A Specialized Data Crawler for Urban Wood Information

R. Edward Thomas
Omar Espinoza

Abstract

The Internet is composed of >50 billion Web pages and grows larger every day. As the number of links and specialty subject areas increases, it becomes ever more difficult to find pertinent information. For some subject areas, special purpose data crawlers continually search the Internet for specific information; examples include real estate, air travel, auto sales, and others. The use of such special purpose data crawlers (i.e., targeted crawlers and knowledge databases), also allows the collection and analysis of agricultural and forestry data. Such single-purpose crawlers can search for hundreds of keywords and use machine learning to determine whether or not what is found is relevant. In this paper, we examine the design and data return of such a specialty knowledge database and crawler system developed to find information related to urban wood utilization—products made from timber harvested in cities and municipalities. Our search engine uses intelligent software to locate and update pertinent references related to urban wood as well as to categorize information with respect to common application and interest areas. At the time of this publication, the urban wood knowledge database has cataloged >700 publications regarding various aspects of urban wood.

Traditionally, logs from trees originating in urban areas of the United States have been disposed of as low-value resources, typically through chipping, burning, or landfilling. There are approximately 74 billion trees in urban areas of the United States, and when trimming or removal is necessary, because of development, disease, or other reasons, they are considered "wood waste." In recent years, new industries have emerged to capitalize on timber from urban trees by offering unique aesthetics, historical significance, sustainability, and sentimentality derived from inimitable wood supplies. Typically, urban wood products have a more character-marked appearance and can symbolize a strong emotional bond between a person and a place or time. The range of urban wood products is diverse, including logs, lumber, slabs, art, furniture, and other wooden household products. This industry is expected to grow significantly (Pitti et al. 2019). Recent work by the authors has revealed that one of the biggest challenges that this emerging industry faces is to create and grow marketing awareness of the benefits and challenges of urban wood utilization. This project addressed the aforementioned challenge by developing a comprehensive knowledge database about all aspects of urban wood utilization. There are many subject areas that revolve around the topic of urban wood, including products, markets, processing, health, safety, forest disease, invasive insects, and fire, among others. Thus, given the broad combined product and subject area of urban wood, it can be difficult to find specific information.

Complicating matters is the sheer size of the Internet. The total number of Web pages on the Internet is unknown; however, estimates place the total between 50 to 60 billion Web pages (de Kunder 2023, Huss 2023). In addition, it is estimated that only 5 billion, or 10 percent of the total Internet size, is indexed by search engines with an estimated 250,000 Web pages added each day (Huss 2023). Hence, finding pertinent data on the Internet is a task that continues to become more complex and time-consuming. When a topic is sufficiently complex, spans several languages, or if there are many ways to refer to a specific topic, searches become even more difficult. In addition, related keywords can change the meaning of the topic being researched, imparting a more specific meaning that can be desirable. For example, adding "economics" and "market" keywords to a search for "urban wood" greatly narrows a search for information. However, attempting to manually perform a series of Web searches using the combinations of potential keywords and acronyms, while excluding other keywords and further trying to collate all the information from the

The authors are, respectively, (ed.thomas2@usda.gov [corresponding author]), USDA Forest Serv., Forestry Sci. Lab., Princeton, West Virginia; and (oaespino@umn.edu), Univ. of Minnesota, Bioproducts and Biosystems Engineering, St. Paul, Minnesota. This paper was received for publication in October 2023. Article no. FPJ-D-23-00045.

Forest Prod. J. 73(4):350–356.
doi:10.13073/FPJ-D-23-00045

various searches together, is time-consuming and frustrating. Temporal data that require collection at regular intervals increase the need for an automated approach to this process.

Special-purpose data aggregators, often referred to as crawlers or spiders, are specialized software programs that browse information on the Internet and catalogs, stores, and if so desired, aggregate select data for a specific purpose. The leading general search engines, such as Google, Bing, DuckDuckgo, and others, use countless crawlers to constantly search the Internet and catalogue their findings in their proprietary databases to be called up when a user searches for a specific keyword or keywords. Hence, it is the data from such proprietary databases that users see when they conduct a search with a general search engine and not the data that are actually available on the Internet. However, as pointed out before, although such general-purpose search engines are good at finding specific information for a narrow topic (such as looking up the meaning of "urban wood"), collating all the available information on a specialized topic (such as looking up all relevant information on the various products derived from urban wood in combination with "urban wood") is difficult and time-consuming. To address this need, specialized data aggregators are developed.

