Trends and Impacts of Digital Library Research in last 20 years:

A case study on International Journal on Digital Libraries

INFO 5810 – Data Analysis and Knowledge Discovery Section 001

Project Submission – 2

Date – 11/01/2022

Group 2 Members:

1. Kaushal Sen - 11587863
2. Venkata Suryasatya Kakarla - 11606141
3. Karthik Rayapati - 11596242
4. Kundana Lanka - 11503082
5. Likith Mallipeddi - 11590166
6. Sandeep Prasad Owk – 11595377

Github Repository Link:
https://github.com/kaushal-sen/Project_5810_Section1_Group2

1. Introduction

Digital Library can be understood as a collection of information in digital formats which can be accessed using a network. Digital libraries came into picture in the early 1990s but have become popular as the readers are now more interested in reading articles and books on a screen rather than going to a physical library. This change of interest in readers is because of the easy accessibility and economically cheaper features that are offered by the digital form of the library. Nowadays everyone is preferring a digital library. A user may be able to access magazine articles, journals, books, papers, photos, music files, and videos, depending on the particular library. The utilization of a digital library on the Internet is increased by a broadband connection like a cable modem or DSL and also due to availability of internet access at the palms of each individual. Plain-text documents and some documents with images can be accessed using dial-up connections, but for complicated files and those with animated video content, a data connection with speed of at least several hundred kilobits per second (kbps) or Mbps can make the user's experience less tiresome and more educational. Digital libraries accessible over the internet can be updated every day.

To understand this transition from traditional library usage to digital libraries is the ultimate research goal by using the variety of works done by different authors in the last 2 decades. For this, we as a group will be doing a case study on International Journal on Digital Libraries (IJDL). The journal contains 23 volumes having 4 issues which in turn has 6-8 articles. This review research will be one of its kind where analysis of all the articles published till date will be considered for information extraction. The results will help identify articles in the IJDL for topic-based analysis and select relevant articles based on the required topics of interest. In this attempt, several methodologies, techniques which were previously published in this journal are analyzed based on their abstract and keywords. Summary extraction will also be performed on the abstract of each article to give a 1-2-line description of the article.

2. Literature Review

We have reviewed and provided analysis of 12 different works in the field of digital library. Each of the works reviewed was published in a recognized journal and have a decent impact factor. Below each of the articles are mentioned with short and crisp detail.

2.1. What Is Usability in the Context of the Digital Library and How Can It Be Measured?

In 2005, Judy Jeng [3] proposed several instruments and models for evaluating the usability of digital libraries to determine the satisfaction rate, learning rate, competence and productivity. Also, an uncertainty has always existed when it comes to determining the intent and utility of a digital library. Several authors previously argued about the limited amount of work done in evaluation in this context. The digital library is also referred to as an individual institution on its own with an infrastructure based on the communication network and computing power. Judy suggested the usability of the digital libraries in terms of usefulness (ability to fulfill specific purposes for intended user), effectiveness (ease to use), satisfaction (performance in long-term). For evaluation different approaches like surveys, user testing, traversal, usage analysis, error rates, metadata, focus group meetings and awareness parameters can be used.

The most common evaluation technique that was used is a questionnaire based survey done by National Taiwan University Library. Depending upon the target groups, the usability results varied like for the hospital based survey digital library turned out to be time consuming due to lack of technical exposure in using the interface for information exchange. For educational purposes, the results suggested that the digital collections meet the required standards for using in class. Inspection of the user interface by using Heuristic evaluation and card sorting techniques were useful in accessing the

structure made by the organization in implementing their library on the web. The formal usability testing technique to measure students' awareness of library resources was used by University of Pacific to test the learning ability of the students to traverse the web and locate books, articles using the references and also finding related missing information using the digital library resources.

Jeng proposed an evaluation model to find the satisfaction, effectiveness and efficiency along with the learnability attributes of digital libraries. According to the article, more work needs to be done in development of more efficient metrics for evaluation of usability and more parameters need to be taken in consideration for future research. This work was highlighting the usability of digital libraries and mentioning the approaches, techniques and efforts that have been used till date.

2.2. Advanced features of Digital library of University of Maribor

Ojsteršek et al [4] discussed the features of the digital library of University of Maribor. The authors highlighted the advanced capabilities like plagiarism detection using natural language processing, specific content extraction etc. The library allows the students to collaborate and share their thesis to a platform where it is publicly available to everyone. It is a web based application acting as a database for searching and knowledge sharing among students. The process of adding and viewing articles in the digital library is explained and it is implemented by mapping the authentication data of the student for login and followed by uploading of the document. Before the document is published a text based natural language processing algorithm checks for plagiarism. The digital library has a simple interface to provide ease of use and is effective for learning which are important factors of usability.

The authors also highlight that the digital library has informative and useful statistics published to show the number of documents and articles based on parameters like keywords, faculty, new articles etc. These statistics are useful in determining the quality of articles and also for choosing the right mentor based on keyword search done. With advantages, some of the cons of digital libraries come into picture. The issue of copying other contents has increased due to availability of free, open and relevant content due to digitalization. To avoid plagiarism, the digital library of University of Maribor makes use of TextProc-natural language based framework for detection using software plugins. The results will be in percentage depicting the amount of plagiarism. Also, the source of the plagiarism is also shown when the contents are copied from a document that is already present in the database. The teams are working on using web crawlers to map the documents available on the total web network to improve the efficiency of the system to detect plagiarism.

The authors were able to highlight the advanced features that are commonly not available on all the digital libraries and also they were able to answer a few of the usability parameters in terms of the user interface and plagiarism detector being an essential instrument for any online content management digital library.

2.3. Impact of New Technologies in the Digital Libraries

The interaction between libraries and information services and final users. The library staff must be better, more specialized, more technically proficient and capable of offering high-tech services. The reader should learn computer skills and build the abilities to take advantage of computerized maximum information services.

The concept of the library has transformed from one of print and paper media to one of digital media as a result of advancements in information technology and the evolution of digital libraries. The computers, communication abilities, and technological expertise of library workers are crucial to the success of a "digital library." In the modern period, we are in the process of moving from a localized physical library to a worldwide digital library. With the appropriate infrastructure, we may start the construction of local digital library resources for improved and new services, which is the logical

extension of technological advancement and the enthusiastic response of the library and information experts to difficulties. Regardless of the location of the computer where the material is stored, the digital library intends to offer access to it "on demand." The function of digital libraries in the modern information world, as well as the new opportunities and risks for library services, particularly in India. An electronic collection of physical or digital resources that may also be found elsewhere may be referred to as a digital library. These materials must be entire works that permit full cognitive and practical interaction from people. The elements that a digital library organizes and houses may be accessed online or offline, and they may comprise multimedia and data in multiple languages. A digital library is not a website or a portal even though it is available online; whereas portals.

2.4. Evaluating the impact of digital library database resources on the productivity of academic research

This research home data was examined through SPSS using the statistical approach of simple linear regression to determine the effect of common database resources on the academic research culture. Our HEC databases that were accessible by 52 universities served as our independent variables during the data analysis process, while articles on the Web of Sciences served as our dependent variables. Resources from digital library databases have a big role in encouraging the research culture in higher education. Understanding intellectual development, research productivity, planning, and the detection of user information demands are all made feasible by the usage of digital databases. This study's objective is to aid management in creating an effective user database resource utilization evaluation and research fantastic academic policy.

With the help of digital databases, this study develops a quantitative way to evaluate the effectiveness of scholarly research. the secondary information drawn from 52 university databases that was provided by Higher Education Commission (HEC) and the literature that was published on the ISI Web of Science. To evaluate the data and determine how the independent variable "digital databases" affected the dependent variable "research productivity," basic linear regression was applied in statistics.

This body of research serves as a helpful tool for managing gaps and fostering the development of necessary actions to build plans and solutions to improve the academic environment. The final application of common database resources can encourage higher academic study to provide original thoughts and enhance the cognitive skills of researchers.

2.5. Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries

According to the author, through quick, distributed access to computer resources, analytical tools, and digital libraries, e-Inquiry promises to quicken the pace of science. While "small scientific" study fields that gather data through meticulous fieldwork continue to handle their data locally, "big science" sectors like physics and astronomy that collaborate around expensive apparatus have built common digital libraries to manage their data and documentation. Researchers studying habitat ecology are having a variety of problems gathering, collecting, and managing vast amounts of data as they start to use embedded sensor networks. We consider how these conclusions might affect data regulation and the design of digital libraries. The Center for Embedded Networked Sensing is associated with the research presented here.

An integrated framework for data management must include automatic support for data description and annotation in order for the data to continue to be readily recognizable, discoverable, and accessible in a usable format. A digital library for scientific data will at the very least include the following tasks: I the definition of a common communication framework (like that offered by XML) for the sharing and exchanging of metadata, both within the immediate research community and with other communities; (ii) the definition of the semantics and syntax of a common metadata schema (i.e., a standard set of metadata elements). The need for better tools is stated in e-Science projects, but

nothing is said about the standards that should be used to create them. More knowledge is required regarding the procedures, attitudes, and motivations involved in the gathering, utilizing, and managing of scientific data. These results are crucial information for creating efficient digital library systems, services, and regulations.

It is assumed that these data will be utilized by other researchers in the creation of digital libraries for scientific data and the policies of funding organizations to encourage data deposit in those systems.

2.6. Digital Libraries: Challenges and Problems.

According to the author, a digital library is a simplified and organized electronic form of documents. It allows for retrieving information or data successfully from different sources. This leads to an increase in demand for information from desktop users. More information can be accessed with a single click on the computer through the internet. Digital libraries provide fast access to information and can solve massive problems by managing a lot of data. But compatibility problems arise in the software and hardware field. In addition to this, data is not secured on websites and social media.

A digital library is one that keeps information in digital form. The essential features of the digital library are the storage of material in digital form, the direct use of communication networks for accessing and getting information, and copying by download or online/offline printing from a master file, among others. The main benefits of the digital library include universal accessibility, access to more information, support for both formal and informal learning, remote access to expensive and rare materials, protection of expensive and rare books that are quickly deteriorating due to overuse or storage conditions, solving massive storage problems, prompt and faster access to information, the ability to manage very large amounts of data, and they also assist in performing searches that are not possible to perform manually.

Contrarily, issues and challenges with digital libraries include information accuracy, hardware/software compatibility, reliability of the information, IPR issues, data security, fair use, convenience of usage, needs for expensive technology, storage life/shelf life, the omnipresence of digital reading and storage devices, and the need for hardware and software. If efforts are made to address these issues and the challenges of the digital environment are properly met, there is little doubt that the digital library will be a great asset to both users and LIS professionals.

2.7. Personalization and Recommender Systems in Digital Libraries

As stated by the author, digital libraries are groups of information that have related services and are made available to user communities using a range of technologies. The information collections can be represented as digital text, image, audio, video, or other types of media and can include scientific, business, or personal data. The information can be born digital or digitized paper, and the services provided on it can range from content operations to rights management, and they can be made available to either individuals or user communities.

