It's in the syllabus: A data infrastructure for innovation studies

Keywords: Science of Science; innovation studies; scholarly data

Extended Abstract

Quantitative studies of science have a long history, dating back possibly to the 1920s and 1950s when Lotka and Shockley respectively analyzed productivity of individual scientists and research laboratories. This field has regained widespread attention recently. Traditional data sources supporting the large-scale, quantitative, cross-disciplinary inquiry of science are commercial databases like the Web of Science, Scopus, and the recently emerged Dimensions. However, the commercial nature of these databases raises numerous issues regarding their data source, coverage, access, and usage. In the last two decades, publicly available scholarly databases have been flourishing, including CiteSeerX, DBLP, PubMed, AMiner, Semantic Scholar, APS, the now-decommissioned Microsoft Academic Graph, OpenAlex, among many others. However, these databases have limited support for innovation studies. For example, they lack (1) the linkages with other data resources that capture diverse aspects of science, such as mentor-mentee relationships; (2) implementations and validations of popular measurements; and (3) flexible ways for users to explore and export desired subsets of data.

Here, we present *it's in the syllabus*, a full-stack, continuously updating data infrastructure suitable for innovation studies. It can be reached at https://thesyllabus.io/. The back end of the system is a large-scale, multi-disciplinary dataset resulted from the collection, integration, and interconnection of multiple publicly available scholarly databases, enabling a wide range of analyses. We further implement a number of widely used measurements and validate our implementations by reproducing key findings in their original proposals. Finally, we build a front-end website to facilitate explorations and harvest of the dataset (Fig. 1A). It adopts an entity-centric view so that users can look up all the relevant information regarding every instance of each type of entities. Currently, we have paper (Fig. 1B), author, journal, institution, and MeSH terms ready for view, and more types of entities will be included in the future. Additionally, the platform features a suit of APIs and a search engine with rich filtering options to provide users with a powerful way to quickly retrieve, curate, and export search results to suit their specific needs (Fig. 1C). We welcome the community's use and feedback as well as contributions of additional data and hope that *it's in the syllabus* will be a common basis used for scholarly data analysis and lower the barrier for researchers who might be less equipped with technical skills handling large data but otherwise interested in quantitative studies of innovation.

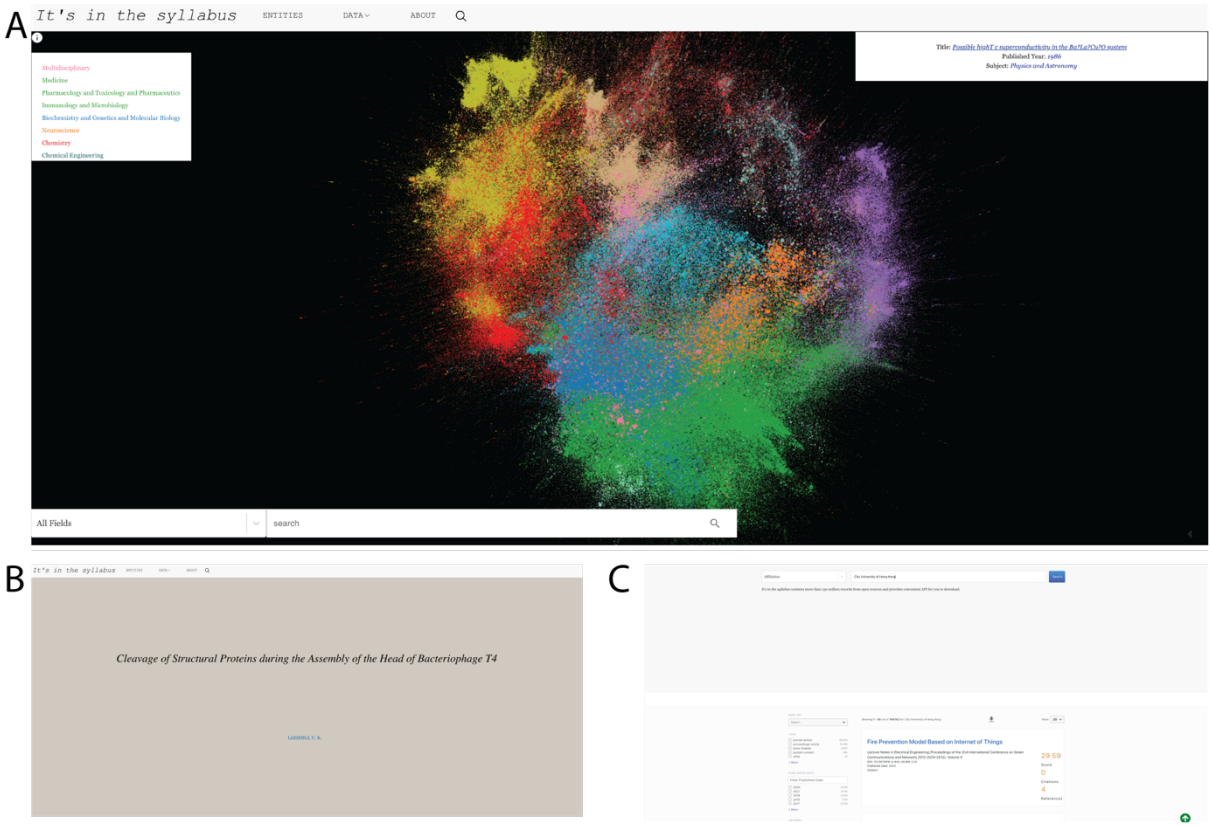

Figure 1. Selected views of *it's in the syllabus*