

Senka Stankovic, Isabella Sun, Kayla Wemp
University of Ottawa, Ottawa, Ontario, Canada

AN EVALUATION OF OPENALEX LEVEL 0 - ART

(Lightning Talk)

Abstract or Résumé:

OpenAlex is an online repository using an automated classification system. While existing studies highlight the overall usability of OpenAlex, this study analyzes the limitations and effectiveness of the Level 0 concept “Art.” Our findings suggest that the system has significant hierarchical and semantic shortcomings identified through a mixed-method research approach. This approach utilized qualitative and quantitative research methods to provide a holistic evaluation of the system, gauging its effectiveness in indexing papers. Moreover, we provide key insights into how improvements can be made to render OpenAlex an effective and usable classification system.

1. Introduction

OpenAlex is an online repository with over 243 million works, of which 43 million are open access. It uses an automated tagging system to classify these works. The system’s structure uses a five-level system to categorise concepts. L0 or the “root level” has nineteen concepts.

The findings highlight the necessity for significant restructuring and quality control within OpenAlex. Our recommendations for improvement include a restructuring of the concepts, the removal of unuseful index terms, and verification that definitions are present and accurate. We also recommend that OpenAlex implements clear procedures in the case of synonymy and homonymy to omit related classification errors.

2. Literature Review and Environmental Scan

OpenAlex uses the dataset from Microsoft Academic Graph (MAG), which was retired at the end of 2021, to provide free data describing scholarly entities and how those entities are connected to each other (Overview, n.d.). There are many studies on how OpenAlex compares to Web of Science and Scopus, and it is well known for its comprehensive coverage of publications (Jiao et al., 2023). Some researchers noted inconsistencies in OpenAlex’s metadata, such as errors in the open access status of works (Jahn et al., 2023), while others successfully used OpenAlex in conjunction with other databases to obtain information about open access books, noting some limitations, such as the lack of accuracy when filtering records classified as “books” (Laakso, 2023). Similarly, scholars found irregularities with document type metadata (Jiao et al., 2023), while others noted gaps in a lack of institutional metadata for dataset authors (Krause & Mongeon, 2023).

While there are studies researching OpenAlex in specific use cases, there has not been a granular analysis of the indexing concepts. Our research addresses this gap through analyzing the hierarchical and semantic relationships within the Level 0 concept of “Art,” and evaluating the application of index terms to works.

3. Evaluation

This evaluation of OpenAlex as a classification system takes a multifaceted approach, analysing its concepts and concept levels (1) qualitatively, with attention to hierarchical and semantic issues; (2) quantitatively, with attention to concept distribution and relevance of concepts; and (3) by evaluating how concepts are applied for indexing of works. The methodology predominantly uses Excel to analyse the data from OpenAlex, which was retrieved in the fall of 2023, thus offering insights and recommendations specific to this time period. It is important to note that OpenAlex indexes new works daily, resulting in changing data sets. Our findings identify a range of shortcomings that ultimately negatively impact OpenAlex’s retrieval accuracy.

-Qualitative

A qualitative analysis of the OpenAlex system identifies several hierarchical and semantic challenges, largely in relation to the overall quality of the structure and inaccurate or unavailable definitions of concepts. The quality of the hierarchical structure is most affected by the scope of concepts in L0 and L1 being either too specific or too broad for their placement within the system. Moreover, there is a lack of predefined methods for discerning how the hierarchical assignment happens. The current structure of L0 and L1 concepts is shown in Figure 1, while our proposed restructuring is presented in Figure 1.1. This restructuring of the higher (L0-L2) levels could increase retrieval efficiency at the lower (L3-L5).