Personalization tools have been very successful outside of digital libraries. For instance, practically all web search engines display customized advertisements in their output pages when it comes to targeted advertising. When using online retail platforms like Amazon.com, we are recommended extra complimentary services and goods. The display layout of menu selections when using a WAP portable device is personalized and catered for different users. Personalization can be used to help a user navigate through expansive online systems like the web or closed collections like courseware, as shown by adaptive hypermedia systems.

The research summarized in this paper is the result of work by a working group formed to investigate the function and potential future of personalization and recommender systems in digital libraries. In this study, we present the findings of the working group, which was supported by the International

Digital Libraries Initiative of the National Science Foundation and the DELOS EU FP5 Network. The following describes the structure of this essay. We have included some background information on our definition of personalization and recommender systems in this section. The overview of our plan for the development of digital libraries and our viewpoint on personalization in such libraries are provided below.

2.8. World Digital Library: A Case Study

According to the author, prior to the development of digital storage technology and the widespread commercial availability of computers, the original vision of what has come to be known as digital libraries was conceived in July 1945.

Bush wrote the article "As we may believe" (1945). Bush proposed the idea of an automated file and library for all the information, correspondence, and books for everyone. J.C.R. Licklider was hired to write a monograph about the future of libraries in the early 1960s. To some extent, this was the first time the conceptual foundation for a digital library had been identified.

"We need to replace the book with a device that will make it simple to transmit information without having to transport material, and that will not only present information to people but also process it for them by following procedures they specify, apply, monitor, and, if necessary, revise and reapply. A library computer is obviously needed to provide those services.

There has been a lot of research done on digital preservation and how it affects preserving cultural assets. Klien (2002) provides details on the Philip S. Hench Walter Reed Yellow Fever Collection, describing it as a two-year Institute of Museums and Library Services (IMLS) grant-backed digital project that includes more than 5.000 papers chosen from the collection. The National Library of Australia has provided guidelines for the protection of digital heritage (2003).

On the digital canvas, WDL authentically replaces the original culture and legacy of the world. The reflections sparkle like the light in the morning and drench the bit and byte's culture and heritage. mode throughout the world. As part of its mandate to conserve materials that add to the collective heritage.

2.9. Recent trends in digital library publications: a scientometric analysis

The article seeks to find the new trends in digital libraries by the usage of the method of scientometric Analysis. In order to fit the needs of the students, librarians use technology on their side to provide the best service possible to develop and provide digital libraries. Sharing resources has become the most important and necessary need for today’s generation, which also gave rise to new ideas and innovation. Many of the current libraries use the “concierge” approach to personalize the delivery and usage of digital libraries. The main idea of the article is to provide and suggest the possible direction for the research of digital libraries.

The scope of the study is to find the trending research article indexed in “Scopus” which is one of the largest citation databases used for searching documents. For the analysis of the study, the number of publications compared from last year to current in the research period shows a strong uptrend in the publications. The study of the number of citations is also conducted, which again shows a strong up-trend nature. The number of people contributing to the articles is tabulated and the geographical locations of the articles are also tabulated to find the topmost countries with the highest number of publications. The most productive articles are found using the biblioshiny, the H-index has been found out and ranked accordingly. The most used keywords are identified and tabulated according to the most searched keywords.

In recent years, researchers found a useful way for using technologies like artificial intelligence, and

machine learning methods to promote their publications effectively. These technologies help with specialized services at any time and at any location the main purpose of digital libraries is to ease the efforts of the users, create library services and preserve the resources in the long term. More articles are being produced than ever before which emphasize the usage and the finding effective ways to optimize the use of technologies for maintaining and offering the services of digital libraries.

2.10. Digital libraries and reference services: present and future.

Current digital libraries spend more effort on information retrieval and access to resources, hardly any effort has been made in the service aspects. Referencing services are one of the important factors of information services, as they are considered as personalized services between the user and the referenced librarians. The less emphasis is mainly due to the narrower definitions of these personal services by digital library researchers. The study mainly focuses on these personalized aspects that are needed to be developed. The analysis of the current state of personalized information services in digital libraries is done, followed by the various online personalized services that are currently in use will be analyzed. And finally, the study proposes some new areas of research that may be undertaken to improve personalized information services in digital libraries.

There are various definitions of personalized information services in the digital library domain, Borgman says “the research focuses on the digital libraries as content collected on behalf of the user, while librarians focus on digital libraries as institutions or services”. The reference services that are available on the web are provided by non-library and commercial organizations. Some of them require payment and some of them are for free. Digital reference libraries that are available on the web are public libraries, which can be accessed by everyone. Digital references services for users of academic libraries and lastly digital references service by cooperative library systems.

Some recent research efforts are directed toward the building of a personalized digital libraries environment. These studies can be classified into three groups, user need assessments, personalized user interface, and personalized digital library. Some universities like North Carolina university library have proposed Mylibrary as a portal application for North Carolina university’s Library. It has two components, Mylink and Myupdate, Mylink is exclusively for the patron’s users for saving information and My update offers the periodic queries to the online catalog and notifies users, using the pre-defined user lists. This is one way to personalize the digital library by offering the users to customize their experience. Some of the other resources of personalized digital libraries are Collaborative Digital reference services (CDRS) automatic reference libraries for the world wide web, SIFTER, and Virtual Reference Desk.

The study re-emphasizes the need for the provision of personalized digital services for the next generation of digital libraries. The emphasis is on filtering mechanisms based on personal users. Technology for automatic information discovery, organizational and retrieval, cooperative models for information sharing and specializations-both content and staff. Further research should be carried out in the areas including testing e-commerce models for testing the payments and cost management aspects and training techniques for expert information problem-solving processes.

2.11. Understanding user requirements and preferences for a digital library Web portal

The research from a user study is reported in this article. To learn more about user needs and preferences for the digital library services provided by The European Library Web portal, research was done as part of the TELplus project. The studies performed revealed potential drivers behind participants' use and avoidance of technology the services offered, enabling better-informed decisions about how to customize, enhance, and deliver Web portal services to their future customers.

Understanding the various, changing demands and preferences of their target user groups and then connecting them to the architecture of digital library Web portals and services offers a significant challenge. The goal was to improve, determine and keep track of user requirements, interests, and preferences for the intended system.
To better tailor software applications to their target users' needs, preferences, and expectations, log data are generally collected to study the usage patterns of various software programs. Although the query log is a less comprehensive source of information for specific events, it contains crucial details for comprehending the range of resources a search engine must make available to users. In the studies they conducted, the analyzed and combined sets of data gave us a chance to gain a better understanding of not only the needs and preferences of the initially involved participants but also of other pertinent user communities that might be interested in using TEL services, as well as of their needs for more individualized offers of these services.

After grouping the most pertinent questions, the evaluation of the questionnaire was carried out. All Web portal and service designers should be interested in this outcome because it shows different mother tongues. Provided research on contrasting sorts of logs in large systems to discover new techniques by utilizing them to assess and customize digital library user services.
As a result of this investigation, we could also talk about the accessibility and utilization of log data. Should collect log data for certain operations, whether general log data has value, whether log files must be made publicly available to researchers, and how more information might be gathered and connected with query log data.

2.12. Teaching digital library concepts using digital library applications

There are several digital library applications, and these applications are employed in classroom settings to demonstrate the operation of digital libraries in addition to the design choices made in creating them. This study identifies and explores DL subjects that may teach most effectively via DL applications in the framework of project-based DL courses. Gaining practical experience utilizing, designing, analyzing, and managing DLs is a crucial part of this learning system. This study uses the function of DL applications as an example to examine topics that might be brought up in a course.
The subject areas that can teach via DL applications are crucial for developing DLs. Similarities and differences between the ingest functionality variations between different DL apps. At a high level, ingest fundamental functionality is the same for everyone. Applications for the DL: The DL receives a digital entity that the object is in some way prepped before being added. The tasks carried out by various DL apps to carry out equivalent duties and consider design choices made by application designers for DL.
The most crucial task a DL can perform with that object application supports without digitalization, there wouldn't be any material for DLs to hold, and without metadata, accessing the content in DLs would be challenging or impossible. Without a user community, it is merely a collection of bits, notwithstanding its skill. Digital content preservation is challenging, and numerous initiatives have been made to handle its many facets. Giving students exposure to and practice with open-source tools is a crucial component of a DL course.
Services for these stakeholders are among the essential components of a DL because they are integral to it. All software development companies, whether for-profit or charity, provide a way for users to get in touch with them for technical help. Teachers can use DL applications for teaching materials in project-based DL courses. Although sharing educational resources is crucial, it is merely one step in improving DL education. Teachers need more materials to use in their classrooms in addition to lesson ideas.

3. Methodology

For achieving the research goals, first a dataset containing the information of each of the articles published in the International Journal of Digital Libraries will be created. The dataset will be in ".xlsx" file format with columns namely author, year of publication, abstract, keywords, DOI information and related relevant attributes from the citations. For analysis of the articles, keywords obtained from the articles will be considered of most importance. This will help in understanding the transitions, development, and growth in the digital library domain over the decade.

Data Collection will be done by web scrapping of "Web of Science" portal for different article details. Following data collection, the next step would be cleaning of data that includes selection of specify attributes from the dataset for proceeding, replacing of missing values, inspection for duplicate entries in the dataset and finally export of a new dataset with fewer but crucial attributes.

For initial stages of exploratory data analysis variety of visualization has been created by taking different combination of the attributes from the cleaned dataset and presented using the pivot table feature of the MS-Excel. Around 15-20 visualization with data insights are planned to be provided by the completion of the project.

Classification of articles based on keywords will also allow the individuals to find articles based on their selected interests easily.