Specialized data crawling is used for numerous purposes. For example, the travel industry has several Web sites that are powered by the products of successful data-crawling aggregation. However, unlike other industries, few examples of Web crawlers collecting agriculture- and forestry-related information exist. A large-scale example is AgroFE, a European Union (EU) project to develop an agro–forestry training knowledge database (Herdon et al. 2014). These authors discuss the use of specialized data aggregators to find information for creating the knowledge database. Another example is the Big Data Europe project (Albani et al. 2016) that targets a wide range of information areas such as agriculture and forestry. Big Data Europe seeks to intelligently combine data from remote sensing (crop type, status, land cover, etc.) with textual data collected from news feeds, social media, and other sources via specialized data aggregators. A recent crawler and knowledge base system was created to collect and categorize information specifically related to the construction and use of Cross-Laminated Timber (CLT; Thomas et al. 2020). This system was developed to improve the level of awareness about CLT in the construction industry and allow stakeholders to access and share knowledge about the state of research and implementation of CLT in the United States and the world.

This paper documents the development and the outcome of a specialized search engine (e.g., a crawler) and knowledge database developed to discover, categorize, store, and disseminate links to information related to urban wood. The information found, summarized, and categorized by the crawler can be accessed at https://urbanwooddatabase.umn.edu. As such, this paper is not a step-by-step guide to developing a Web crawler and enabling the search for the required content, because all informational and crawler projects will differ somewhat depending on the subject matter and the type(s) of data in question. However, the approach and the methodology described in this publication can serve as a guide to collecting data regarding other forestry and agricultural subject areas. For those wishing to develop such a system for a specific topic, there are numerous resources and guides available, such as the introduction to spider development by Heaton (2002).

Methods

The objective of this project was to develop a specialized search engine and knowledge database that operates and collects relevant links and data related to urban utilization wood with minimal human oversight. The most important requirement was that the crawler must have the ability to keyword-search Websites and Web documents for potentially hundreds of keywords, categorize information found, and store those links in a knowledge database. Hence, the crawler was tasked to search not only Web pages, but all types of documents commonly found on the internet, including, but not restricted to, Portable Document Format (PDF), text documents, presentations, spreadsheets, and others. This required the crawler to be capable of processing a wide array of file formats. Each source file needs to be searched for the keywords and all relevant documents need to be stored in the knowledge database. Another important requirement is that the system must allow users to easily search for and view relevant knowledge based upon their search criteria.

To assure relevance and timeliness, the knowledge database system operates multiple crawlers at once to build the database, and to maintain and verify links and knowledge. However, finding and maintaining knowledge is just one critically important activity, another one is the ability to assess the relevance of the knowledge found and to categorize it according to the system developed—in this case, into 19 different subtopics established by the research team. Finally, to ensure that relevant, high-quality results are provided to user queries, an administrator manually verifies documents that are added to the system's knowledge database before they are officially added to the knowledge database and made available to users.

The urban wood knowledge database uses the MySQL Enterprise database system (Oracle Corp. 2023a) to store information. MySQL is a high performance; robust, secure, and reliable database management system. These key features make MySQL well-suited for the urban wood knowledge database project. The knowledge database stores the Uniform Resource Locator (URL), a brief synopsis or abstract, and the page title, author, and keywords found. In addition, any links to other sites found at the URL are also stored. This enables the knowledge database to build a library of referral links that permits the database to determine how many different URLs refer to any specific page. The referral count data combined with the number of keywords found on a site are key components for the quality ranking of Web pages.

The urban wood knowledge database system was developed using the Java programming language (Oracle Corp. 2023b) in combination with several programming libraries, (e.g., modular blocks of code written to do specific tasks). These libraries provided features vital to the project and incorporating them greatly reduced development time. The software libraries that were most important to the successful development of the crawler system were

- *Apache HttpCore*: The Apache HttpCore (Apache Software Foundation 2022a) classes and interfaces were used to support the connection and transport of data and information from Web servers to the crawler processes.

Table 1.—Core keywords used by the Urban Wood databases.

