Aggregation and Search: Baskets for Berrypicking

Joseph Busch, Taxonomy Strategies

Abstract

The goal of search is to reliably find what you are looking for, to be able to type in a highly variable query and return the most relevant result or the right answer every time. These days, effective search relies to a large extent on natural language processing and analytics. The purpose of KOS is not to find items or answers, but to group or aggregate content into collections for review or further refinement. This can be pre-search to build a collection to search on rather than the whole universe, or it can be post-search to characterize the search result set, or refine the results. It's important to consider the kind of search result user experience when the KOS is designed. The aggregation scenario means a broad and shallow scheme with discrete categories is needed. The focus needs to be on designing the baskets for gathering "berries" rather than the berries themselves that users will be picking. This paper lays out some use cases for this aggregation scenario and presents some examples.

Introduction

In her iconic 1989 "berrypicking" paper, Marcia Bates [1] reflects on how (footnote chasing, citation searching, journal run, area scanning, subject searches in bibliographies and A&I services, and author searches), where (in a physical library), and who (trained users) is doing research in imagining how research would be able to be done in a full-text digital library. She was excited by the prospect of being able to easily simulate the way users can readily switch from one information retrieval mode to another in a physical library, but online. Up until that time, online information retrieval was conducted in a very formal and limited way where the user needed to learn the structure and terminology of document representation to formulate their queries in order to find a match.

There are two key contexts that are assumed in Bates' vision. First, she assumes that the audience is trained users doing research. Second, she assumes the content of the digital library will be similar to a physical library—books, magazines, newspapers, research and reference tools, etc. While Bates does foresee that searching will be full text rather than (or in addition to) catalog and index metadata, she does not foresee the reach of the World Wide Web or the "Cloud" that will enable ubiquitous access to content by anyone who has a connected device.

Bates also did not foresee the new barriers to information retrieval where the promise of everything digital both flooded the commons with undifferentiated information, and also restricted access to much of the published work behind pay walls. Unless one has access to a digital library in a research or educational institution as staff or student, many users are left to use what can be gleaned for free on the Internet where even a preview of a full text journal article can be very expensive.

Besides cost, the culture of ubiquitous web access has dramatically changed information seeking behavior from the time when Bates was imagining a digital library utopia. While it is now easy to type in a search to verify certain facts on a mobile device; engaging in the type of berrypicking research that Bates reflects on, still requires access to a research library (though not a physical one). In the offices of companies, non-profit organizations, and government agencies libraries have disappeared. At the same time, the generation and use of digital content inside organizations has

continued to explode. As the amount of available data has grown, the problem of managing the information has become more difficult, leading to information overload. [2]

The following table lists methods for handing large amounts of information that are presented to users when they use a web search engine or enterprise search interfaces, and identifies which of the berrypicking search interfaces imagined by Bates they are similar to.

WWW/Enterprise Search Interfaces	Berrypicking Search Interfaces
Natural language processing and analytics	Citation searching
Find an expert	Footnote chasing
Guided navigation	Subject searches
Search results as collections	Area scanning
Visualizing collections	
Knowledge graphs	

Each of these methods is discussed briefly with examples in the following section.

Natural Language Processing and Analytics

Since the last decade of the twentieth century, web search engines have been developed that use web crawlers to systematically browse websites and index all the accessible public information. A similar method is used in enterprises to crawl servers and sites within an organization to browse and index information. Natural language processing (NLP) methods are used to analyze the semantic and syntactic features of text to identify and index meaningful entities beyond simple term frequency and document length. Various measures of site and content use are also recorded and analyzed so that more popular information is promoted. Finally, many content providers pay to promote their information so it is more highly ranked in search results, or simply featured. Aside from the unintended consequences of this marketplace mirroring the biases in society, over the past 25 years this ecosystem has matured and people type in questions on their mobile device and for the most part get answers useful whenever they feel like it.

The availability of computing resources to power this enormous information infrastructure have been central to advancing the development and use of natural language processing. But the central role of analytics to rank and promote one information item over another are sometimes overlooked. This idea was central to Google's early success, and it is similar to citation searching. Citation searching involves finding an information item of interest and identifying what other information items refer to it. In scholarly research, citations are footnoted sources that may also be listed in a bibliography. The more frequently a source is cited, the more important it is generally considered. On the WWW, citations are manifested in various ways. One is via hyperlinks. The frequency with which a source is hyperlinked, plus the nature of the source (i.e., is it commercial, educational, or governmental) increases the ranking. The second factor related to WWW citation is the frequency with which a webpage is visited. Additional factors may also be considered in webpage or content item ranking in online search interfaces, particularly the content type or genre. For digital content, genre is a complicated matter that requires a separate paper itself.