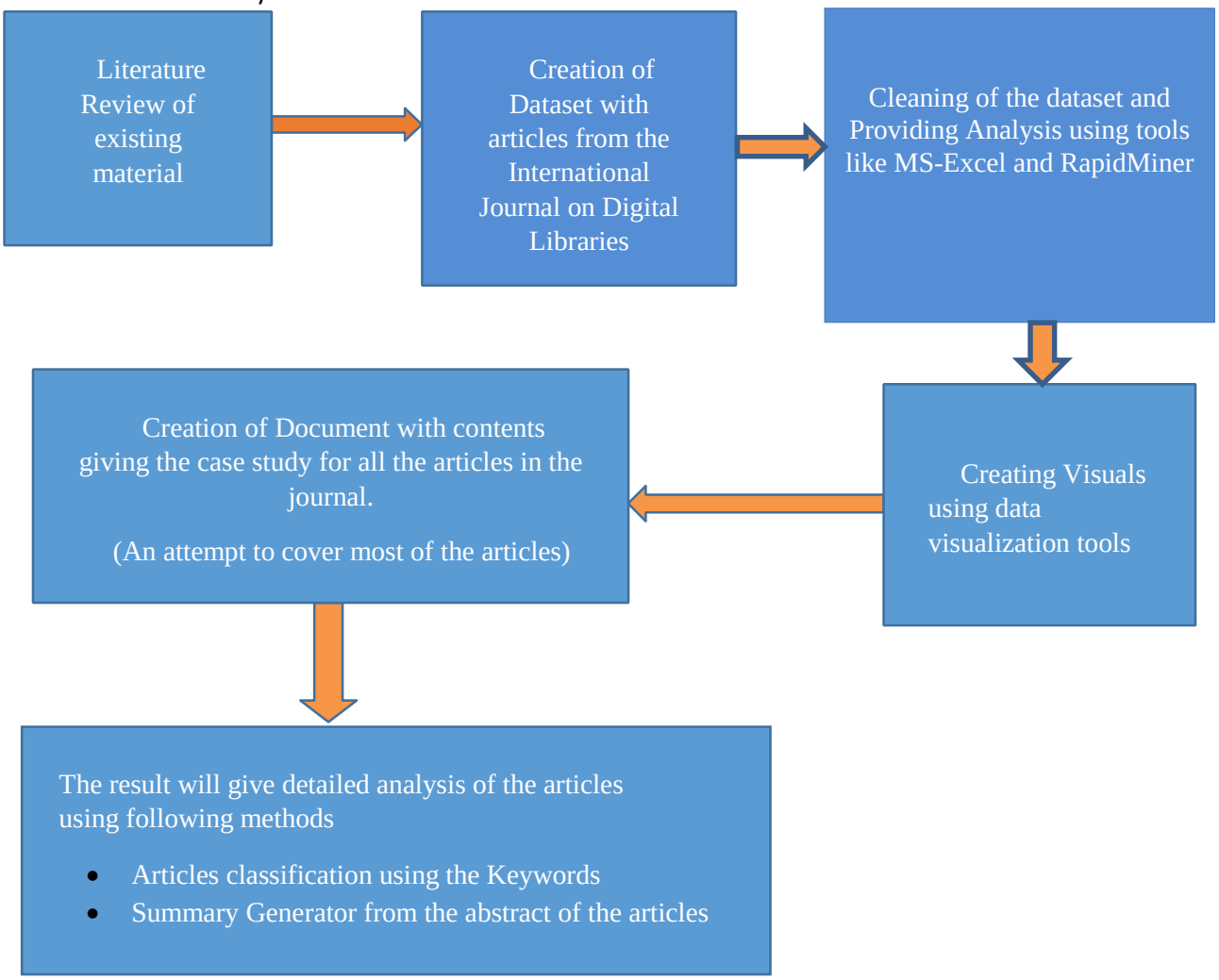


Figure 1-Pipeline for research activity and analysis task

4. Data collection and Cleaning

4.1. Data Collection

A dataset of scholarly articles from the journal named “International Journal on Digital Libraries'' has been used for this case study. The dataset contains details of the articles published in the last 20 years i.e. between 2002-2022. Total number of data records in the dataset equals 420. The data is collected by using the web scraping techniques for collecting all the details displayed on the webpage about the articles published in the journal. Also another source of the data is a database named “Web of Science'' from which the citation count of the articles is gathered. For some instances, “Crossref '' – a public digital object identifier platform is used to get the meta data of the articles using the DOI information.

The dataset “Article Details from IJDL.xlsx” containing details of the articles have 72 different attributes. The data types of most of the variables are nominal (text) and integers.

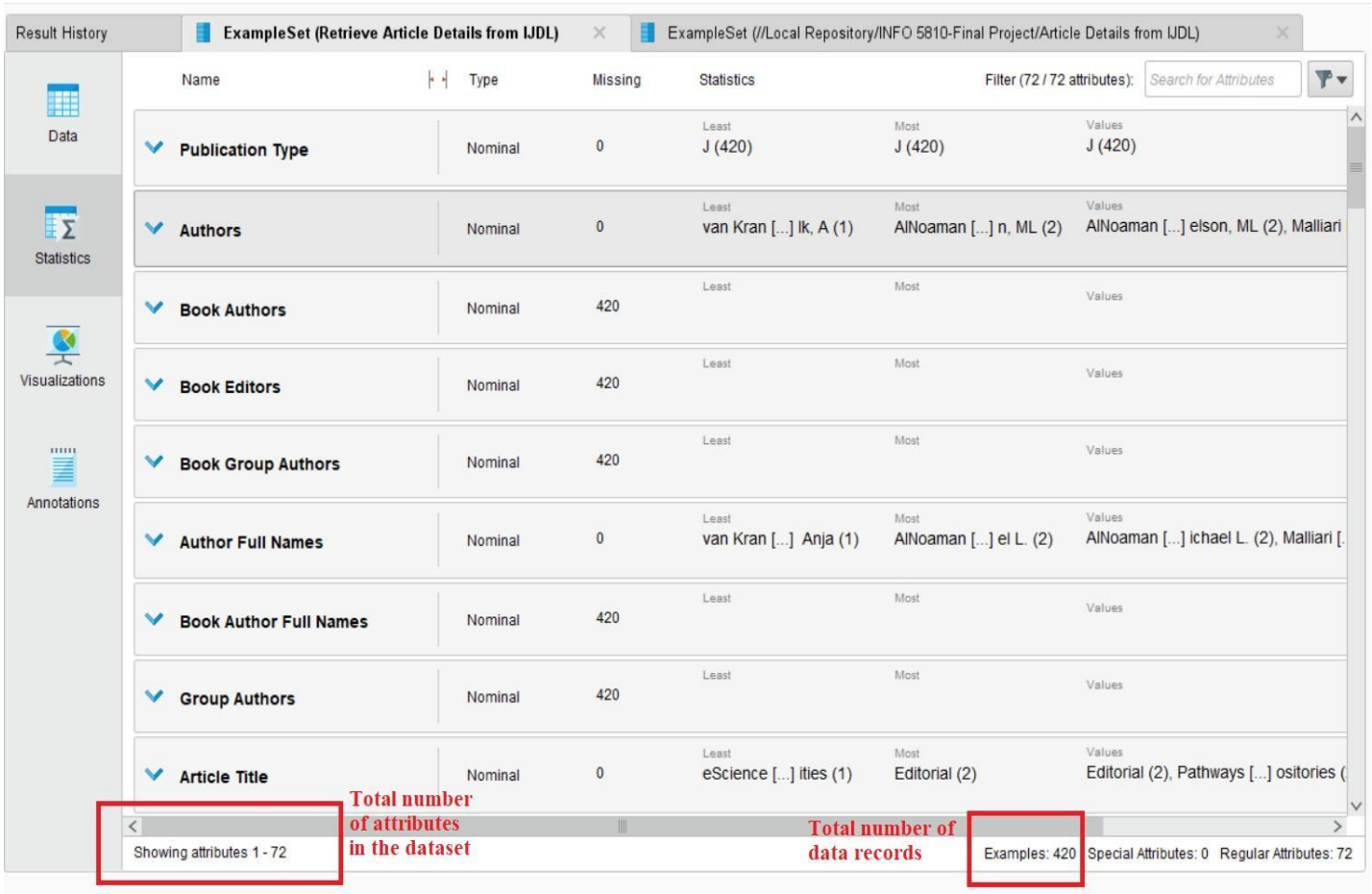


Figure 2 – RapidMiner Statistics Tab View for the Original Dataset Attributes count

Attributes List:

Publication Type	Publication Type	Authors	Book Authors	Book Editors	Book Group Authors
Author Full Names	Book Author Full Names	Group Authors	Article Title	Source Title	Book Series Title
Book Series Subtitle	Language	Document Type	Conference Title	Conference Date	Conference Location
Conference Sponsor	Conference Host	Author Keywords	Keywords Plus	Abstract	Addresses
Reprint Addresses	Affiliations	Email Addresses	Funding Text	Cited References	WoS Core
ORCIDsFunding Orgs	Funding Name Preferred	Cited Reference Count	Times Cited	Times Cited	All Databases
180 Day Usage Count	Count, Publisher	Since 2013 Usage	Publisher City	Publisher Address	ISSN
eISSN	ISBN	Journal Abbreviation	Journal ISO Abbreviation	Publication Date	Publication Year
Volume	Issue	Part Number	Supplement	Special Issue	Meeting Abstract
Start Page	End Page	Article Number	DOI	DOI Link	Book DOI
Early Access Date	Number of Pages	WoS Categories	Web of Science Index	Research Areas	IDS Number
Pubmed Id	Open Access Designations	Highly Cited Status	Hot Paper Status	Date of Export UT (Unique WOS ID)	Web of Science Record.

The dataset doesn't contain the full text of the articles. It has the entire abstract text, author specified keywords which can be used for dynamic text mining, or for designing some machine models to analyze the data. All the articles are published in English language so translation of the text is required.

4.2 Data Cleaning

The data collected from the articles published in the journal contains a lot of attributes which are empty or contains data that is monotonous. The attributes which are not needed for further analysis of the dataset will be removed from the dataset using the "Select Attributes" operator available in RapidMiner Studio. Out of the 72 attributes in the original dataset, 19 different attributes were selected for the cleaned dataset. The missing values in the columns were replaced as per column data types.

Authors	Article Title	Source Title
Document Type	Author Keywords	Abstract
Affiliations	Funding Orgs	Cited Reference Count
Times Cited, WoS Core	180 Day Usage Count	Since 2013 Usage Count
Publisher City	Publication Year	Volume
Issue	Special Issue	DOI

Interesting attributes from the cleaned dataset are the "**citation reference count"** giving the detailed number of times the article was cited by other authors in their articles. This attribute also shows the reach of the idea from that particular article.

- The column of **"180 Day Usage Count"** and **"Since 2013 Usage Count"** gives the usage count of the articles in that particular time frame.
- **"Special Issue" and "Document Type"** attributes can identify specific articles classification.

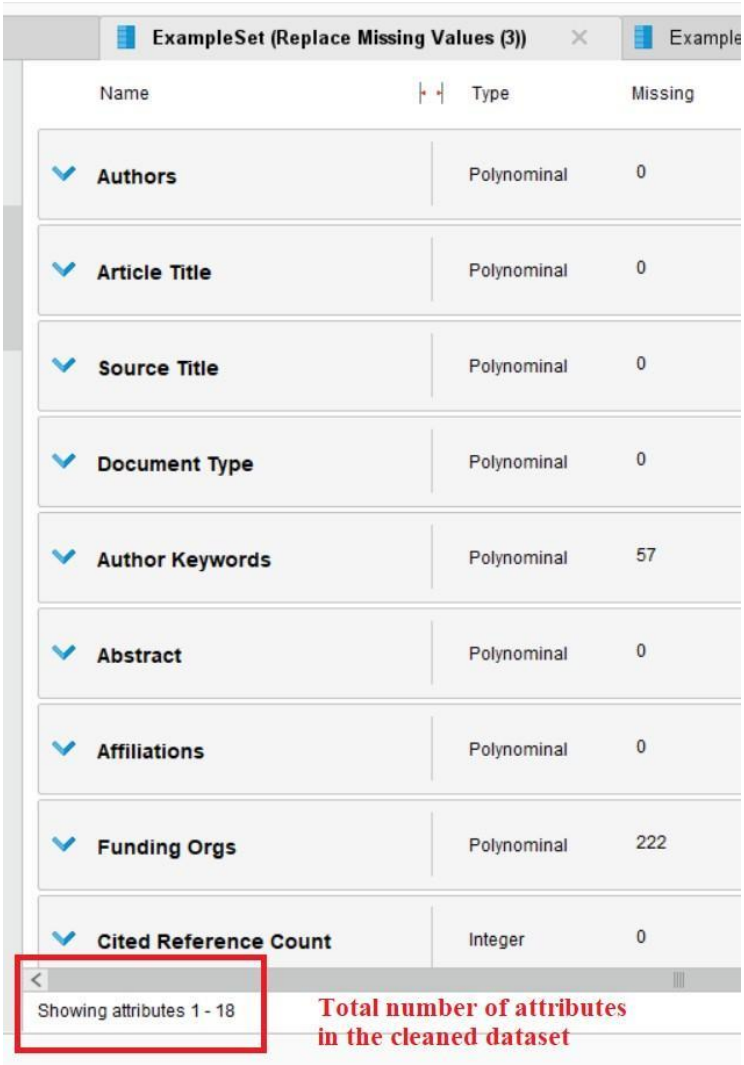


Figure 3 – RapidMiner Statistics Tab View for the Cleaned Dataset Attributes count

5. Data Analysis Plan

Analyzing the number of articles published in 2 decades based upon the abstract and selected keywords. This would allow us to relate the change in interests of the researchers over time. Also, we plan to show a visual distribution of the articles using charts based on the combination of attributes from the cleaned dataset.