Urban forest	Urban wood	Urban forestry
Community forestry	Public	Street tree
Urban tree	City	Cities
Salvaged	Upcycled	Resawn
Reused	Recycled	Repurposed
Recovered	Reclaimed	Urban timber
Urban lumber	Urban interface	Urbanization
	Wildland urban	

- *Apache Tika*: The Apache Tika (Apache Software Foundation 2022b) toolkit detects and extracts metadata and text from over one-thousand different file types.
- *GROBID*: A machine-learning software system for extracting bibliographical information from scholarly documents (GROBID 2019).
- *weka*: A collection of machine learning algorithms for data analytics (classification) and predictive modeling (Frank et al. 2016).

The selection of keywords for the crawler searches is critically important because the set of keywords directly affect the information resources that the crawler finds. To determine the most meaningful and common keywords associated with urban wood information, a series of Rapid Automatic Keyword Extraction (RAKE; Rose et al. 2010) analyses were performed. RAKE is an algorithm that processes documents to determine the keyword set and the frequency each keyword occurs. We used a python implementation (Medelyan 2018) to process a sample of 200 urban wood documents that included research papers, news and magazine articles, and product and information bulletins that represented the complete subject area of urban wood. RAKE compiled a list of descriptive keywords for each document. Pooling the keyword lists of the 200 documents, 21 core keywords were found that were common across subject areas and most documents. Once the crawlers were operating, we discerned that the keyword "urbanization" was sometimes associated with relevant links and it was added to the core keyword list (Table 1). The RAKE analyses identified an additional 405 product, use, and subject area keywords that were commonly found in urban and reclaimed wood literature. An additional 170 tree species names are included as keywords. Thus, the crawler searches each link for a total of 597 keywords.

To avoid including Web pages containing off-topic information, even though they may contain one or more of the keywords we are searching for, we required that at least one core keyword as well as two additional keywords be found. In addition, we use a list of 151 exclusionary keywords, such as "Medication," "Pharmacy," "Dating," "Food," and "Comic," among many others to avoid off-topic Websites. These keywords are discovered and added during crawling as potential links are found that match target keywords, but are off-topic. An examination of these links often reveals a common keyword that can be used to exclude those links. In addition, to avoid crawling some sites entirely, a set of excluded URLs are maintained. On some Web sites there are segments of the URL that indicate Web pages for which no relevant links will be found. Some examples of this are "contact-us," "alumni," or "shopping-cart." To avoid searching these sections of a URL, we use a sub-URL exclusion that examines the URL for these and other URL keywords to avoid. The crawler system has 647 excluded sites and 266 sub-URL exclusions. The keyword and URL exclusions allow the crawlers to stay on topic and avoid wasting resources on sites and Web pages that will yield no pertinent data. In addition, the knowledge database does not crawl or catalog links from social media Websites. These steps reduce the number of links found that contain relevant keywords, but are nonrelevant to the topic area of urban forestry to approximately 15 percent. With the latest developments in artificial intelligence, such as ChatGPT (OpenAI 2023), comes the potential to better filter out information that contains relevant keywords, but is not relevant to the topic of urban wood.

The design of our crawler requires a starting point, for which we used a list of 550 'seed' links. The 'seed' links were obtained using current general-purpose search engines. A search was conducted for 11 subject-category keyword strings ("invasive insect," "products," "markets," "ecosystem," "jobs," "slabs," "flooring," "economy," "health," "lumber," and "flooring") combined with an urban wood keyword string ("urban wood," "urban forest," "community forestry," "recycled wood," and "reclaimed wood"). For each search, the first 50 links returned by the search engine were used to seed the crawler.

Figure 1 depicts the flow of decisions performed to analyze the content at a given URL. The process begins with a crawler retrieving a URL that needs to be processed from the MySQL database. If the URL directly points a document, such as a PDF, the document is downloaded using the Java URL Connection class and converted to a text document using the *Apache Tika Java* class library. Otherwise, the link contents are downloaded and text content extracted using the *Apache HttpCore* library methods.

The text document from the URL (either the Website content or the content of the file downloaded) is first searched for exclusionary keywords that could indicate that the link is off topic. If off topic keywords are found, the link is marked as such in the MySQL database and processing of the URL is halted. Otherwise, the text is searched for urban wood–specific keywords. Recall that to improve selectivity, we required that a core keyword along with two subject area keywords be found before a Web page is further processed. These requirements allow the crawler to focus tightly on the information resources that visitors to the urban wood knowledge database will find most valuable.