Find an Expert

In the third decade of the twenty-first century, when information seekers encounter and use a search box, they expect it to function like Google. This means the user assumes that they are doing a full-text search. This is true on the WWW, and it is also the case inside organizations where users encounter what they assume is enterprise search. Users also assume that the collection they are

searching includes the information that they are seeking. So, if the search doesn't return a useful result within the first few search result pages, they will most often abandon rather than reformulate the search. In many cases, they will engage another channel, such as contacting an account representative, support, or a colleague to get the answer to their question or complete their transaction. [3]

Expertise directories that help employees learn "who knows what" is a baseline knowledge management application. Expertise finding is most often done via social networks such as LinkedIn or ResearchGate, or a knowledge base that holds resources that provide a detailed response to common questions (more than simple FAQs). Seeking an expert could be considered a form of footnote chasing. There are no longer many reference librarians available to turn a research question into a teachable moment for guiding a user through the research process. Instead, chatbot dialog systems have become a common channel for delivering initial customer support online.

Guided Navigation

Given the current information ecosystem that exists in corporations and non-profits today, it's important to figure out how to take advantage of ubiquitous search as an entry point for browsing or exploring the collection scoped by a user's search. What methods can be used to break the paradigm that the relevant result must be near the top of the search results page (SERP) or within just a few scrolls down, especially when the search result contains hundreds or thousands or millions of items. If rich metadata exists, as it did in Bates' CD ROM utopia, and in Vannevar Bush's Memex [4], abstracting and indexing (A&I) databases, online library catalogs, and eCommerce websites provide a model for refining a very large text search collection in a few clicks to a manageable set. Steve Papa, the former CEO of Endeca, an early faceted search engine, coined the term "guided navigation" to describe the process of refining a rich metadata search result. [5] The metadata-controlled vocabularies don't need to be that large or complex, to provide the granularity to accomplish this task. For example, four metadata-controlled vocabularies of 10 values each have the same discriminatory power as one taxonomy of 10,000 values. An example is shown in Figure 1.

Content Types	Health Topics	Industries	Substances
FAQs	Children's Health	Agriculture	Allergens
Forms & Applications	Food Safety	Automobile Repair	Biological Contaminants
News & Announcements	Health Advisories	Chemical	Carcinogens
Policies & Procedures	Health Effects	Construction	Chemicals
Publications	Health Risks	Dry Cleaning	Explosives
Presentations	Occupational Health	Electronics & Computer	Liquid Waste
Regulated Product Information	Pesticide Effects	Energy	Microorganisms
Reports	Seniors' Health	Extractive	Ozone
Tools & Databases	Sun Protection	Food Processing	Pesticides
Transcripts & Statements	Toxicity	Leather Tanning & Finishing	Radioactive Waste

Figure 1-Example of four metadata-controlled vocabularies.

Broad and shallow faceted knowledge organization schemes or KOS have great utility, and are easier to build, maintain, and apply than narrower and deeper KOS.

Guided navigation or faceted search is a method that augments subject searching by inviting the user to narrow their search by using multiple filters that are the facets of the KOS. Guided navigation is familiar to users because it is how they shop online. Guided navigation is less frequently used on content websites, but can be used with the same effect as on shopping sites. Figure 2 is an example of guided navigation applied on a content site epa.gov.