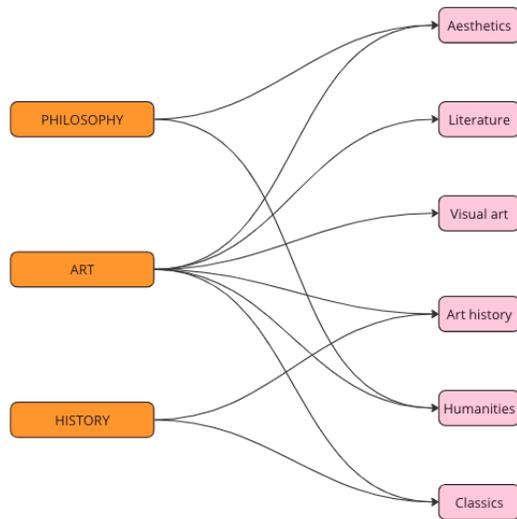


Figure 1.0 Current Structure of Concepts, L0 to L1

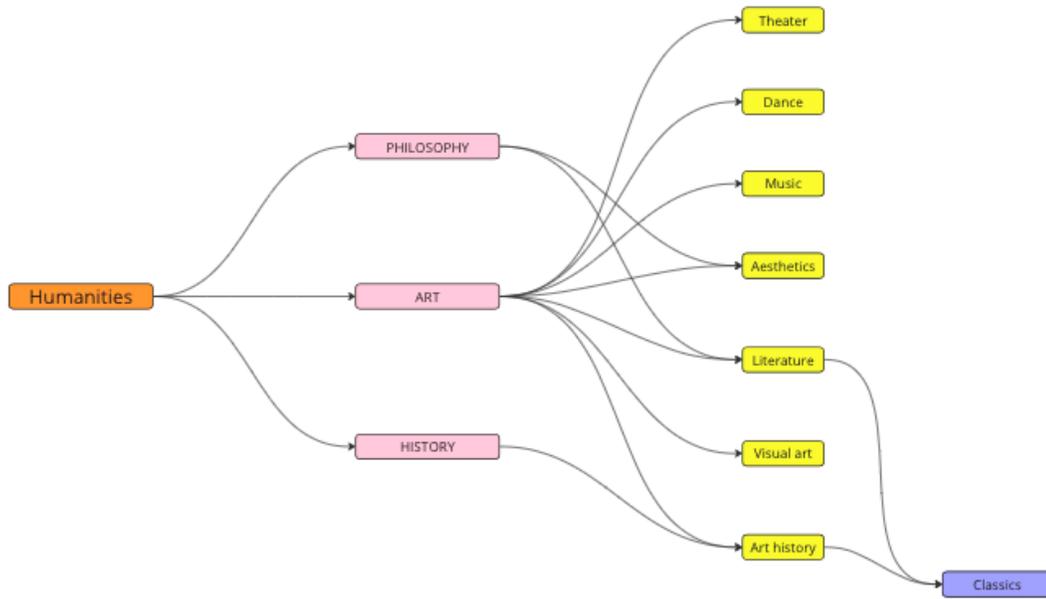


Figure 1.1 Suggested Structure of Concepts, L0 to L3

-Quantitative

The quantitative analysis of OpenAlex examines the numeric discrepancies in the works, and concept relevancy and levels. Through querying the API, we found that the majority of concepts are concentrated at L2, with little to no L3 to L5 concepts. Since there is no hierarchical structure to the concepts, we discovered that concepts with a high amount of works did not have accurate descriptors, which poses a risk of indexing errors in the classification of works. Moreover, our findings show the low relevancy rates between children and parent concepts (Figure 2.0). Therefore, the concepts available to a user searching OpenAlex would not only retrieve inaccurately indexed works, but also yield irrelevant results for their search.

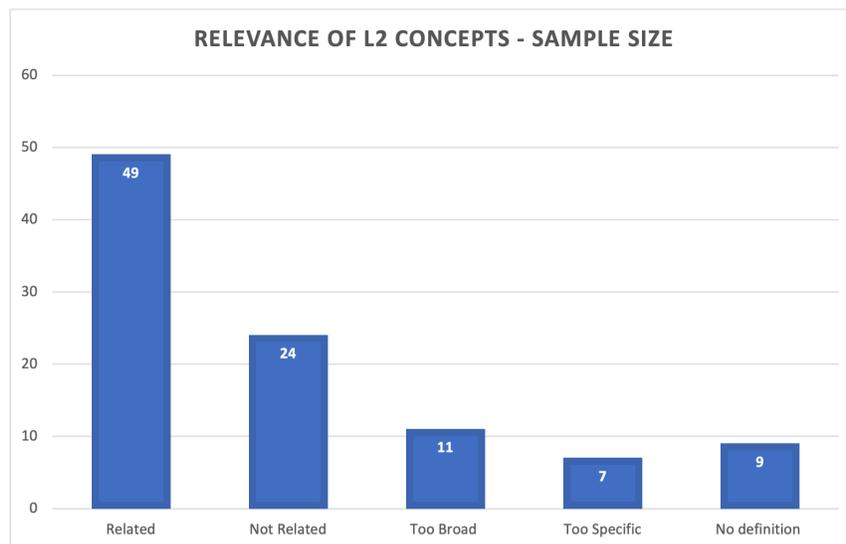


Figure 2.0: Relevance of 10% of L2 concepts randomly sampled from each L1 concept

