Mapping the Field of Data Ethics

Jonathan Reeve, Isabelle Zaugg, Tian Zheng

Abstract

From the spread of disinformation via social media, to class-biased dynamic pricing, to racial profiling in online systems that lead to "real-world" harms, teaching data ethics has never been more urgently needed. This paper explores current topics within data ethics syllabi. It analyzes the overlaps and divides between various approaches to teaching data ethics, from emphasizing FAT (Fairness, Accountability, Transparency), to the Public Interest Technology movement, to calls to reimagine "digital justice" from Critical Race and Digital Studies scholars, to applying AoIR's ethical Internet research guidelines, to the way philosophical texts on ethics from different cultural contexts are being applied to the digital age. We present a semantic, linked open data graph describing the relations between texts, courses, professors, and universities involved in teaching data ethics. While patterns in data ethics education will to some degree emerge organically from the data, we also use a "human in the loop" approach to identify and label these patterns. Patterns highlighted include the most-cited and most-assigned works in the field of data ethics, whether data ethics courses include computational training for data ethics problem-solving, whether courses target data science students or cohorts from other fields (highlighting a potential inroad to the data science pipeline), and institution and geographic location of courses. This tool was designed to inform our own approach to teaching data ethics, but we envision that it will be of use to other data ethics educators wishing to approach the subject in an exploratory way that highlights both consensus in the field and outliers. Our data and framework are open-source, and users can submit their own syllabi to update the map. Furthermore, our framework may be used to help map the institutional knowledge structures of other disciplines.

Keywords

Data ethics, education, semantic web

Introduction

Methods

We begin by collecting data surrounding university courses. Using course lists such as the tech ethics curriculum list provided by Casey Fiesler et al, and Dennis Tennen et al.'s Open Syllabus Project, we are able to derive preliminary information: the names of the courses, the instructors, and the host departments, among others (see Fiesler, 2019; Nowogrodzki, 2016). Fiesler et al's list is a openly-editable Google Sheets spreadsheet, called Tech Ethics Curriculum, containing data about roughly three hundred courses. The spreadsheet tracks courses titles, universities, departments, course description URLs, syllabus URLs, and

Syllabus URLs are only present for [TODO: X%] of

The semantic web, also known as "Web 3.0" or linked open data, is a relatively new system of conventions for standardizing and encoding graph data, such that it is universally interoperable, in a language known as RDF, or the Resource Description Framework. Some of the most well-known projects in the field include DBPedia, the set of parsed and inferred data from Wikipedia, and Wikidata, the data set which proposes to be the knowledge basis for Wikipedia. At its most basic, RDF data may be represented as a series of subject-verb-object triples, where each node has a stable URL. Social relationships between people, for instance, may be described as , where the angle-bracketed entities resolve to URIs. There exist a number of ontologies, or pre-defined sets of relations, which may be used to describe entities within their domains. We use a number of ontologies in conjunction: the Curriculum Course Syllabus Ontology (CCSO) describes relations between courses, universities, syllabi, professors, and learning materials such as texts; the Bibliographic Ontology (Bibliontology) describes metadata for articles, books, videos, and other media; and the Citation Typing Ontology (CiTO) describes citation relations between texts. Fig. 1 shows an example directed graph visualization, illustrating relations between these entities.

Given a course syllabus URL, we are able to download the syllabus, convert it to plain text where necessary, and quasi-automatically identify its bibliographic references. Since existing automated approaches to the extraction of bibliographic metadata are written for scholarly articles, rather than syllabi, we must supplement their use with manual extraction. Once these bibliographic items are extracted, we query bibliographic APIs such as CrossRef and Semantic Scholar, which allow us to resolve these texts to stable identifiers like DOIs. These APIs also provide us with rich metadata surrounding the citation networks of these texts, both outgoing (references to other texts) and incoming (citations of the article in later works).

We then resolve universities and departments to their Wikidata identifiers, which allows us to retrieve additional information with which we can organize our data, for instance the geographic coordinates of the university and its founding date.