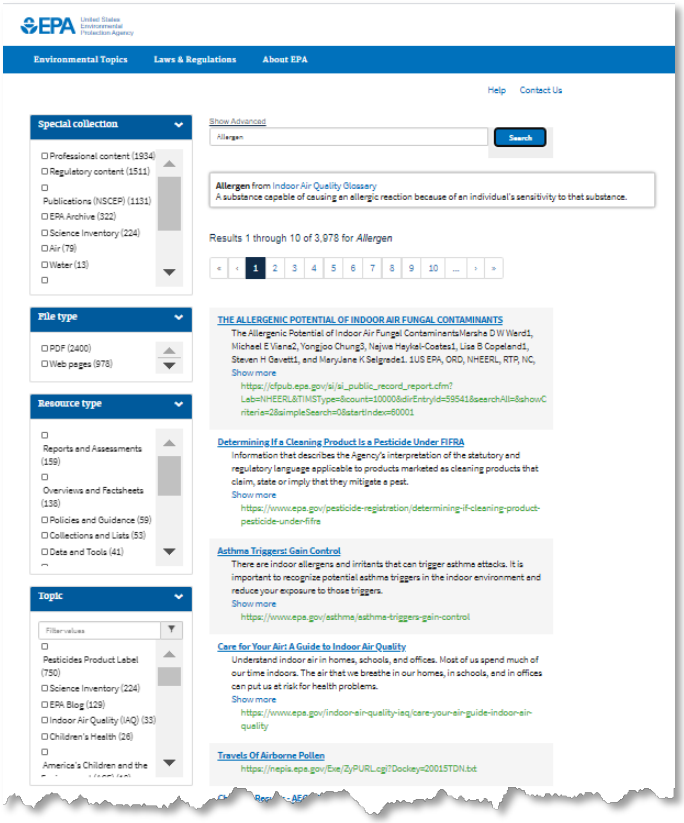

Figure 2-Example of guided navigation on a content website epa.gov.

Search Results as Collections

Every search should be thought of as a collection of results. Instead of presenting text search results as a list (of bibliographical references) where only the first few scrolls are likely to be reviewed, a representation of the whole search result collection could be presented. This provides the user an overview of the available information, and invites them to refine or start with a new search. Refining the search offers a new collection with clues as to what is in that collection. Figure 3 shows results for the search term "rover" (as in Mars Rover) from four NASA collections. While the search returns more than 200,000 items, the results are presented as a collection with the top occurring categories selected from a faceted KOS (the NASA Taxonomy [6]).

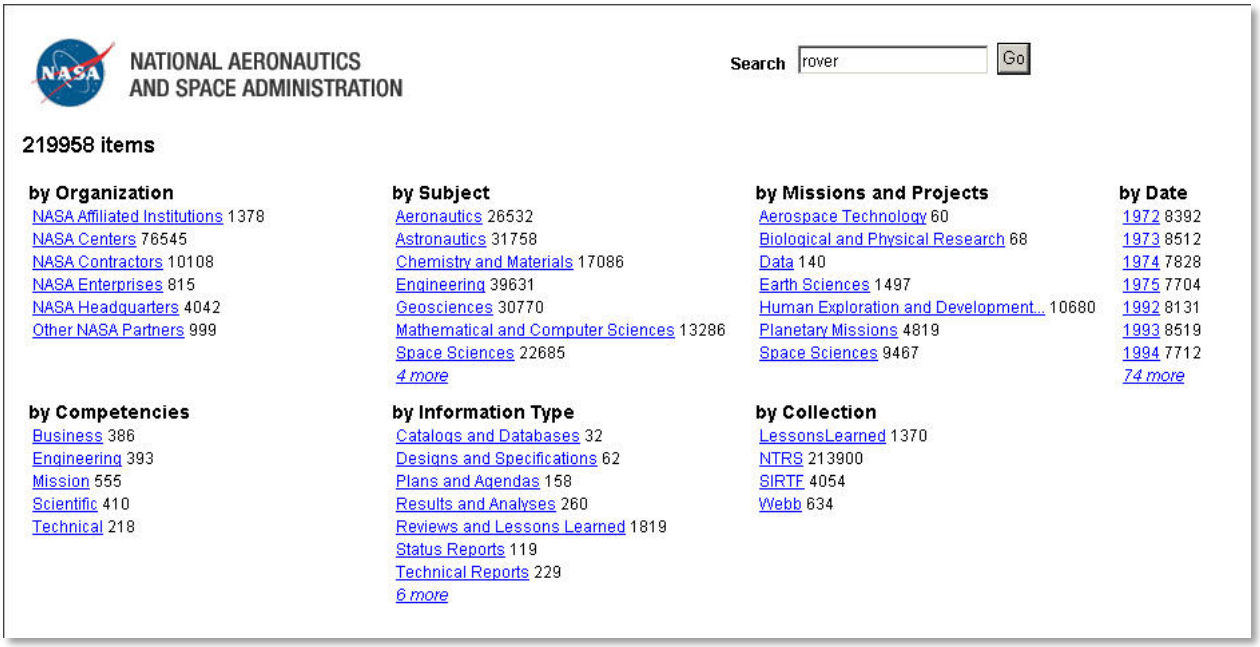

Figure 3 - Example of search results from four NASA collections presented as a collection with contextual categories and counts (November 2004).

