A Specialized Data Crawler for Cross-Laminated Timber Information Resources

Ed Thomas
Omar Espinoza
Rahul Bora
Urs Buehlmann

Abstract

The Internet is composed of more than 6.2 billion Web pages and grows larger every day. As the number of links and specialty subject areas grows, 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 key words and use machine learning to determine if what is found is relevant. In this article, we examine the design and data return of such a specialty knowledge database and crawler system developed to find information related to cross-laminated timber (CLT). Our search engine uses intelligent software to locate and update pertinent references related to CLT as well as to categorize information with respect to common application and interest areas. At the time of this publication, the CLT knowledge database has cataloged nearly 3,000 publications regarding various aspects of CLT.

Cross-laminated timber (CLT) is a relatively new structural building system based on the use of large multilayered panels made from solid dimensional lumber that is glued together with alternating layers perpendicular to each other (Laguarda-Mallo and Espinoza 2015). These panels are made to specification in CLT factories and then transported to the construction site, where they are put into place with cranes. Walls and floor systems are joined using metal connectors. Additional insulation layers can be applied to CLT walls and ceilings or left bare to take advantage of the warmth and aesthetics of wood. CLT's attractiveness as a building system originates in part from the speed at which CLT buildings can be raised, with considerable savings in labor and minimal disturbance to the site's surroundings (Crespell and Gagnon 2011).

CLT is a fledgling industry facing challenges such as a low level of awareness in the construction industry and compliance issues with existing building codes. To overcome such challenges, the number of research projects about the design and performance of CLT has increased dramatically during the past few years, as has the number of people involved in CLT research and implementation. For the success of CLT, as a product in the early stages of market adoption, it is important for stakeholders (i.e., researchers, manufacturers, construction professionals, developers, government agencies, and the public in general) to be able to *access* and *share* knowledge about the state of research and implementation of CLT in the United States and the world.

From our explorations into the subject of CLT, we have found 27 different terms used to refer to CLT products, including ''cross-laminated timber,'' ''cross-laminated pan-

The authors are, respectively, Research Computer Scientist, USDA Forest Serv., Princeton, West Virginia (ed.thomas2@usda.gov [corresponding author]); Associate Professor, Bioproducts and Biosystems Engineering Dept., Univ. of Minnesota, St. Paul (espinoza@umn.edu); Software Engineer, Foundation Engineering, Dell Technol., Unstructured Data, Seattle, Washington (rahul.bora@dell.com); and Professor, Dept. of Sustainable Biomaterials, Virginia Tech Univ., Blacksburg (buehlmann@gmail.com). The use of trademarked corporate names, products, and services is used only for the convenience of the reader and is not an endorsement by the University of Minnesota or the USDA Forest Service over other products and services that may be suitable. This paper was received for publication in April 2020. Article no. 20-00017.
©Forest Products Society 2020.
Forest Prod. J. 70(3):256–261.
doi:10.13073/FPJ-D-20-00017

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26

el,'' ''mass timber,'' ''massive timber,'' ''Kreuz Lagen Holz,'' ''KLH,'' and ''KLT,'' among others. Given the newness of CLT, the multiple ways of referencing it, and the breadth of subject areas, it can be difficult to develop effective queries. This is especially so when one considers the vastness of the data posted on the Internet.

At the time of this writing, the Internet is composed of approximately 6.2 billion Web pages (de Kunder 2020), and its growth shows no signs of slowing. Hence, finding pertinent data on the Internet is a task that continues to become more complex and time consuming. When the information topic is sufficiently complex or spans several languages or if there are many ways to refer to a specific topic, searches become even more difficult. In addition, related key words can change the meaning of the topic being researched, imparting a more specific meaning, which can be desirable. For example, adding the key words ''economics'' and ''efficiency'' to a search for ''CLT'' greatly narrows a search for information regarding CLT. However, attempting to manually perform a series of Web searches using the combinations of potential key words and acronyms while excluding other key words and then trying to collate all the information from the various searches is time consuming and frustrating. If the data are temporal and require collection at regular intervals, then the need for an automated approach is even more pressing.

Special-purpose data aggregators, often referred to as crawlers or spiders, are specialized software programs that browse information on the Internet and catalog, store, 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 catalog their findings in their proprietary databases to be called up when a user searches for a specific key word. 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 actually available on the Internet. However, as pointed out before, while such general-purpose search engines are good at finding specific information for a narrow topic (such as looking up the meaning of the word ''CLT''), collating all the available information on a specialized topic (such as looking up all relevant information on ''earthquake performance of CLT structures'') is difficult and time consuming. To address this need, specialized data aggregators have been 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 data aggregators collecting agriculture and forestry–related information exist. One recent example is AgroFE, a European Union project to develop an agroforestry 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), which targets a wide range of information areas, including 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.

