Founded by Salman Khan in 2006, Khan Academy is an online education platform with the stated goal of providing “a free world-class education for anyone, anywhere.”¹ The platform provides over 50 academic courses on a wide array of subjects and also hosts courses that were developed in collaboration with other non-profit organizations in order to educate on more specialized topics. Khan Academy has developed an organizing system for its educational resources that employs a hierarchical classification scheme and sequential inter-resource relationships that is meant to both make the discovery of these resources feasible, and to add value by placing these resources in the context of a broader curriculum.

What? - At the most granular level, the bulk of the primary resources for this organizing system are video lectures, which serve as the main content delivery mechanism for the courses. Videos are generally between 3 and 15 minutes in length. The academic content covered in these videos ranges from traditional grade-school curriculums (algebra, biology, American civics) to standardized test prep (SAT, GMAT, MCAT) to sponsored curriculums from Khan Academy’s partner organizations (The Museum of Modern Art, California Academy of Sciences, NASA). The fact that these resources are videos, as opposed to text files or static images, represents their primary domain. Their purpose as educational materials could be considered their secondary domain.

¹ https://www.khanacademy.org/about
Besides their suite of videos, Khan Academy also hosts a collection of instructional texts for some of their courses, as well a wide of array interactive educational units. These interactive units are primarily used for Khan Academy’s mathematics and computing courses – domains that are generally learned through practice. By replacing videos with interactive content, students are able to put to use the skills they are learning, and receive immediate feedback on their understanding of the material.

Why? – As noted in its mission statement, Khan Academy was developed in order to provide a free, high quality education to anyone with an Internet connection. In his book, *The One World Schoolhouse: Education Reimagined*, Salman Khan discusses how educational paradigms have not substantially evolved since the industrial revolution. He sees Khan Academy, and other online education resources, as tools that enable a more personalized and self-paced form of learning. Their use can make mastery learning (an education method whereby students must master concepts before moving on to more advanced one) feasible.

How Much? – Khan Academy uses a hierarchical classification scheme to organize its courses. At the top-level of this hierarchy, there are categories for broad academic domains (Math, Arts and Humanities, Science) as well as broad categories for courses that are not as strictly academic (Test Prep, College Admissions). These top-level categories are based on traditional academic disciplines that all students will be familiar with. This familiarity is crucial, as students often do not know the proper jargon for the material they are seeking to
learn. Beginning with common terms makes content discovery - one of the key interactions of this organizing system - easier.

The next level of the hierarchy breaks the large domains into more specific subjects. For example, the ‘Economics and Finance’ domain includes the subjects Microeconomics, Macroeconomics, and Entrepreneurship. One level deeper, the subjects are divided into topics. For the Microeconomics subject, this includes topics such as Elasticity and Forms of Competition. Finally, within these courses are even more granular modules that contain the actual videos and exercises. The aforementioned ‘Forms of Competition’ course contains three modules - ‘Perfect Competition’, ‘Monopoly’, and ‘Between Perfect Competition and Monopoly’. These three modules are organized in a sequential relationship, in order to convey the idea that a student should cover the first two before approaching the third. Furthermore, the individual modules have an internal structure. The videos within each module are also arranged in a sequential manner. Structuring resources in such a way adds additional semantic value, because it conveys to the student which materials are more basic and could be considered prerequisites, and which are more advanced. This additional information, embedded within the inter-resource relationships, enhances the single most important interaction that Khan Academy supports – enabling students to learn from their collection of resources.

*When?* – Originally, the process of adding resources to Khan Academy’s platform was relatively ad-hoc. Salman Khan would put together educational videos and post them to YouTube, without much additional organization. However, as the platform has grown and matured over the past few years, the organizational process
has formalized. Now, a course is carefully crafted, and the internal structure of its component videos and exercises precisely ordered, before being becoming publicly available within the context of the broader organizing system. While new courses are not frequently added, the most common impetus for the addition of new materials is a partnership with an outside organization. For example, Khan Academy is currently collaborating with The College Board to develop new SAT prep materials, which will be released in Spring 2015. Additionally, once resources have been added, they generally do not get reorganized.