The following table shows the NASA Taxonomy which is a broad and shallow KOS, consisting of 11 facets. The five facets used in the search results shown in Figure 3 are asterisked.

Facets	Descriptions
Access Controls	Sensitivity and access control
Audiences	Who is the content intended for
Business Purpose	Why the content was created
Content Types*	The genre of the content
Instruments	Flight payloads that yield science
Locations	Sites where work occurs – on and off Earth
Missions and Projects*	NASA's lines of business and work
NASA STI Subject Categories*	The topic of the content
NASA Workforce Competencies*	What field or discipline is relevant
Organizations*	NASA organizations
Work Breakdown Structure	Work components

Further refinements can be made by source collection (not to be confused with search results collection) and year of publication.

Visualizing Collections

Collections of search results with rich metadata can also readily be visualized, for example, using maps and charts. A web page that uses content from more than one source is sometimes called a mashup. Dashboards are a particular type of mashup that provides a summary of key-performance indicators (KPI's). Mashups and dashboards often use an application programming interface (API) to quickly integrate data from a source into the web page, or a data visualization application such as Tableau to explore tables in a relational database. Because these webpages are integrated with data

sources, the visualization is often interactive so that user can explore the data source, or drill down to see more specific components of the visualization. Figure 4 is a map that provides a national visualization of themes (topics) for all U.S. states, as well as a drill-down to a state with county/city items. Figure 5 is a visualization that shows total and KPI amounts awarded by lines of business and in summary for the whole enterprise. The user can click on any data bar to see the source table of items that make up the total.

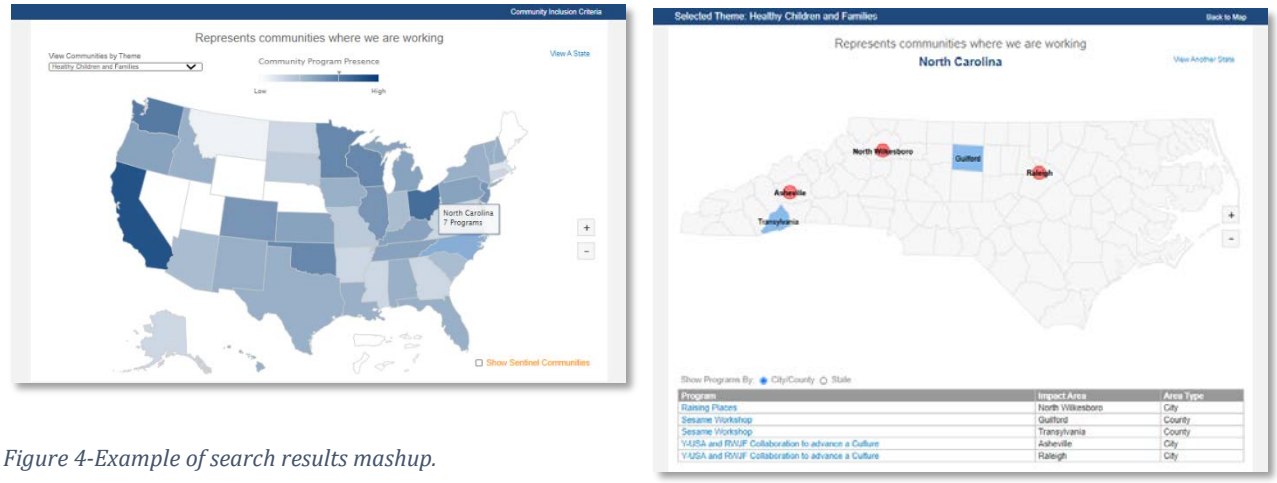

Figure 4-Example of search results mashup.