If the text from the URL meets the keyword requirements, then it is processed using a *Weka*-based classifier (Frank et al. 2016). Weka is a collection of machine learning and data analysis software that supports the classification methods used by the crawler. The classifier is trained to classify information into the following subject areas (Table 2). The classifier was trained using a set of 308 documents and links that covered the subject areas targeted by the knowledge system. The classifier operates by assigning a probability to each of the subject areas. If the highest probability is with the "other" subject areas, then the information is regarded as off topic. Otherwise, the two subject areas with the highest and next highest probabilities are recorded as the primary and secondary subject areas, respectively.

If the text is from a downloaded file (any form of PDF, MS Word, OpenOffice, etc.), then the text is processed using *GROBID* (GROBID 2022) to extract the title, abstract, keyword list,

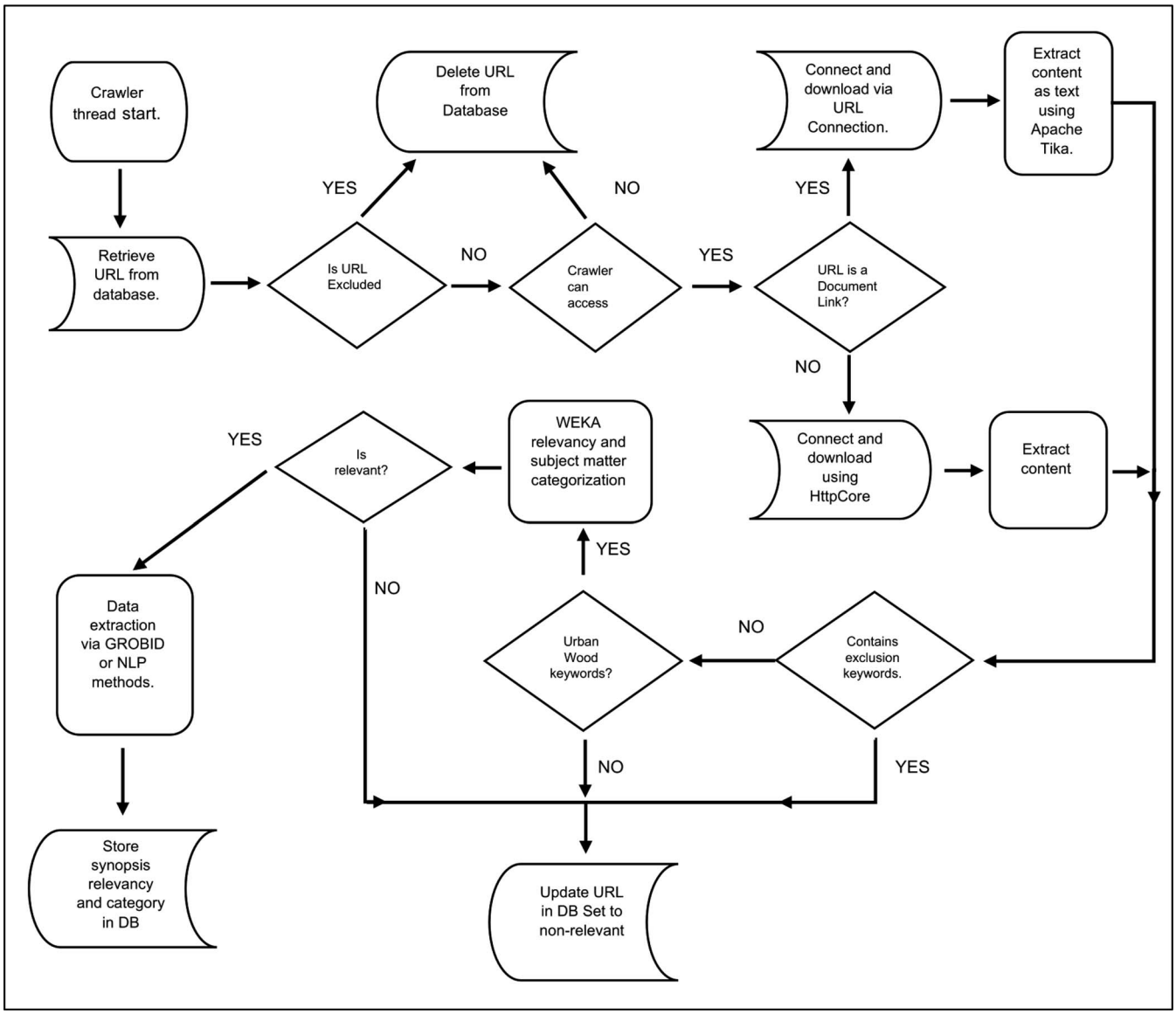

Figure 1.—Flowchart showing URL processing steps.