This article 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 CLT (also referred to as mass timber). The information found, summarized, and categorized by the crawler can be accessed at https://masstimberdatabase.umn.edu. As such, this article is not a step-by-step guide to developing a Web crawler and having it search for the required content, as 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 here can serve as a guide to collecting data regarding other forestry and agricultural subject areas.

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 CLT with minimal human oversight. The most important requirement was that the crawler have the ability to search Web sites and documents for potentially hundreds of key words, categorize the information found, and store those links in a knowledge database for easy and fast retrieval later. Hence, the crawler was tasked to search not only Web pages but also all kinds of documents referred to on these Web sites. Such documents are posted in countless formats, including but not restricted to Portable Document Format (PDF), text documents, presentations, spreadsheets, and others, requiring the crawler to be able to process a wide array of file formats. Each source needs to be searched for the key words, all relevant documents need to be stored in the knowledge database, and, most important, the system needs to find and show the relevant knowledge on the users entering a search term in an easy-to-use interface.

To ensure 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. Just as important 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, documents that are added to the system's knowledge database can be manually checked by an administrator before they are officially added to the knowledge database.

The CLT knowledge database uses the MySQL Enterprise database system (Oracle Corporation 2019a) to store information. MySQL is a high-performance, robust, secure, and reliable database management system. These key features make MySQL well suited for the CLT knowledge database project. The knowledge database stores the Uniform Resource Locator (URL), a brief synopsis or abstract of the knowledge found, page title, author, and key words. 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, permitting the database to determine how many different URLs refer to any specific page. The referral count data combined with the number of key words found on a site are key components for the quality ranking of Web pages.

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26

The CLT knowledge database system was developed using the Java programming language (Oracle Corporation 2019b) 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 their incorporation greatly reduced development time. The software libraries that were most important to the successful development of the crawler system were the following:

1. *jsoup*: A Java library for working with real-world HTML that provides an application programming interface (API) for extracting and manipulating data (Hedley 2019).
2. *Apache Tika*: The Apache Tika (Apache Software Foundation 2019) tool kit, which detects and extracts metadata and text from more than 1,000 different file types.
3. *Aylien*: Information retrieval, machine learning, and natural language APIs for text analysis and extraction. Aylien is used to generate summaries to enable easy information abstraction and category assessments (Aylien Ltd 2019).
4. *GROBID*: A machine learning software system for extracting bibliographical information (title, author, abstract, citation, year, etc.) from scholarly documents (GROBID 2019).
5. *weka*: A collection of machine learning algorithms for data analytics (classification) and predictive modeling (Eibe et al. 2016).

The selection of key words that the crawler searches for was critically important, as the set of key words directly impacts the information resources that the crawler finds. In the CLT knowledge database project, the crawlers search for 267 key words related to CLT. These key words have been assembled by the research team and 21 outside experts (members of the intended audience for the knowledge database, such as architects, structural engineers, developers, wood scientists, government officials, educators, and researchers) based on their professional expertise. The key words were then adapted on the basis of the feedback from the system over time. Currently, the core of the key words is made up of 27 different key words commonly used for CLT products, including ''cross-laminated timber,'' ''cross-laminated panel,'' ''mass timber,'' ''massive timber,'' ''Kreuz Lagen Holz,'' ''KLH,'' and ''KLT,'' among others. The remaining key words were divided into subject areas and tree species.

The key word ''CLT'' itself is problematic. For example, CLT is the abbreviation used by the Federal Aviation Administration (2020) for Charlotte-Douglas International Airport as well as the abbreviation for the central limit theorem. To avoid including Web sites devoted to these and other off-topic Web pages, we used a list of exclusionary key words. Other exclusionary key words, such as ''medication'' and ''pharmacy,'' target off-topic sites. In addition, to avoid crawling some sites entirely, a set of excluded URLs is maintained. The key word and site exclusions allow the crawlers to stay on topic and avoid wasting resources on sites and Web pages that will yield no pertinent data.

Figure 1 shows 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 to 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 is extracted using the *jSoup* class library.

The text document from the URL (either the Web site content or the content of the file downloaded) is first searched for exclusionary key words that could indicate that the link is off topic. If off-topic key words 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 CLT-specific key words. To improve selectivity, we required that multiple occurrences of product name key words (such as ''CLT,'' ''cross-laminated,'' ''KLH,'' etc.) as well as a subject area key word (such as ''seismic,'' ''fire,'' ''market,'' ''safety,'' etc.) 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 CLT knowledge database will find most valuable.