We also resolve instructors and authors to their ORCIDs, where applicable, and

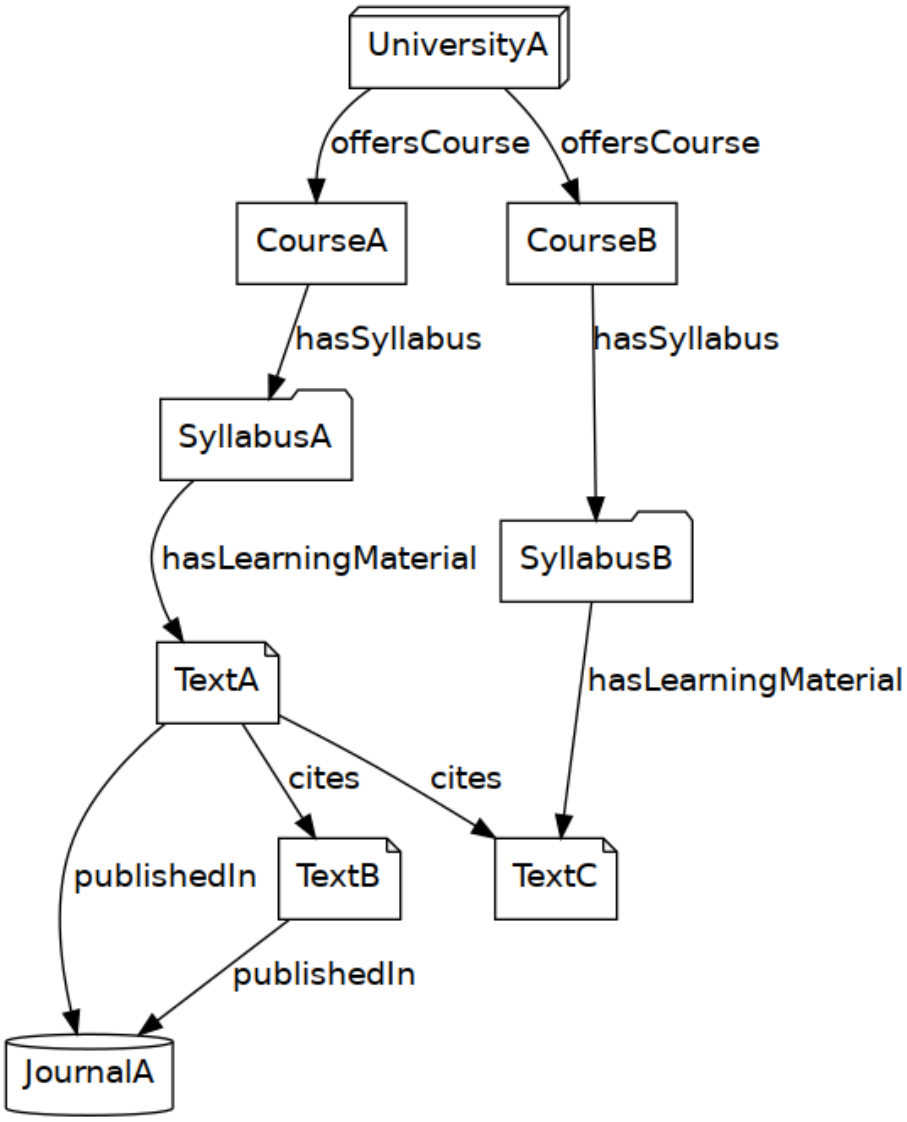

Figure 1: Flow chart of ontology data

this allows us to retrieve an author's other publications, and their past and present institutional affiliations.

At this point, we are left with a number of duplicates, for texts and authors, since there are a number of ways of citing a text in natural language. For this, we use a deduplication algorithm, which collapses texts which pass a certain similarity threshold.

References

Fiesler, C. (2019), "Tech Ethics Curricula: A Collection of Syllabi", *Medium*, November, available at: https://cfiesler.medium.com/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18 (accessed 18 January 2021).

Nowogrodzki, A. (2016), "Mining the secrets of college syllabuses", *Nature News*, Vol. 539 No. 7627, p. 125.