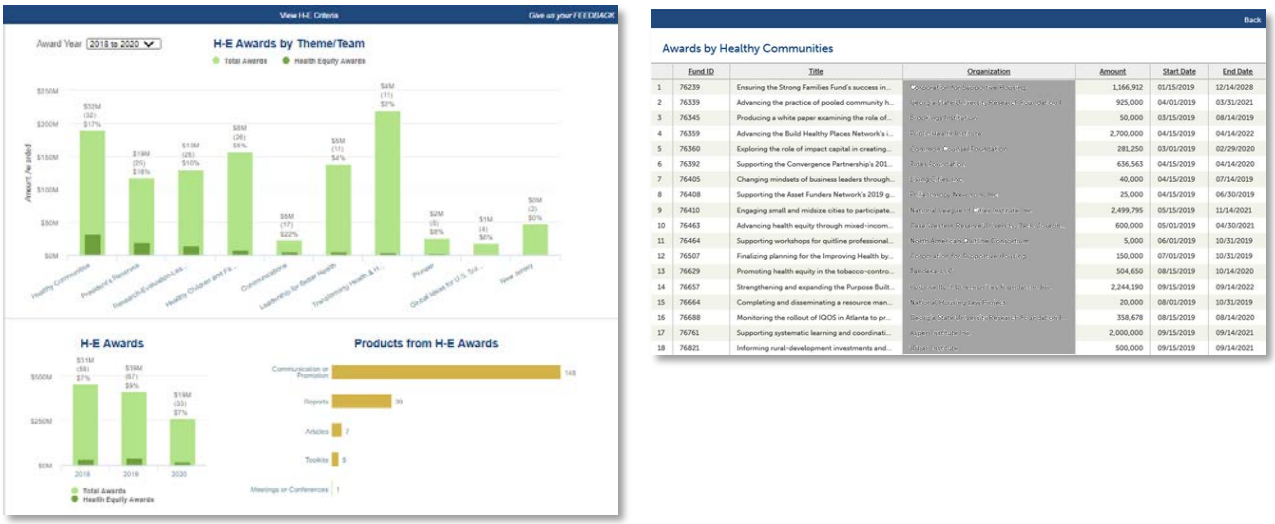

Figure 5-Example of search results dashboard.

Visualizations of source data by category was not foreseen by Bates in her berrypicking model, but this interactivity is exactly what she was excited about. The hard work here is designing a KOS that can be implemented to enable and support quick and easy integration with visualization API's. There is even more work to curate such resources so that they can be sustained, discovered, and used over time.

Knowledge Graphs

An enterprise knowledge graph is a representation of an organization's knowledge domain and artifacts that is understood by both humans and machines. It represents an organization's

knowledge assets, content, and data—people, places, documents, multimedia, data, etc.—and how these things are related to each other. Those things and relationships are represented using a metadata model specific to the industry or developed for the organization. Typically, this is an ontology that defines classes for the things, properties for the things, and relationships among the things. It is a lot of work to build up the infrastructure for an enterprise knowledge graph, and then establish and implement the policies and procedures to link up a critical mass of an organization's materials.

Figure 6 compares an ontology for the physics domain designed to support the submission and publication of scholarly and conference papers [7] with the knowledge graph for the same concept designed to be presented on the Google search results page. The PhySH figure shows Research Areas related to the selected Discipline "Fluid Dynamics", as well as related research areas where they have been identified. Similarly, Physical Systems, Properties, Techniques, and Professional Topics related to the selected Discipline would be shown in those tabs in this user interface. The Google knowledge graph presents a Wikipedia definition, images, books, and related topics that people searched for.

