific publications: A longitudinal, comprehensive, and large-scale analysis on Microsoft Academic Graph. *Information Processing Management*, *59*(2), 102859. Retrieved from https://www.sciencedirect.com/science/

article/pii/S0306457321003307 doi: https://doi.org/10.1016/j.ipm.2021.102859

Hug, S. E., & Brändle, M. P. (2017, December 01). The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551-1571. Retrieved from https://doi.org/10.1007/s11192-017-2535-3 doi: 10.1007/s11192-017-2535-3

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017, April 01). Citation analysis with Microsoft Academic. *Scientometrics*, *111*(1), 371-378. Retrieved from https://doi.org/10.1007/s11192-017-2247-8 doi: 10.1007/s11192-017-2247-8

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984-998. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23734 doi: https://doi.org/10.1002/asi.23734

Kocsis, P., Súkeník, P., Brasó, G., Nießner, M., Leal-Taixé, L., & Elezi, I. (2024). The unreasonable effectiveness of fully-connected layers for low-data regimes. In *Proceedings of the 36th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Lee, W. (2017, 09). Conflicts of semantic warrants in cataloging practices. *NASKO*, *6*, 231-238. doi: 10.7152/nasko.v6i1.15242

Li, W., Liu, X., Ajmal Khan, M., & Yamaguchi, S. (2005, June 01). The effect of plant growth regulators, nitric oxide, nitrate, nitrite and light on the germi-

nation of dimorphic seeds of suaeda salsa under saline conditions. *Journal of Plant Research*, *118*(3), 207-214. Retrieved from https://doi.org/10.1007/s10265-005-0212-8 doi: 10.1007/s10265-005-0212-8

Library of Congress. (2016). Assigning and constructing subject headings [H 180]. Retrieved from https://www.loc.gov/aba/publications/FreeSHM/H0180.pdf

Liu, J., Chen, H., Liu, Z., Bu, Y., & Gu, W. (2022). Non-linearity between referencing behavior and citation impact: A large-scale, discipline-level analysis. *Journal of Informetrics*, *16*(3), 101318. Retrieved from https://www.sciencedirect.com/science/article/pii/S1751157722000700 doi: https://doi.org/10.1016/j.joi.2022.101318

Mai, J.-E. (2005). Analysis in indexing: document and domain centered approaches. *Information Processing & Management*, *41*(3), 599-611. Retrieved from https://www.sciencedirect.com/science/article/pii/S030645730300116X doi: https://doi.org/10.1016/j.ipm.2003.12.004

Mai, J.-E. (2011). The modernity of classification. *Journal of documentation*, *67*(4), 710–730. doi: 10.1108/00220411111145061

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024, October). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models.
doi: 10.48550/arXiv.2410.05229

National Information Standards Organization (NISO). (2010). *Guidelines for the construction, format, and management of monolingual controlled vocabular-*

ies (ANSI/NISO Z39.19-2005 (R2010)). Retrieved from https://www.niso.org/

Olson, H. A. (2001). The power to name: Representation in library catalogs. *Signs: Journal of Women in Culture and Society*, *26*(3), 639–668. doi: 10.1086/495624

Olson, H. A. (2004). *Bacon, warrant, and classification.* dLIST. Retrieved from http://hdl.handle.net/10150/105397

OpenAlex. (n.d.-a).

OpenAlex. (n.d.-b). *Openalex: End-to-end process for concept tagging.* Google Docs. Retrieved from https://docs.google.com/document/d/1q3jBlEexskCZaSafFDMEEY3naTeyd7GS/edit?usp=sharing&ouid=112616748913247881031&rtpof=true&sd=true

OpenAlex. (2023). openalex-concept-tagging. Retrieved from https://github.com/ourresearch/openalex-concept-tagging

OurResearch. (n.d.). Our projects. Retrieved from https://ourresearch.org/projects