The outcomes of this review analysis of the entire journal will also assist individuals who are interested in the study of digital libraries to have a simple dataset with 1-2 lines summary of the articles and can look for articles of choice easily.

5.1. Exploratory Data Analysis & Visualizations using Pivot Table in Excel.

EDA is an approach for data analysis that employs a variety of techniques. For the exploratory data analysis of the data of the articles collected from the "International Journal on Digital Libraries' several different graphical representations have been created using the tools of MS- Excel, Rapid Miner.

Chart-1: Representation of number of articles published in each year between the years 2002-2022.

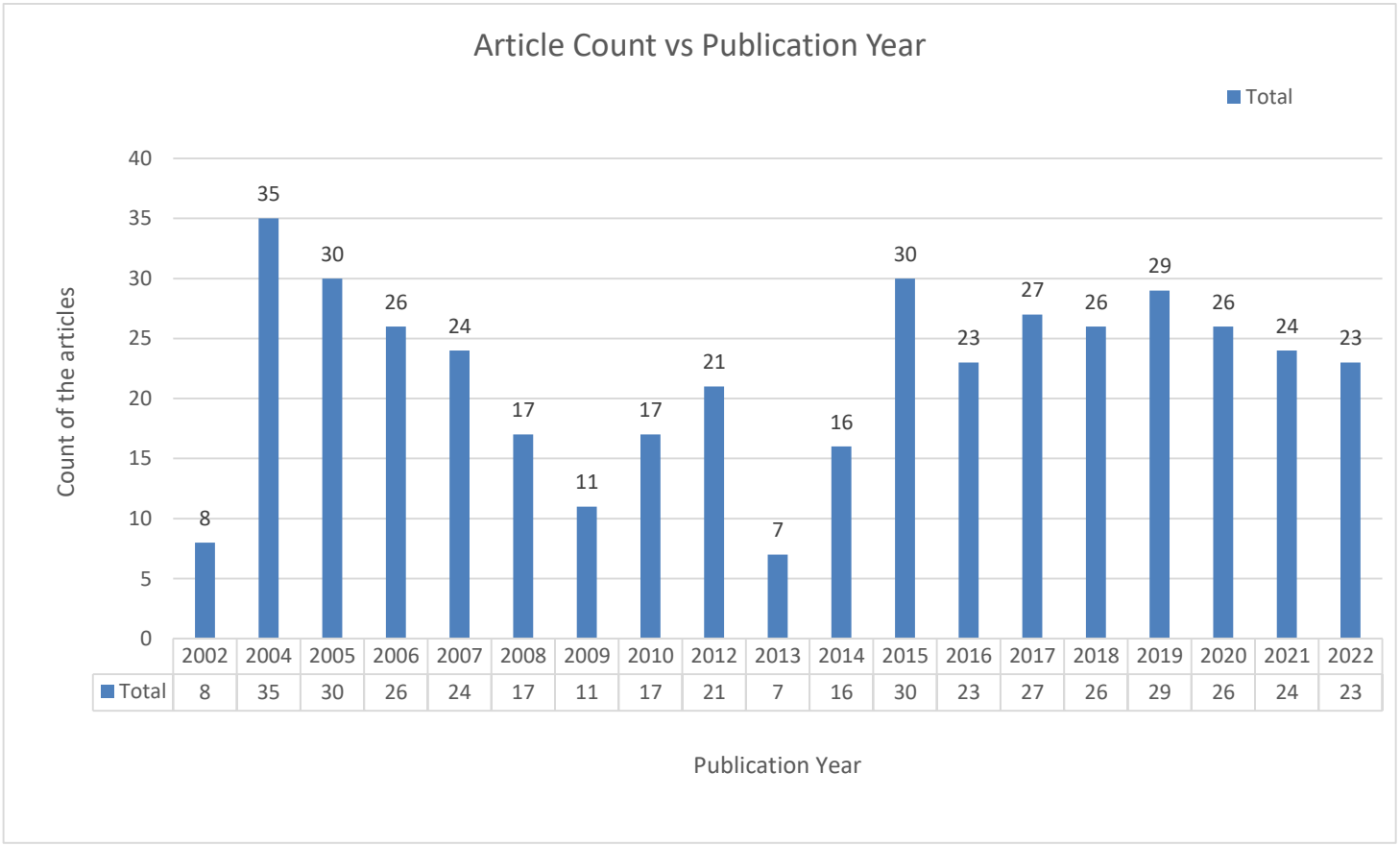


Chart-2: Representation of number of articles published in each issue of every year between the year 2002-2022.

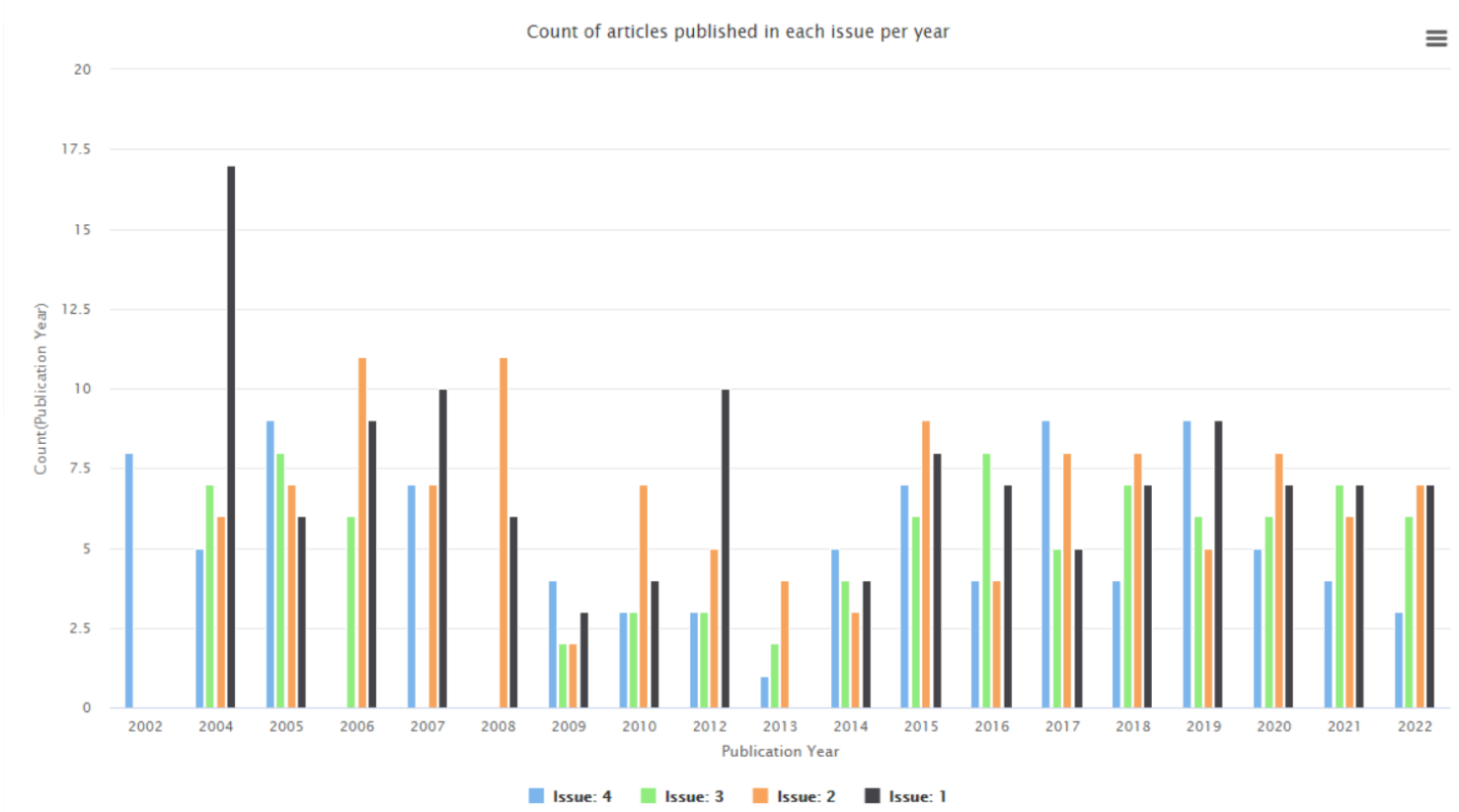


Chart-3: Representation of Citation Reference count of articles every year between 2002-2022.