If the text from the URL meets the key word requirements, then it is processed using two *weka*-based classifiers (Eibe et al. 2016). The first classifier determines if the text is relevant to the overall CLT subject area and returns a Boolean value, true or false. The second *weka* classifier determines the subject category of the text and returns one value from the possible set, including ''fire,'' ''seismic,'' ''raw materials,'' ''moisture,'' ''markets,'' ''vibration,'' ''connectors,'' and others. When the text is from a downloaded file (any form of PDF, Microsoft Word, OpenOffice, etc.), the text is processed using *GROBID* (GROBID 2019) to extract the title, abstract, key word list, 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. In these cases, the *Aylien* programming library and server (Aylien Ltd 2019) is used to create a brief abstract or summary of the Web page.

Once processing is complete, the URL's data are updated in the MySQL database. References to the key words found are also stored, as are 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 CLT knowledge page at https://masstimberdatabase.umn.edu.

Results

The crawlers began their search in February 2018 with a list of 400 ''seed'' links to use as starting points. These seed links were obtained using current general-purpose search engines. A search was conducted for eight subject category key words (''adhesive,'' ''design,'' ''economic,'' ''environment,'' ''fire,'' ''moisture,'' ''raw material,'' and ''seismic'') combined with a CLT product key word string (''CLT cross-laminated mass timber KLH panel''). For each search, the first 50 links returned by the search engine were used to seed the crawler. At the time of this publication, the crawlers have explored more than 7 million links and indexed another 20 million to explore. However, these lists are summarized and manually examined quarterly to make sure that the crawlers are staying on topic, and excluded URLs and key words are added to the database.