and authors. All files are deleted immediately after processing. If the text was extracted directly from a Web page, i.e., from html, then *GROBID* will normally not be able to find standard bibliographical items. For these instances we developed a natural language processing extractor that was based on methods created by Oumghar (2021). The approach resulted in an extractor that recognizes complete sentences and assembles them to form a brief synopsis.

Once processing is complete, the URL's data are updated in the MySQL database. References to the keywords found are stored as well as the links to other Web pages that were discovered at that URL. The URL is then queued in the administrator console for approval, and edited by a human operator if necessary. Once the administrator approves the URL, it will appear in users' search results on the urban wood knowledge Web page at https://urbanwooddatabase.umn.edu.

Results

The production UrbanWood KMS server and crawlers became operational and began searching for information on August 31, 2021. At the time of writing, the crawlers have

Table 2.—Subject area categories.

Urban forestry	Education	Invasive species
Fire risk management	Environment & sustainability	Health & safety
Urban wood utilization	Sawing & sawmills	Reclaimed wood
Log & lumber quality	Drying	Companies & products
Success stories & reports	Management & marketing	Networks & associations
General	Other	

Table 3.—Counts of resources discovered by primary subject category.

Category	Count	Percent
Urban forestry	258	36.4
Reclaimed wood	81	11.4
Networks and associations	73	10.3
Education	66	9.3
Urban wood utilization	59	8.3
Fire risk management	49	6.9
Health and safety	35	4.9
Environment and sustainability	33	4.7
Invasive species	24	3.4
Companies and products	20	2.8
Other	6	0.8
Management and market	1	0.1
Log and lumber quality and grading	1	0.1
Sawing and sawmills	1	0.1
Success stories and reports	1	0.1
Total	708	

explored >577,900 links and indexed another 154,500 links to explore. However, these lists are summarized and manually examined quarterly to make sure that the crawlers are focusing on the topic, and URLs are excluded and pertinent keywords are added to the database. On average, a single crawler process can fully explore approximately 2,000 links/d. In the first 30 days, 362 relevant links were processed; 99 more relevant links were processed in the next 30 days. Excluding the first 60 days, and a 6-month period, when the crawlers were stopped for revision, the average number of relevant links found per day is approximately 1.6. However, whenever the crawler discovers new repositories (sites with numerous sources about urban wood), as occurred once, it indexed 33 new publications in one day. A key benefit of the crawler approach to find data and information related to urban wood is that the search process is continuous, as opposed to a human researcher who would need breaks and would likely become bored.

Table 3 lists the number of resources found in the primary subject areas created for the urban wood knowledge database. The subject area classification of resources is performed by an artificially intelligent, Weka-based classifier (Frank et al. 2016) that assigns the category based on the presence and combination of multiple keywords. The most

Table 4.—Counts of resources discovered by secondary subject category.

Secondary category	Count	Percent
Urban forestry	113	16.0
Environment and sustainability	71	10.0
Companies and products	66	9.3
Education	56	7.9
Urban wood utilization	41	5.8
Networks and associations	41	5.8
Health and safety	25	3.5
Management and market	16	2.3
Success stories and reports	14	2.0
Reclaimed wood	9	1.3
Other	8	1.1
Invasive species	8	1.1
Management and marketing	2	0.3
Sawing and sawmills	1	0.1

Table 5.—Counts of resources discovered by type.