-Evaluating Index Terms

Two methods were used to evaluate index terms. Our first method employed an analysis of the concepts assigned to one specific work (**Table 1**), where we gauged the quality and relevance of the concept. This method was applied with two sample works. Our second method used a random sample of 15 works assigned to one concept to see how many works are accurately indexed with those concepts (**Figure 3.0**). This method was applied with three sample concepts. Both of these approaches demonstrate the potential for high amounts of inaccuracies when it comes to OpenAlex’s concepts being used as index terms for works, in part due to issues with homonyms or similar concepts, lack of specificity, illogical concept hierarchies, or simply choosing irrelevant words from titles and abstracts.

Concept	Level	Relevant to work?	OpenAlex Definition
Volute	3	Y	curved funnel that increases in area as it approaches the discharge port; casing in a centrifugal pump that receives the fluid being pumped by the impeller, maintaining the velocity of the fluid through to the diffuser
Iconography	2	Y	use of symbols, themes, and subject matter in the visual arts
Frieze	2	Y	wide central section part of an entablature
Macedonian	2	N	South Slavic language mostly spoken in North Macedonia and its neighbouring countries
Bronze	2	Y	Metal alloy
Context (archaeology)	2	N	glossary for archaeological terms

Table 1:

An analysis of the concepts applied to The Derveni krater : Masterpiece of classical Greek metalwork.

<https://api.openalex.org/works/W594641071>

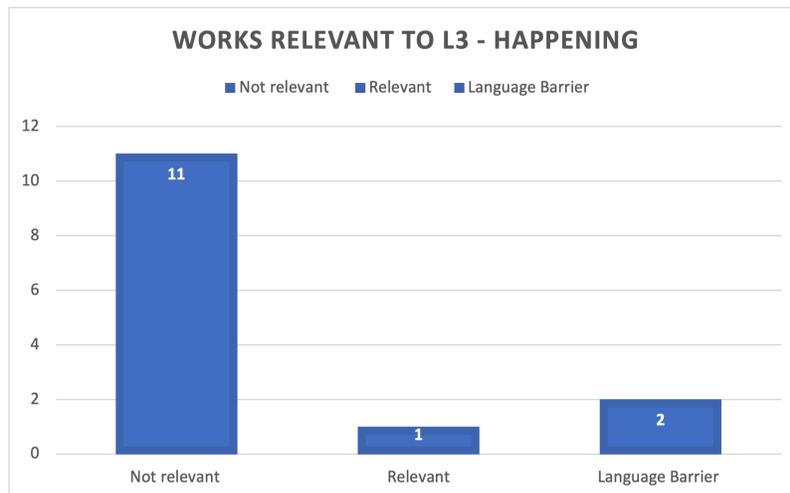


Figure 3.0 Count of relevant works indexed with *Happening-L3*

4. Conclusion

OpenAlex has an impressive ability to index a high volume of works due to its automated tagging system. While this is a compelling step forward in the realm of open access work and automated systems, there is significant progress to be made to ensure that OpenAlex is a functional tool that researchers can use to their benefit. The qualitative analysis demonstrates that hierarchical and semantic discrepancies increase classification inaccuracies, and suggests possible edits to be made to increase the effectiveness of the concept structure. The quantitative analysis demonstrates numeric discrepancies and concludes that the system leaves substantial room for error. The evaluation of indexing indicates that works are often indexed inaccurately, and do not compare in accuracy to human-indexed works. This leads to low retrieval accuracy and affects the usability of OpenAlex for research.

References

- Jahn, N., Haupka, N., Hobert, A. (2023, Nov. 7). Scholarly communication analytics: Analysing and reclassifying open access information in OpenAlex. Retrieved from https://subugoe.github.io/scholcomm_analytics/posts/oalex_oa_status/
- Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data*, 10(1), 737–737. <https://doi.org/10.1038/s41597-023-02625-x>
- Krause, G. & Mongeon, P. (2023). Measuring Data Re-Use Through Dataset Citations in OpenAlex [preprint]. *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*, Leiden, Netherlands. <https://doi.org/10.55835/6442d8d30f5efe988a0e1d67>
- Laakso, M. (2023). Open access books through open data sources: Assessing prevalence, providers, and preservation. *Journal of Documentation*, 79(7), 157–177. <https://doi.org/10.1108/jd-02-2023-0016>
- Overview. OpenAlex. (n.d.). <https://docs.openalex.org/>