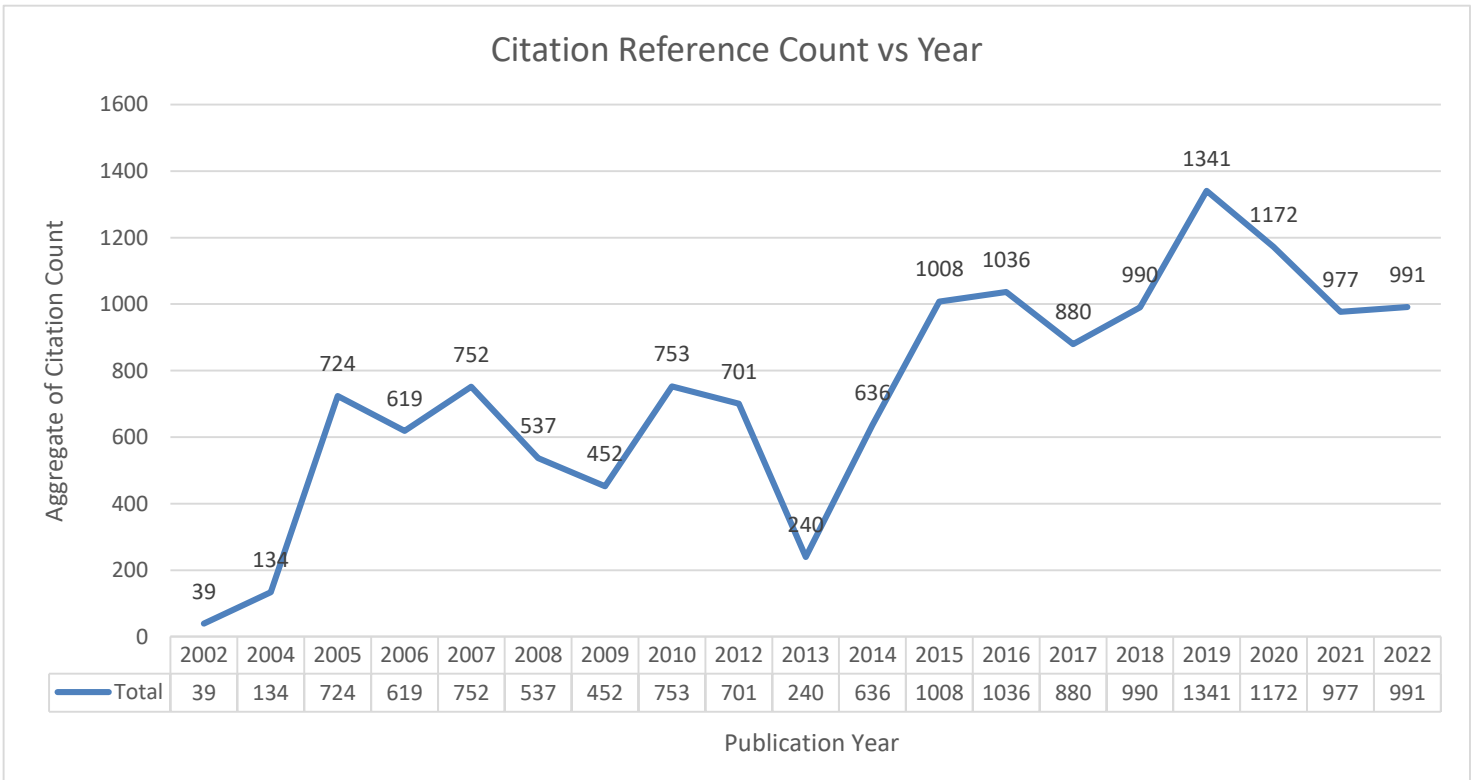


Chart-4: Representation of Document Category Type of the published paper in 2002-2022.