Link type	Count	Percent
Web page	353	49.9
Journal article	105	14.8
Report	103	14.5
Unclassified/Other	48	6.8
Magazine/Newspaper article	36	5.1
Seminars/Webinars/Conferences	27	3.8
Presentation	11	1.6
Brochure/Product sheet	11	1.6
Book/Book section	8	1.1
Conference paper	3	0.4
Standard	2	0.3
Thesis/Dissertation	1	0.1

common, 36.4 percent of all links, are general information articles about urban forestry. Reclaimed wood is the second most common (11.4%), and is closely followed by links about urban wood networks and associations (10.3%). Table 4 lists the number of resources that were assigned to each of the secondary subject areas. Of the 708 total links, only 471, or 66.5 percent were assigned a secondary category. General information about urban forestry was the most commonly assigned secondary category, 16 percent of all links. Thus, 50.4 percent of all links discovered contain general urban forestry information.

Table 5 shows the counts of the different types of publications currently contained in the urban wood knowledge database. Web page articles are the most numerous (49.9%)

Table 6.—Top 30 most commonly found keywords on urban wood-relevant URLs.

Keyword	Count	Percent
URBAN FOREST	494	69.8
PUBLIC	446	63.0
TREE	429	60.6
FOREST	425	60.0
CITY	421	59.5
STATE	316	44.6
URBAN FORESTRY	303	42.8
URBAN TREE	296	41.8
PROGRAM	270	38.1
HEALTH	260	36.7
CITIES	259	36.6
WOOD	243	34.3
LOG	230	32.5
QUALITY	228	32.2
BENEFITS	183	25.8
STAIN	182	25.7
FOREST SERVICE	176	24.9
PROCESS	167	23.6
CANOPY	165	23.3
PARK	161	22.7
COST	144	20.3
REMOVAL	136	19.2
BUILDING	135	19.1
RECLAIMED	131	18.5
COMMUNITY FORESTRY	125	17.7
URBAN WOOD	125	17.7
DEVELOPMENT	123	17.4
ASSESSMENT	121	17.1
EDUCATION	121	17.1
PLANNING	121	17.1