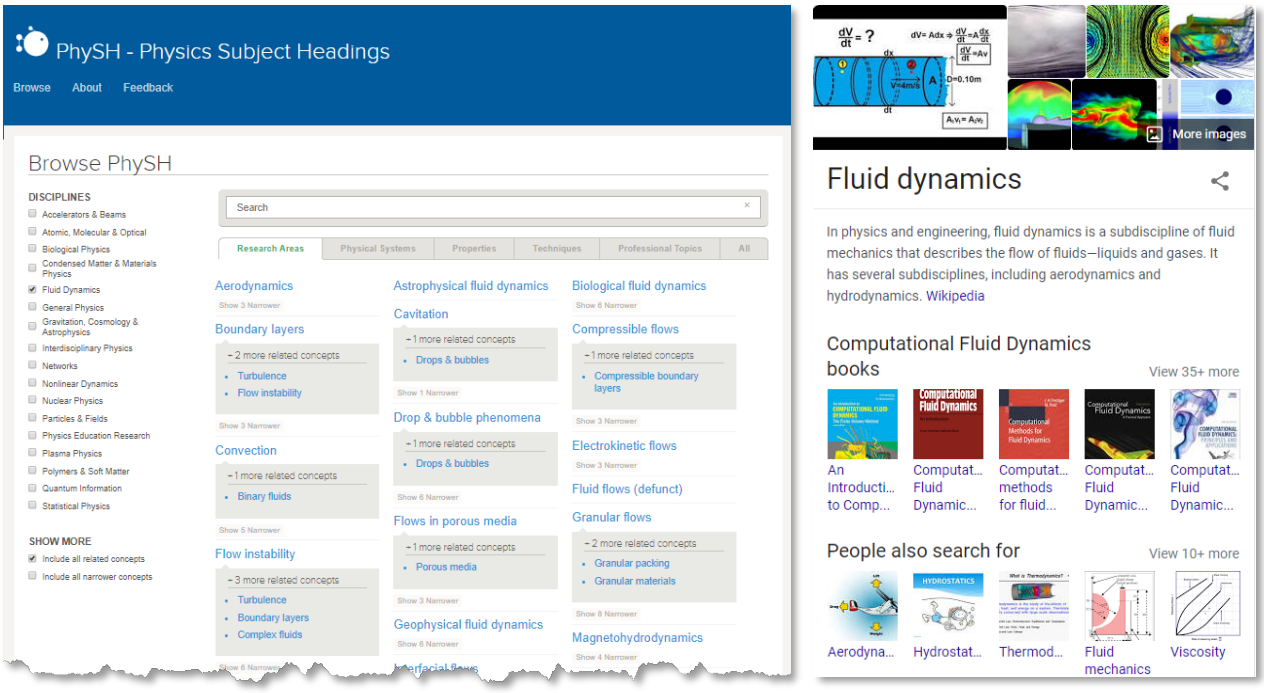

Figure 6-PhySH ontology and Google knowledge graph for Fluid Dynamics.

The Google initiative to provide succinct responses to queries for well-documented named entities such as famous people, locations, creative and built works, theories, etc. gathers information from many different sources and presents it in the right rail of the Google search results pages when appropriate.

Today there is not an agreed upon definition of knowledge graph, but (despite Google) it is often considered to be the same as ontology. The goal of a knowledge graph whether for Google or an enterprise is to improve search results. A knowledge graph or ontology is a type of KOS. Bates would consider the renewed purpose of ontology to improve the search interface a great opportunity to engage the user in an interactive research process.

Conclusion

The purpose of KOS is not to find items or answers, but to group or aggregate content into collections for review or further refinement. This can be pre-search to build a collection to search on rather than the whole universe, or it can be post-search to characterize the search result set, or refine the results. It's important to consider the kind of search result user experience when the KOS is designed. The aggregation scenario means a broad and shallow scheme with discrete categories is needed. The focus needs to be on designing the baskets for gathering "berries" rather than the berries themselves that users will be picking.

References

[1] M. Bates. "The design of browsing and berrypicking techniques for the online search interface." 13(5) *Online Review* (1989) pp. 407-424, and in: M. Bates. *Information Searching Theory and Practice: Selected Works*. Vol. 2. Berkeley: Ketchikan Press, 2016. pp. 257-278.
[2] "Information explosion." From: *Wikipedia*. https://en.wikipedia.org/wiki/Information_explosion. Last checked: 6/8/2020.
[3] APM Music customer interviews, May-June 2020. Medline Industries customer interviews, March 2019. Robert Wood Johnson Foundation program staff interviews, November 2016.
[4] V. Bush. "As we may think." *The Atlantic* (July 1945). https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/. Last checked: 6/8/2020.
[5] S. Papa. "The faceted navigation and search revolution." *KM World* (March 23, 2006) https://www.kmworld.com/Articles/White-Paper/Article/The-Faceted-Navigation-and-Search-Revolution-15378.aspx. Last checked: 6/8/2020.
[6] *NASA Taxonomy*. Last updated: 05/08/2012. https://vocabularyserver.com/nasa/. Last checked: 6/9/2020.
[7] *PhySH – Physics Subject Headings*. American Physical Society. https://physh.aps.org/. Last checked: 6/11/2020.