Pena-Rocha, M., Gomez-Crisostomo, M., Guerrero-Bote, V., & Moya-Anegon, F. (2024, October). *Bibliometrics effects of a new item-by-item classification system based on reference reclassification.* doi: 10.48550/arXiv.2410.23792

Portenoy, J. (2023, December 19). *Upcoming changes to concepts in openalex.* Google Groups. Retrieved from https://groups.google.com/g/openalex-users/c/2yE1jie_D3s/m/c3j9UYiLBgAJ

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823-1835. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21086` doi: https://doi.org/10.1002/asi.21086

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, , & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140–157. Retrieved from `https://doi.org/10.1080/19312458.2018.1455817` doi: 10.1080/19312458.2018.1455817

Scheidsteger, T., & Haunschild, R. (2023, March). Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *El Profesional de la información*, *32*. doi: 10.3145/epi.2023.mar.09

Shen, Z., Ma, H., & Wang, K. (2018, July). A Web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations* (pp. 87–92). Melbourne, Australia: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P18-4015` doi: 10.18653/v1/P18-4015

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, *60*(3), 571-580. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/`

asi.20994 doi: https://doi.org/10.1002/asi.20994

Sutter, M., Kocher, M. G., Glätzle-Rützler, D., & Trautmann, S. T. (2013, February). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, *103*(1), 510–31. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.103.1.510 doi: 10.1257/aer.103.1.510

Tan, J., Pan, X., Kavulya, S., Gandhi, R., & Narasimhan, P. (2008, January). Salsa: Analyzing logs as state machines..

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378-2392. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22748 doi: https://doi.org/10.1002/asi.22748

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020, February). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. Retrieved 2024-07-14, from https://doi.org/10.1162/qss_a_00021 doi: 10.1162/qss_a_00021

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., ... Rogahn, R. (2019). A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, *2*. Retrieved from https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2019.00045 doi: 10.3389/fdata.2019.00045

Welles, B. F. (2014). On minorities and outliers: The case for making Big Data small.

Big data & society, *1*(1), 205395171454061. doi: 10.1177/2053951714540613

Xie, W., Ma, S., Wang, Z., Wang, E., Wang, B., & Su, J. (2024). Do large language models truly grasp mathematics? an empirical exploration. *ArXiv*, *abs/2410.14979*. Retrieved from https://api.semanticscholar.org/CorpusID:273502126

Xu, H., Liu, M., Bu, Y., Sun, S., Zhang, Y., Zhang, C., ... Ding, Y. (2024). The impact of heterogeneous shared leadership in scientific teams. *Information Processing Management*, *61*(1), 103542. Retrieved from https://www.sciencedirect.com/science/article/pii/S0306457323002790 doi: https://doi.org/10.1016/j.ipm.2023.103542

Zafar, H. (2024a). *Datasets for Masters Thesis, Huma Zafar.* Borealis. Retrieved from https://doi.org/10.5683/SP3/29QGMP doi: 10.5683/SP3/29QGMP

Zafar, H. (2024b, December). *Huma Zafar Masters Thesis Scripts.* Retrieved from https://github.com/hzafar/masters-thesis-scripts

Zafar, L., Masood, N., & Ayaz, S. (2023, April 01). Impact of field of study (fos) on authors' citation trend. *Scientometrics*, *128*(4), 2557-2576. Retrieved from https://doi.org/10.1007/s11192-023-04660-2 doi: 10.1007/s11192-023-04660-2

Zhang, H., Li, L. H., Meng, T., Chang, K.-W., & den Broeck, G. V. (2022). On the paradox of learning to reason from data. In *International joint conference on artificial intelligence.* Retrieved from https://api.semanticscholar.org/CorpusID:248986434

Zhang, Y., Jin, B., Zhu, Q., Meng, Y., & Han, J. (2023). The effect of metadata

on scientific literature tagging: A cross-field cross-model study. In *Proceedings of the acm web conference 2023* (p. 1626–1637). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3543507.3583354 doi: 10.1145/3543507.3583354