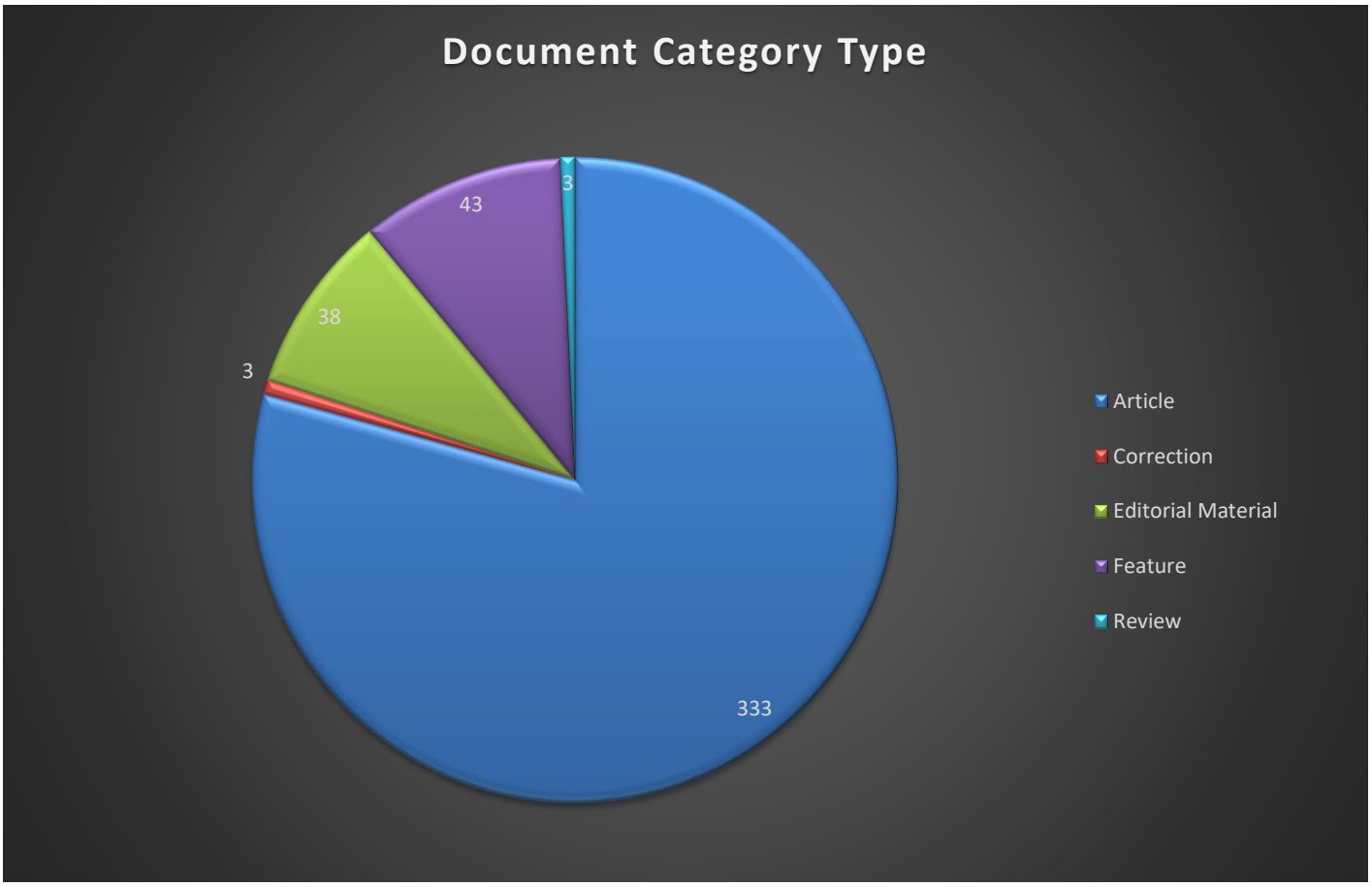


Chart-5: Representation of Funding Sources for each year

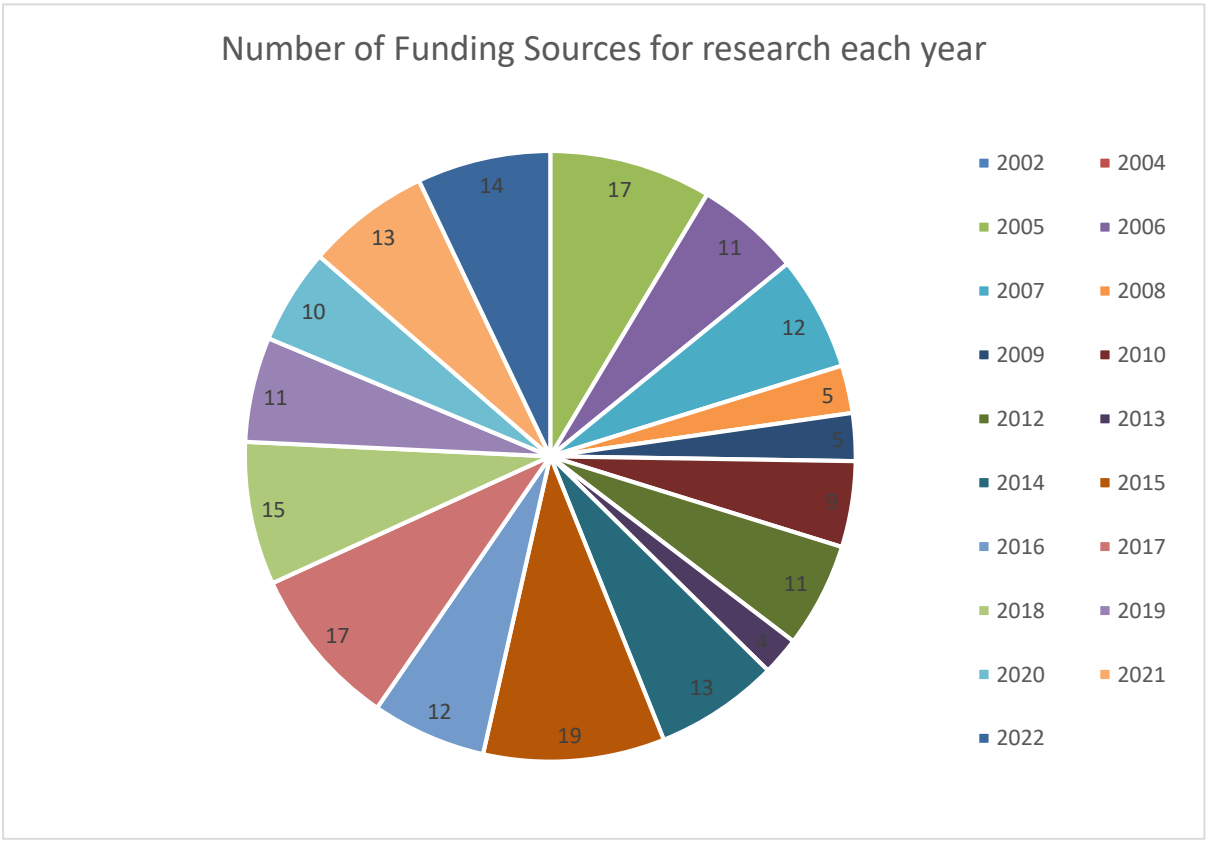


Chart-6: Funding from different Organizations

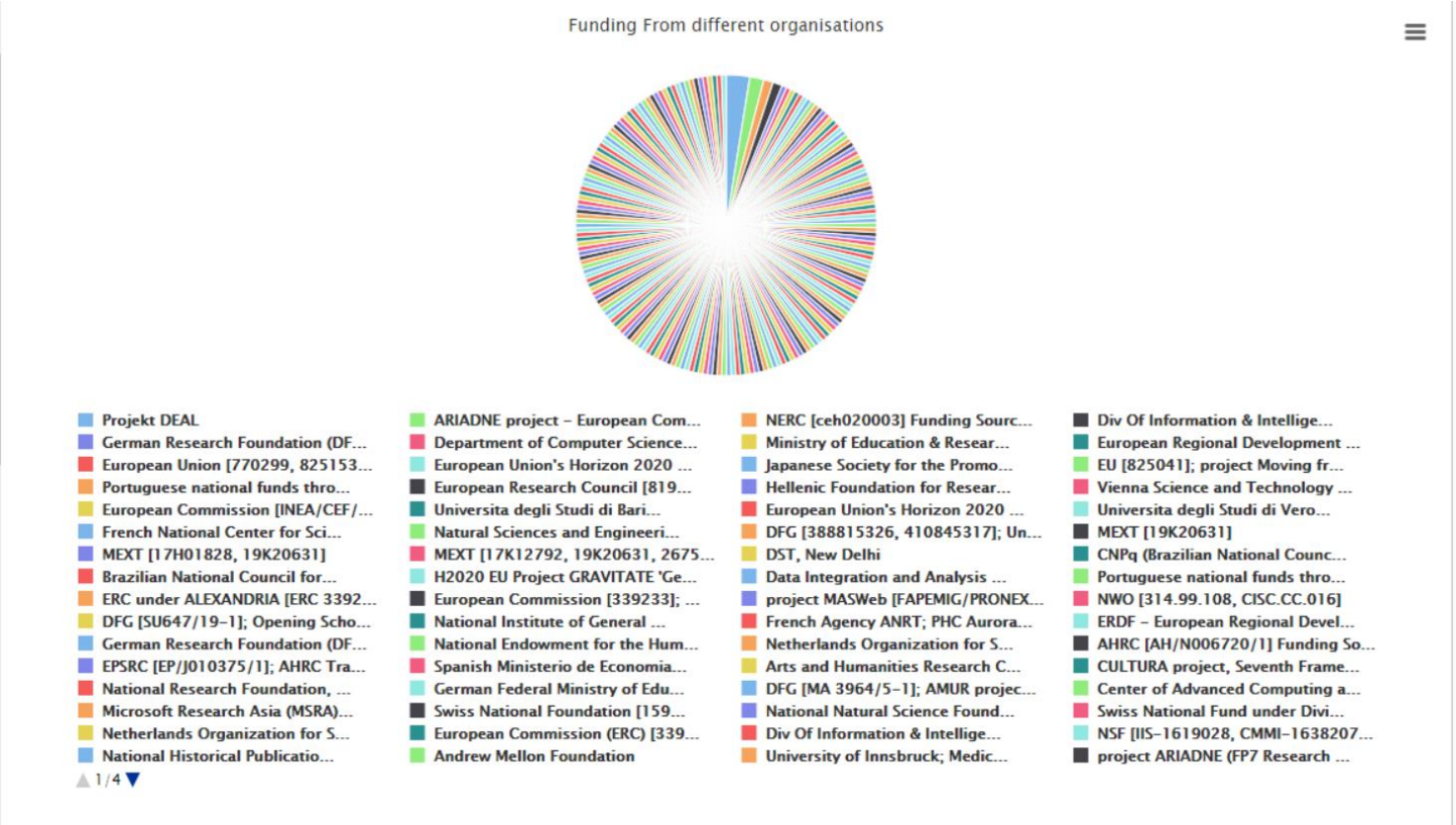


Chart-7: Representation of Special Issues published in the journal

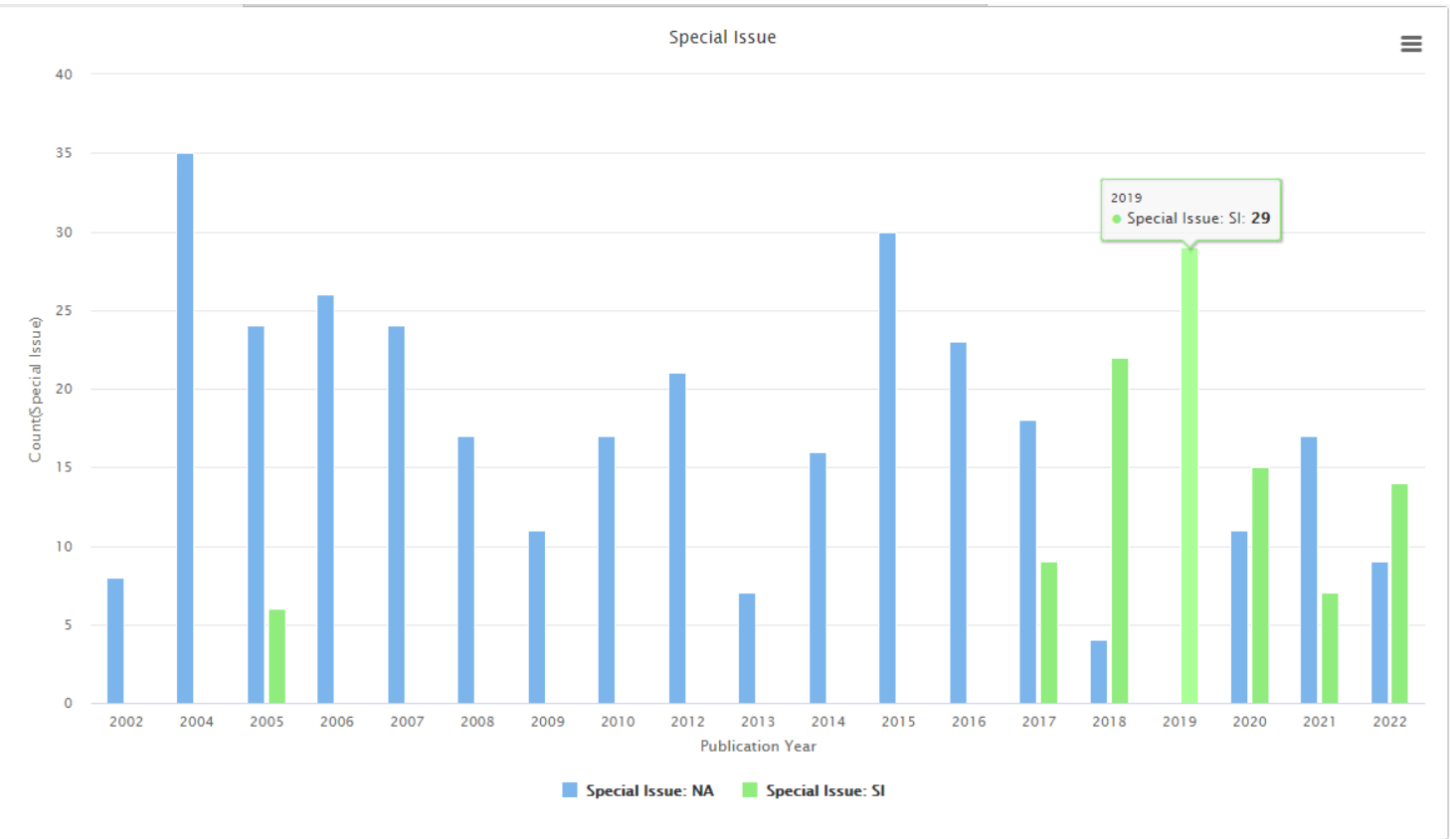


Chart-8: Representation of citation in last 180 days of the articles

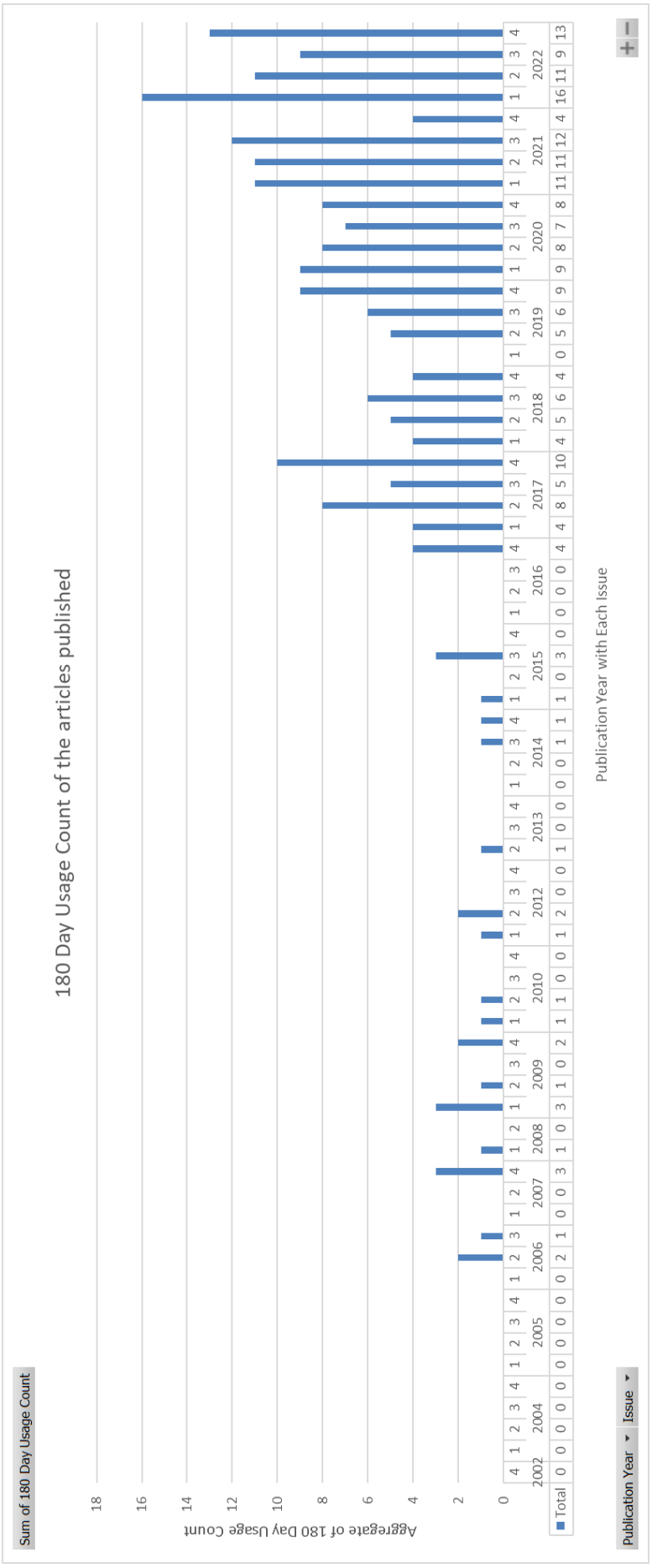


Chart-9: Representation of citation in since 2013 of the articles

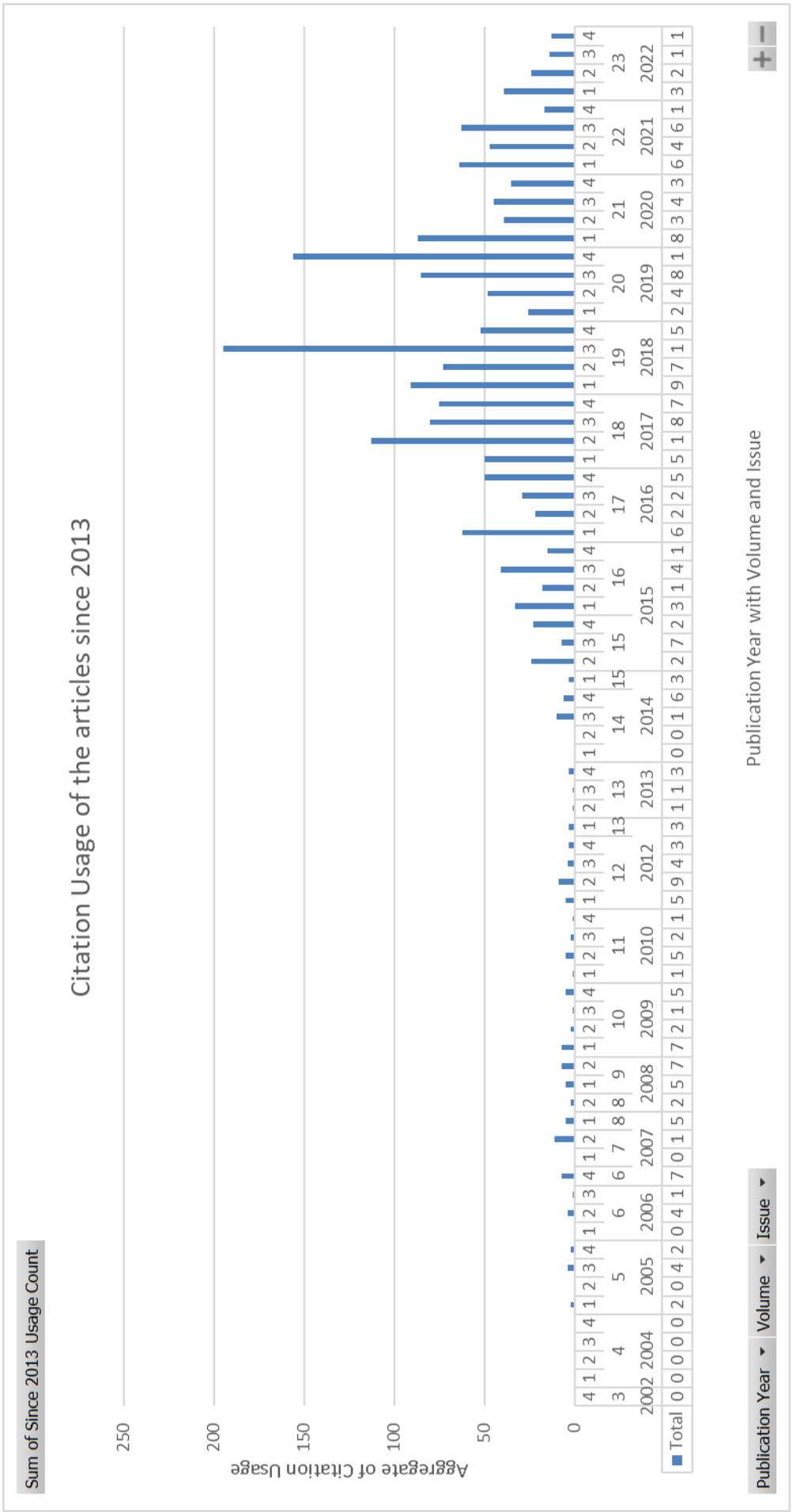


Chart-10: Representation of count of articles published in each volume from 2002-2022.

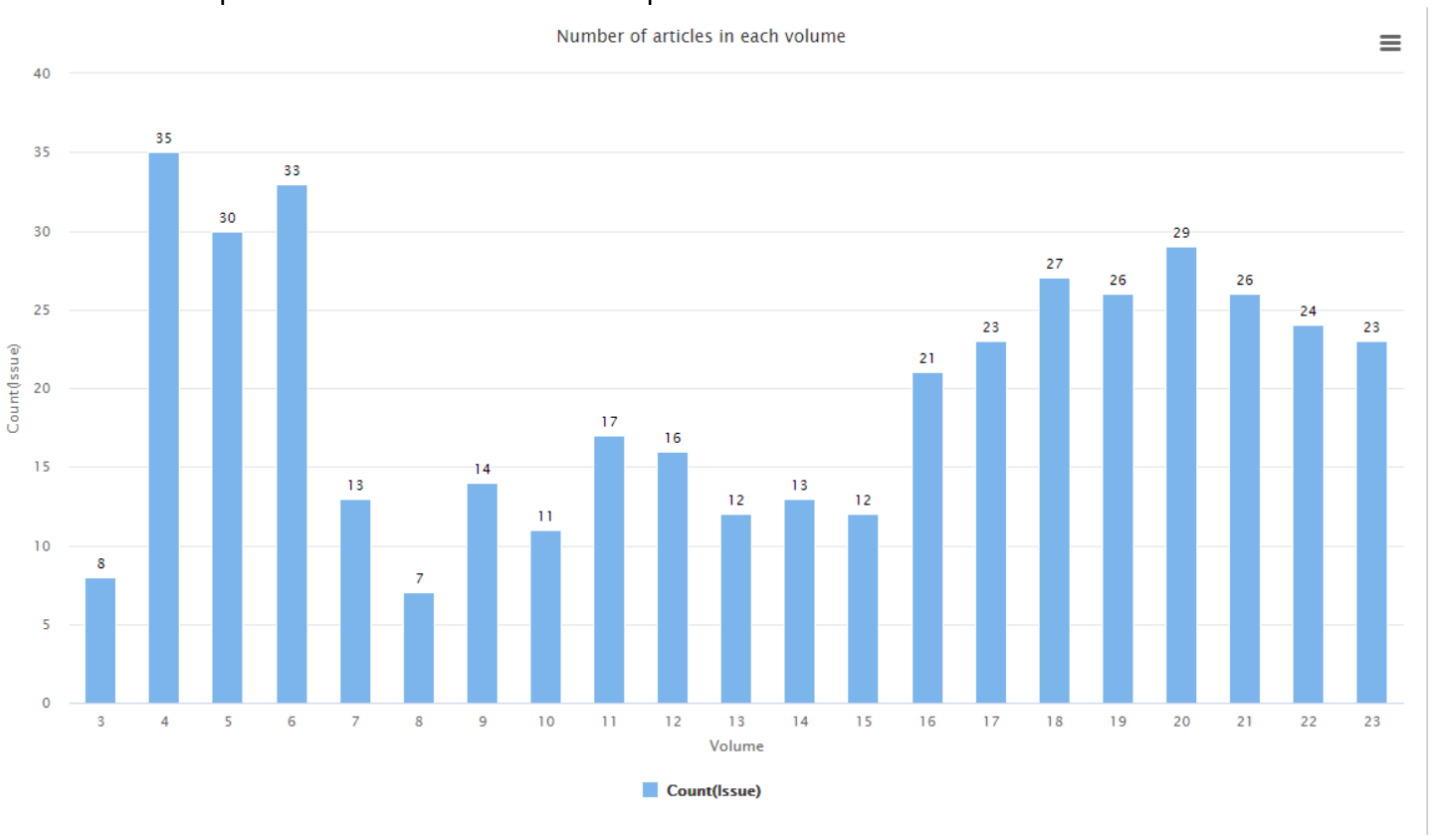
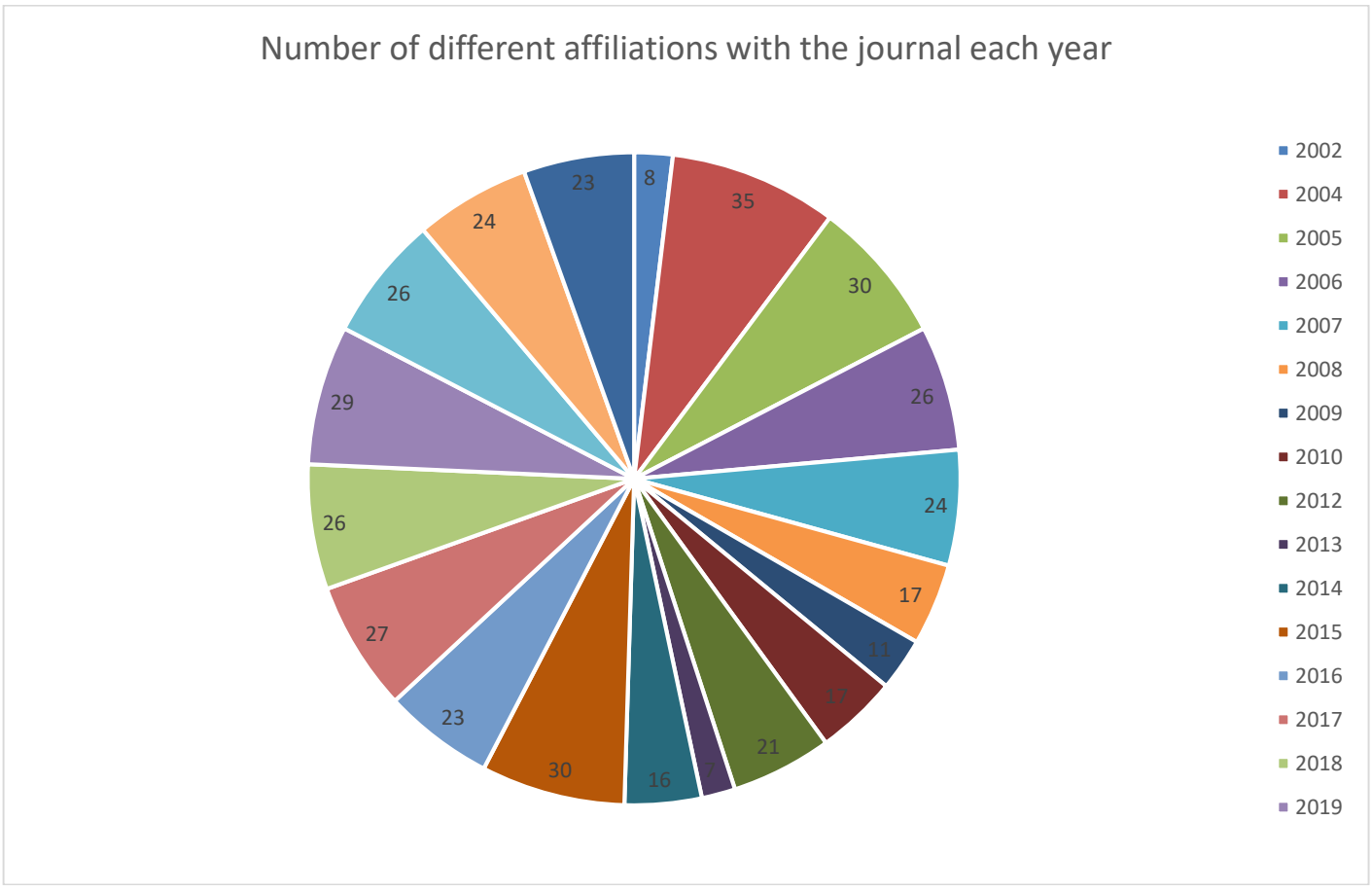


Chart-11: Representation of count of different university or organization affiliations with the articles published each year.



Author Keyword Analysis using Excel:

Approach 1: Including stop words and special characters as well.

Total Number of Keywords in the dataset: - 1378
Total distinct words as author keywords: 3378

Below is the attached screenshot of the analysis done on the author given keywords to the articles in the dataset.

Row Labels	Sum of Count		Cummulative Percentage of Keyword
Digital	159	159	5%
data	85	244	7%
web	70	314	9%
information	69	383	11%
libraries	53	436	13%
retrieval	46	482	14%
library	45	527	16%
preservation	41	568	17%
metadata	39	607	18%
analysis	39	646	19%
Semantic	34	680	20%
user	28	708	21%
system	28	736	22%
search	26	762	23%
music	24	786	23%
management	23	809	24%

- Top 15 words implies 708 out of the 3378 distinct words make up 20% of the Keyword population.
- There are 3378 unique words but the volume consisting half of them i.e. top 95 words that make only 2.81% of the total words forms the majority of the total keywords in the dataset.