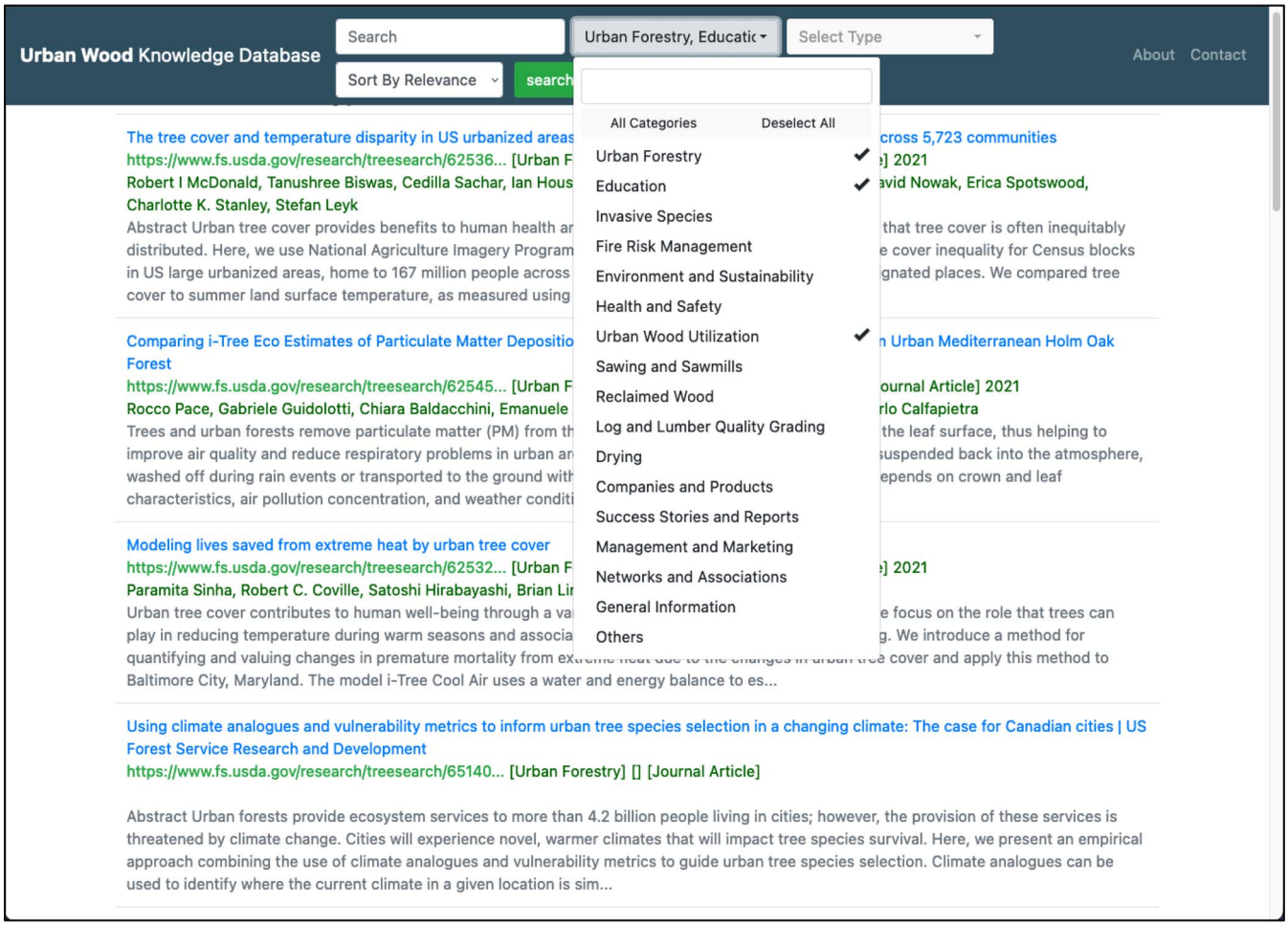

Figure 2.—UrbanWood KMS search engine screen shot showing search category and keyword.

publication type indexed in the system. In addition, there are 105 (14.8%) journal articles and 103 (14.5%) general reports in the knowledge database. There are 74 links (10.5%) in the knowledge base related to providing information and data to the public via magazines, newspapers, seminars, Webinars, conferences, and presentations, indicating the importance of this topic to the general public. Overall, these data are a good indicator of the types of quality information the knowledge database has been able to locate, process, and classify for its user base.

The top 30 most commonly encountered keywords found in links relevant to urban wood and their frequency are shown in Table 6. "URBAN FOREST" is the most common keyword set appearing in 494, or 69.8 percent of all relevant links. The closely related keyword sets "URBAN FORESTRY," "URBAN TREE," "COMMUNITY FORESTRY," and "URBAN WOOD" occurred in 42.8 percent, 41.8 percent, 17.7 percent, and 17.7 percent of all links, respectively. The top 5 keywords are each found in ≥60 percent of all relevant links. Additionally, 9 of the 22 core keywords used to identify information related to urban wood are in the top 30 list of keyword occurrences. After the top 10 keywords, the occurrence percentage drops to 36.6 percent and 17.1 percent for the 30th-most-common occurring keyword. These keywords identify papers in several of the categories listed in Table 1. Overall, occurrences of the remaining keywords are decreased in relation to the information categories with which they are associated (Tables 3 and 4).

Conclusions

There are few published examples of crawler and data aggregation -based knowledge database systems being developed for the collection of knowledge related to agricultural, forestry, or wood products, especially knowledge for scientific use. However, the ability to collect and classify thousands of documents focused on a narrow subject area greatly simplifies literature searches. There are several distinct benefits to the use of specialized crawler and data aggregation -based knowledge database systems, among them the following:

- Searches for literature are based on deep Internet searches involving hundreds of keywords.
- Web documents are classified by subject area and publication type.
- Enhances awareness of the subject area.

However, a key benefit to using the urban wood knowledge base system is that the data search can be limited to a single category, multiple categories, or all categories (Fig. 2). This allows users to quickly find specific information that intersects specific categories. The knowledge base system determines the data that intersect the specified keyword and

specified categories, and presents the results to the user. By default, the knowledge base searches a set of papers that can match one of many different means of referring to urban wood while excluding off-topic data.

Traditional marketing studies use surveys to identify important trends and elements with respect to a specific subject. Our approach using Web crawlers to discover data and published papers regarding all aspects of urban wood approaches this from a different perspective. A Web crawler depends upon the information being published on the Web, and is therefore subject to a delay between the conception of an idea and when the results are published. Surveys depend upon someone responding to them in an intelligible manner, but can capture those trends before they are published somewhere. The crawler captures the data postpublication, so the trends and papers it reveals have been validated and established.