How/By Whom? – Again, the job of organizing Khan Academy’s resources originally fell on the shoulder of its founder, Salman Khan. But as the organizing process has moved from being an impromptu one to a structured one, that responsibility has been given to a team of content experts that Khan Academy employs. Now, individuals who are specialized in the academic curriculum that is being taught via Khan Academy are responsible for the organization.

---

2 https://www.khanacademy.org/sat
For our IO Lab final project, my team decided to build a platform that aggregates online educational resources. I took it upon myself to develop the organizing system for our website, and to source all of the information that would populate it.

First, we had to develop a classification scheme for the courses that we would be featuring. As none of our group members could be considered experts in curriculum design, we decided that it would be best to use an existing system as a starting point. For this, we chose the hierarchical classification that Khan Academy uses, and made several substantial alterations. As we wanted our organizing system to be general enough to accommodate resources from a wide variety of sources, and enhance the general level of recall, we had to reduce the specificity of our categories. We moved from Khan Academy’s 10 top-level categories to only 5 purely academic ones. Next, we limited our hierarchy to only two levels of categories – generic domains and more specific topics. Some effort was put in to reducing the amount of topics as well. For example, Khan Academy uses a different category for each grade level in math (3rd-grade math, 4th-grade math, 5th-grade math, etc.). We consolidated all of these topics into a single ‘Grade School Math’ category. We also wanted to add some faceted classification to our organizing system. To achieve this, we categorized each course based on a presumably required skill level (beginner, intermediate, proficient, or expert). We decided on this so users would be able to select courses not only based on the content that they were covering, but also on the
user’s own skill level. This enabled users to increase precision when searching through our courses.

Being aware of our lack of expertise in online courses, we also wanted to enable our organizing system to hold some crowd-sourced information. To achieve this, for each of the courses in our organizing system, we built fields for ‘rating’ (the quality of the course), ‘rigor’ (the difficulty of the course), and tags. These three fields could all receive input from the front-end of our website.

Next, we had to decide on which educational resources we would be including in our organizing system. We eventually settled on three platforms – Khan Academy, Udacity, and Coursera – to source information from. The decision to use these platforms was based on two important features. Firstly, because these three providers have established themselves as leaders in the online education space, we felt comfortable trusting that the resources they produce are of high quality. Secondly, these three platforms all provide publicly available APIs that we could utilize. Other platforms that were considered – namely Udemy and Lynda.com – had to be abandoned because they did not provide publicly available APIs with which we could access their organizing systems.

After selecting our three course providers, we obtained course catalogs via their respective APIs. All course catalogs were contained within a JSON format. A Python script was written for each provider that extracted the relevant data points, and reformatted them to fit the schema of our organizing system. At this point, we ran into several interoperability issues. The most notable of these issues came from how the different course providers store their course data. As mentioned in the
accompanying case study, Khan Academy uses a purely hierarchical classification scheme. Not only is this demonstrated via their interface, but in their API as well. All of the Khan Academy’s course data is stored in a JSON files that uses a tree structure. The JSON begins with a single node that contains dictionaries for the top-level curriculum domains (Math, Science, Humanities, etc.) which then contain dictionaries for more granular topics, and all the way down through the individual videos and exercises. While this structure lends some semantic value, it is also cumbersome to navigate. Conversely, Udacity and Coursera use faceted classification schemes for their courses. Again, this is mirrored in their APIs, which provide JSON files that use a much simpler list structure.

Interoperability issues also arose from different platforms reporting data with different formats. For example, the schema for our platform requires a field called ‘length’ that indicates the amount of time a student must spend to complete a course. Coursera provides that exact piece of information, while the other two platforms provide tangentially related information. Each Udacity course only has data on an estimated weekly workload for the course, and each Khan Academy course only provides data on how many modules it contains. Because of these kinds of discrepancies, we often had to transform information into the proper format, before inserting it into our organizing system.

Upon completion, we had developed an organizing system that held data for over 1,200 online courses that were taught in 12 different languages. These courses cover a wide variety of material, ranging from the art history, to genetics, to data visualization. While there is still much that could be done to refine aspects of our
organizing system, namely in terms of more accurately classifying our courses, we have still achieved our initial goal of enabling users to more effectively find the best online course for their needs.

The final product can be seen here: https://io-db.herokuapp.com