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26

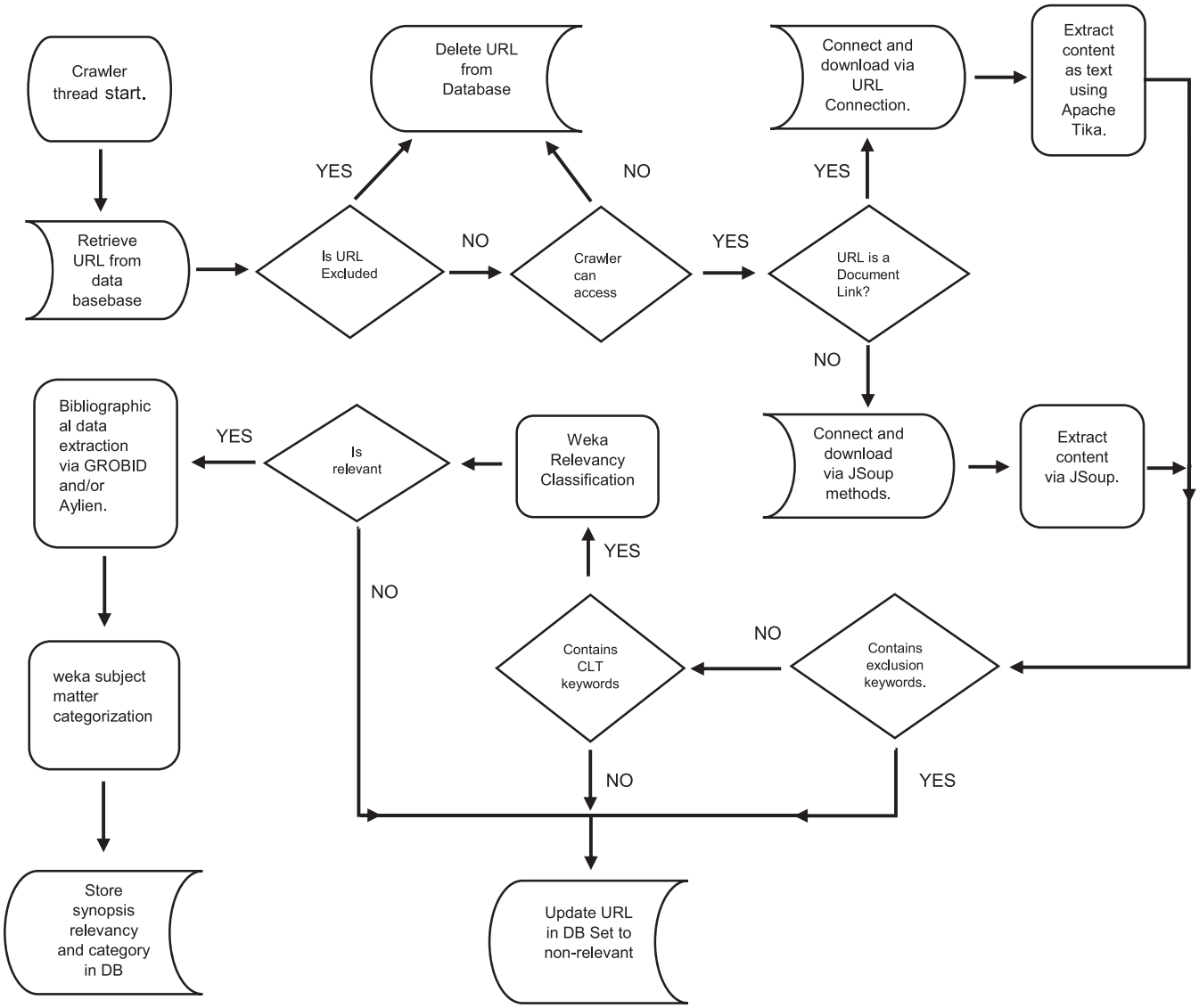

Figure 1.—Flowchart showing Uniform Resource Locator (URL) processing steps. DB = database; CLT = cross-laminated timber.

On average, a single crawler process can fully explore approximately 10,000 links a day. Given that three crawler processes are normally running, the system explores about 30,000 links a day. In the first 60 days, the crawlers discovered 1,264 links that contained relevant information that is currently included in the knowledge database. After these first 60 days, the average number of relevant publications found has been around three per day. However, the crawler may discover new repositories (sites with numerous sources about CLT), as it did once, thus indexing 57 new publications in 1 day.

Table 1 lists the number of publications found in the subject areas created for the CLT knowledge database. The most common publications are general information articles. Fire performance is the second most common and is closely followed by publications about CLT case studies. While publications regarding the economics of CLT seem to be the least numerous, one should consider that economics is often a key component of case studies. Thus, the ''economics'' category consists of papers that concentrate only on the economic aspects of CLT, such as building costs, energy consumption, or economy of design, and are not based on a single case study.

Table 2 shows the counts of the different types of publications currently contained in the CLT knowledge database. Journal article pages are the most numerous publication type indexed in the system. In addition, there is a total of 762 technical reports, conference papers, research theses, and dissertations in the knowledge database. This is a strong indicator of the types of quality publications that the knowledge database has been able to locate, process, and classify for its user base.

The top 25 most commonly encountered key words found in CLT-relevant publications and their frequency are shown in Table 3. Notice that CLT is represented by the key word set of ''CLT,'' ''cross laminated timber,'' and ''cross-laminated timber,'' thus illustrating the potential need to search for multiple spellings of the main key word (e.g., ''CLT'') to find relevant publications using a general-purpose search engine. The next most common key words—''construction,'' ''design,'' and ''structural''—are related to construction aspects of CLT. These key words identify

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26

Table 1.—Counts of resources discovered by subject category.

Category	No. (%)
General information	297 (10.65)
Fire performance	265 (9.50)
Case study/project	249 (8.92)
Mechanical performance	234 (8.39)
Market	230 (8.24)
Seismic performance	223 (7.99)
Connectors	184 (6.59)
Commercial/company information	139 (4.98)
Design/architectural aspects	119 (4.27)
Tall buildings	108 (3.87)
Environmental performance	107 (3.84)
Raw materials	108 (3.87)
Standards/building code	96 (3.44)
Vibration/acoustic performance	94 (3.37)
Moisture/durability	87 (3.12)
Alternative/hybrid configurations	75 (2.69)
Bonding/adhesives	27 (0.97)
Economics/costs	26 (0.93)
Other	122 (4.37)
Total	2,790 (100.00)

papers in several of the categories listed in Table 1. Overall, occurrences of the remaining key words are somewhat evenly distributed (Table 3).

Conclusions

There are few published examples of crawler and data aggregation–based knowledge database systems being developed for the collection of agricultural-, forestry-, or wood products–related knowledge, 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 key words.
- Web documents are classified by subject area and publication type.
- Awareness of the subject area is enhanced.

However, a key benefit to using the CLT knowledge database system is that the data search can be limited to a single category, multiple categories, or all categories. This allows users to quickly find specific information that intersects specific categories. For example, suppose a user was interested in the engineering aspects of CLT as they related to economics and case studies. One could easily construct a query for this by selecting the ''economics'' and ''case study'' categories and then key word searching for ''engineering.'' The CLT knowledge database system determines the data that intersect the specified key word and specified categories and presents the results to the user. By default, the knowledge database searches a set of papers that can match one of 27 different ways of referring to CLT while excluding off-topic data.