The crawler and data aggregation -based knowledge database system presented in this manuscript focused on providing an easy-to-use resource relating information regarding all aspects of urban wood. Urban wood utilization offers potential for job creation and to use underutilized timber for a high-value-added product. The industry is in the early stages of market adoption, so it is important for stakeholders such as researchers, design professionals, manufacturers, developers, government agencies, and public in general to be able to access, review, and share knowledge about the state of research and implementation of urban wood utilization in the United States and the world. The creation of an easily accessible, free knowledge database may foster collaboration between parties and prevent duplication of efforts.

Literature Cited

Albani, S., M. Lazzarini, M. Koubarakis, E. Taniskidou, G. Papadakis, V. Karkaletsis, and G. Giannakopoulos. 2016. A pilot for big data exploration in the space and security domain. *In:* Proceedings of 2016 Conference on Big Data from Space (BiDS'16), P. Soille and P. G. Marchetti (Eds.). Publications Office of the European Union. 4 pp. http://publications.jrc.ec.europa.eu/repository/handle/JRC100655. DOI: 10.2788/854791. Accessed October 31, 2023.

Apache Software Foundation. 2022a. Apache HttpCore Overview. https://hc.apache.org/httpcomponents-core-4.4.x/index.html/. Accessed July 27, 2022.

Apache Software Foundation. 2022b. Apache Tika—A Content Analysis Toolkit. Apache Software Foundation. https://tika.apache.org/. Accessed July 27, 2022.

de Kunder, M. 2023. Daily estimated size of the World Wide Web. https://www.worldwidewebsize.com/. Accessed April 18, 2023.

Frank, E., M. A. Hall, and I. H. Witten. 2016. The WEKA workbench. Online appendix for: Data Mining: Practical Machine Learning Tools and Techniques. Fourth ed. 2016. Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf. Accessed Nov 1, 2023.

GROBID. 2022. GROBID: GeneRation Of BIbliographic Data. https://github.com/kermitt2/grobid. Accessed July 27, 2022.

Heaton, J. 2002. Programming Spiders, Bots, and Aggregators in Java. Sybex Publishing, San Francisco, California. 516 pp.

Herdon, M., C. Burriel, J. Tamás, L. Várallyai, P. Lengyel, and J. Pancsira. 2014. AgroFE—Collaborative environment and building learning knowledge base for agro–forestry trainings. World Conference on Computers in Agriculture and Natural Resources, University of Costa Rica, San Jose Costa Rica, July 27–30, 2014. Sponsor or publisher name, city, country. # pp. http://CIGRProceedings.org. Accessed Month day, year.

Huss, N. 2023. How many websites are there in the world? https://siteefy.com/how-many-websites-are-there/. Accessed April 18, 2023.

Medelyan, A. 2018. Python implementation of the Rapid Automatic Keyword Extraction (RAKE) algorithm. https://github.com/zelandiya/RAKE-tutorial. Accessed July 28, 2022.

OpenAI. 2023. Pioneering Research. OpenAI, San Francisco, California. https://openai.com/research/overview. Accessed September 14, 2023.

Oracle Corp. 2023a. MySQL Enterprise Edition. Oracle Corporation. https://www.mysql.com/products/enterprise/. Accessed April 18, 2023.

Oracle Corp. 2023b. Oracle Technology Network for Java Developers. Oracle Corporation. https://docs.oracle.com/en/java/javase/18/index.html. Accessed April 18, 2023.

Oumghar, K. 2021. Text-Summarizer. https://github.com/karimo94/Text-Summarizer. Accessed July 27, 2022.

Pitti, A., O. Espinoza, and R. Smith. 2019. Marketing practices in the urban and reclaimed wood industries. *Bioprod. Bus*. 4(2):15–26. https://doi.org/10.22382/bpb-2019-002

Rose, S., D. Engel, N. Cramer, and W. Cowley. 2010. Automatic keyword extraction from individual documents. *In:* Text Mining: Theory and Applications. M. W. Berry and J. Kogan (Eds.). John Wiley & Sons. https://doi.org/10.1002/9780470689646.ch1

Thomas, E., O. Espinoza, R. Bora, and U. Buehlmann. 2020. A specialized data crawler for cross-laminated timber information resources. *Forest Prod. J.* 70(3):256–261.