The crawler and data aggregation–based knowledge database system presented in this article focused on providing an easy-to-use resource relating knowledge regarding all aspects of CLT. For CLT to become a mainstream structural system, several barriers must be overcome, including low awareness by design/building professionals and the difficulty of building code compliance. Efforts by numerous individuals and organizations are aimed at generating design/testing data to accelerate CLT adoption by the construction industry. The crawler and data aggregation–based knowledge database system fosters awareness of the potential of CLT, leading to the growth of CLT use by disseminating knowledge and facilitating collaboration among stakeholders, while reducing the risk of duplication of efforts. By improving the availability of information related to CLT, we believe that manufactures and their suppliers, researchers, design professionals, code officials, government agencies, and other stakeholders can directly benefit from the tool presented here, thereby supporting the increased use of CLT as a timely construction material.

Table 2.—Counts of resources discovered by type.

Resource type	No. (%)
Journal article	739 (26.49)
Web page	681 (24.41)
Report	338 (12.11)
Conference paper	299 (10.72)
Magazine/newspaper article	249 (8.92)
Presentation	129 (4.62)
Thesis/dissertation	125 (4.48)
Brochure/product sheet	57 (2.04)
Book/book section	42 (1.51)
Standard	11 (0.39)
Others	120 (4.30)
Total	2,790 (100.00)

Table 3.—Top 25 most common key words found on cross-laminated timber (CLT)–relevant Uniform Resource Locators.

Key word	No. (%)
CLT	2,334 (83.66)
Construction	2,031 (72.80)
Design	2,010 (72.04)
Structural	1,748 (62.65)
Load	1,663 (59.61)
Buildings	1,660 (59.50)
Research	1,549 (55.52)
Cross-laminated timber	1,479 (53.01)
Fire	1,378 (49.39)
Architect	1,268 (45.45)
Strength	1,148 (41.15)
Projects	1,127 (40.39)
Properties	1,108 (39.71)
Innovation	1,096 (39.28)
Glue	1,094 (39.21)
Cross laminated timber	1,062 (38.06)
Cost	1,049 (37.60)
Shear	981 (35.16)
Connection	968 (34.70)
Seismic	948 (33.98)
Span	800 (28.67)
Thickness	792 (28.39)
Suit	781 (27.99)
Sustainability	755 (27.06)
Mass timber	738 (26.45)

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26

Acknowledgment

The work on which this article is based was funded in whole or in part through a grant awarded by the Wood Innovations Program, USDA Forest Service.

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.), March 15–17, 2016, Santa Cruz de Tenerife, Spain; Publications Office of the European Union, Brussels. pp. 196–199.

Apache Software Foundation. 2019. Apache Tika—A content analysis toolkit. https://tika.apache.org. Accessed March 27, 2020.

Aylien Ltd. 2019. Aylien Text Analysis API. https://aylien.com/text-api. Accessed March 27, 2020.

Crespell, P. and S. Gagnon. 2011. Cross-laminated timber: A primer [PowerPoint presentation]. FPInnovations, Vancouver, British Columbia, Canada.

de Kunder, M. 2019. The size of the World Wide Web (the Internet). https://www.worldwidewebsize.com. Accessed March 27, 2020.

Eibe, F., M. A. Hall, and I. H. Witten. 2016. The WEKA Workbench. Online Appendix for ''Data Mining: Practical Machine Learning Tools and Techniques.'' 4th ed. Morgan Kaufmann, Burlington, Massachusetts.

Federal Aviation Administration. 2020. Airport and facility codes. https://www.faa.gov/nextgen/cip/airport_facility. Accessed March 21, 2020.

GROBID. 2019. GROBID: GeneRation Of BIbliographic Data. https://github.com/kermitt2/grobid. Accessed July 17, 2019.

Hedley, J. 2019. jsoup: Java HTML Parser. https://github.com/jhy/jsoup. Accessed March 21, 2020.

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. Presented at World Conference on Computers in Agriculture and Natural Resources, July 27–30, 2014, University of Costa Rica, San Jose.

Laguarda-Mallo, M. F. and O. Espinoza. 2015. Awareness, perceptions and willingness to adopt cross-laminated timber in the United States. *J. Cleaner Prod*. 94:198–210.

Oracle Corporation. 2019a. MySQL Enterprise Edition. https://www.mysql.com/products/enterprise. Accessed March 21, 2020.

Oracle Corporation. 2019b. Oracle Technology Network for Java Developers. https://docs.oracle.com/en/java/javase/14. Accessed March 21, 2020.

http://prime-pdf-watermark.prime-prod.pubfactory.com/ | 2024-12-26