Row Labels	Sum of Count		Cummulative Percentage of Keyword
Digital	159	159	5%
data	85	244	7%
web	70	314	9%
information	69	383	11%
libraries	53	436	13%
retrieval	46	482	14%
library	45	527	16%
preservation	41	568	17%
metadata	39	607	18%
analysis	39	646	19%
Semantic	34	680	20%
user	28	708	21%
query	7	1655	49%
mapping	7	1662	49%
indexing	7	1669	49%
documents	7	1676	50%
Electronic	7	1683	50%
interfaces	7	1690	50%
Event	7	1697	50%
humanities	7	1704	50%
architecture	7	1711	51%
Collaborative	7	1718	51%
Temporal	6	1724	51%
usability	6	1730	51%
Scientific	6	1736	51%
Topic	6	1742	52%
word	6	1748	52%

Replace Missing Values (3) | Article Count vs Year | Citation Count vs Year | Sheet11 | Citation Since 2013 | Document Cal

Approach 2: Excluding stop words, numbers and special characters as well.

Total Number of distinct Keywords in the dataset: - 1120

Below is the attached screenshot of the analysis done on the author given keywords to the articles in the dataset.

	A	B	C	D	E
1	**Word**	**Frequency**		**Cummulative Percentage**	
2	digital	159	159	5%	
3	data	85	244	7%	
4	web	70	314	9%	
5	information	69	383	11%	
6	libraries	53	436	13%	
7	retrieval	46	482	14%	
8	library	45	527	16%	
9	preservation	41	568	17%	
10	analysis	39	607	18%	
11	metadata	39	646	19%	
12	semantic	34	680	20%	
13	system	28	708	21%	
14	user	28	736	22%	
15	search	26	762	23%	
16	music	24	786	23%	
17	knowledge	23	809	24%	
18	management	23	832	25%	
19	document	22	854	25%	
20	and	21	875	26%	
21	research	21	896	27%	
22	systems	21	917	27%	
23	classification	20	937	28%	
24	archiving	20	957	28%	
25	cultural	20	977	29%	
26	archives	20	997	30%	
27	linked	19	1016	30%	

The keyword percentage is almost the same even after removing the unnecessary recurring stop words.
So, these top 10-20 keywords can be used by any new user to search for most of the articles in the International Journal on Digital Libraries.

Chart 12: Word Cloud for the Author Keywords

5.2 Data Analysis Using RapidMiner

P1) FP-Growth and Association Rule

The purpose of using the FP-Growth algorithm along with association rule mining is to get the frequently found combination of word in the abstract text column. In RapidMiner Studio, there are operators that can assist in performing these operations as per requirements.

Operator 1: **FP-growth Operator** efficiently calculates all frequently-occurring itemsets in a data set, using the FP-tree data structure.
Operator 2: **Association Rule operator** generates a set of association rules from the given set of frequent itemsets.

RapidMiner Process Canvas:

Figure 4 - Process canvas for the FP-growth and Association Rules operators

Process Documents from Data Operator:

This operator is used to create word vectors which are a collection of numerical values for each and every token in the dataset. As we are trying to generate association rules as well we will be using here Binary Term occurrences has been used as a parameter this checks the entire dataset for the attribute and then comes back to check every abstract data records. If the attribute is found it enters 1 else 0.

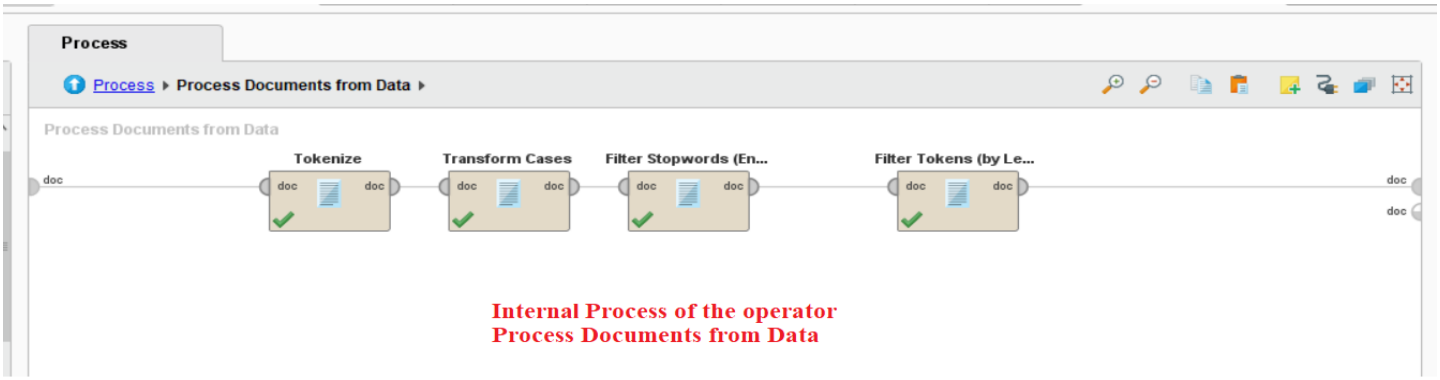


Figure 5 - Sub-internal process canvas for Process Documents from data operator

Results from the FP-growth model:

AssociationRules (Create Association Rules for abstract) | WordList (Process Documents from Data) | WordList (Process Documents from Data (2)) | AssociationRules (Create Association Rules for Keywords)

Word	Attribute Name	Total Occuren...	Document Occurences
ability	ability	11	11
able	able	26	24
abstract	abstract	5	4
abstracts	abstracts	4	4
abundance	abundance	3	3
academic	academic	18	15
acceptable	acceptable	2	2
acceptance	acceptance	3	3
accepted	accepted	3	3
access	access	77	60
accessed	accessed	5	5
accesses	accesses	2	2
accessibility	accessibility	7	7
accessible	accessible	13	12
accessing	accessing	6	6
accommodate	accommodate	4	3
accompanied	accompanied	2	2

WordList Generated from the abstract attribute in the dataset

AssociationRules (Create Association Rules for abstract) | WordList (Process Documents from Data) | WordList (Process Documents from Data (2)) | AssociationRules (Create Association Rules for Keywords)

Word	Attribute Name	Total Occurences ↓	Document Occurences
digital	digital	166	138
data	data	87	65
information	information	73	62
libraries	libraries	55	54
library	library	46	41
retrieval	retrieval	46	43
preservation	preservation	42	34
analysis	analysis	39	33
metadata	metadata	39	35
semantic	semantic	34	33
user	user	31	26
system	system	29	25
content	content	26	23
management	management	26	23
search	search	26	21
knowledge	knowledge	24	20
music	music	24	14

WordList generated from the Keywords attributes in the dataset.

This list of keywords and this occurences is similar to the analysis done previously in excel using pivot table.

Association Rule Mining for the Abstract Attributes:

AssociationRules (Create Association Rules for abstract) | WordList (Process Documents from Data) | WordList (Process Documents from Data (2)) | AssociationRules (Create Association Rules for Keywords)

Show rules matching

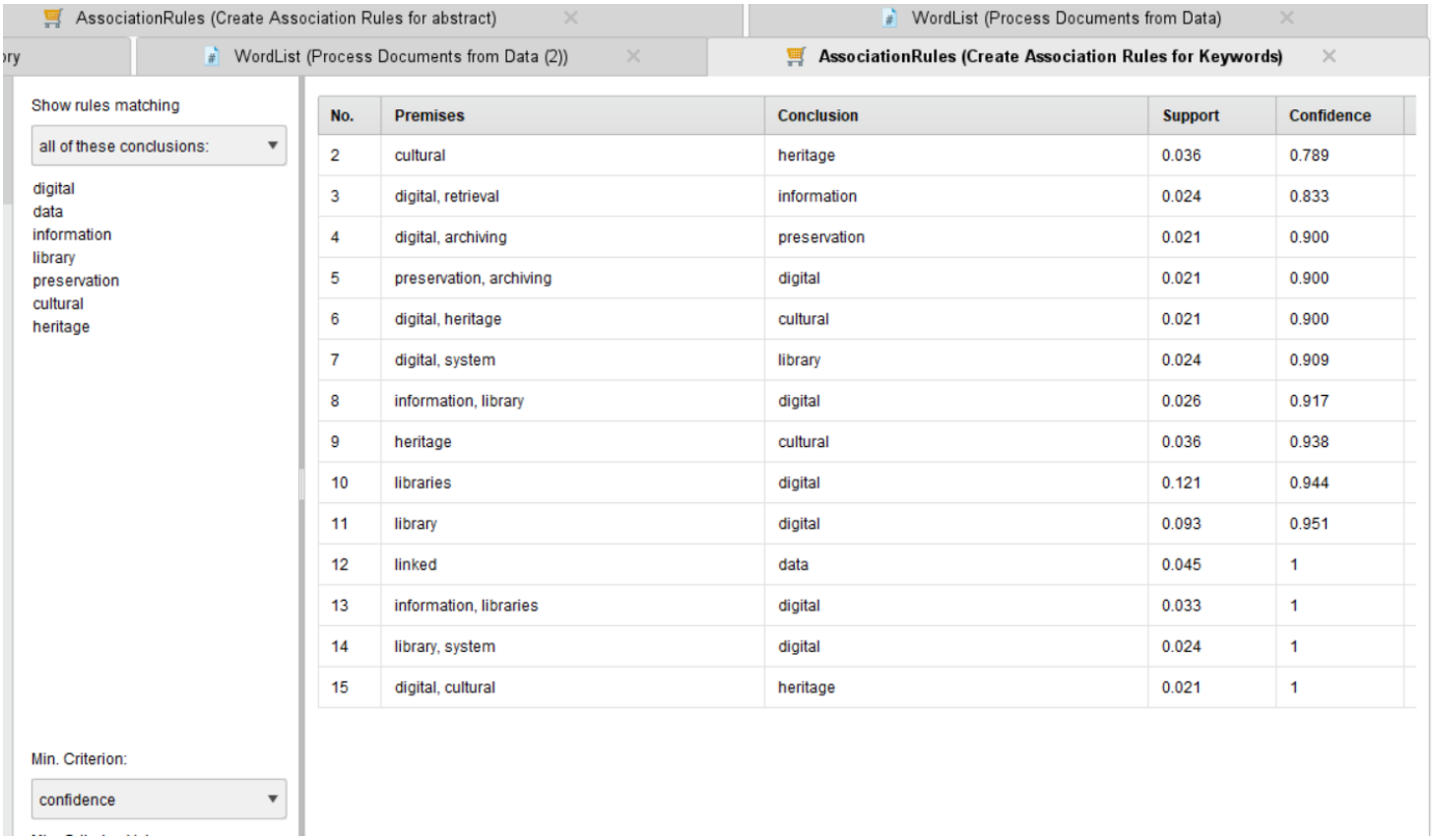
all of these conclusions:

digital

No.	Premises	Conclusion	Support	Confidence
3	data, library	digital	0.105	0.917
4	paper, libraries	digital	0.160	0.931
5	paper, library	digital	0.131	0.932
6	library	digital	0.226	0.941
7	information, library	digital	0.105	0.957
8	libraries	digital	0.274	0.958
9	data, libraries	digital	0.119	0.962
10	information, libraries	digital	0.129	0.964
11	libraries, library	digital	0.145	0.968
12	based, libraries	digital	0.124	1

```
Association Rules
[resources] --> [digital] (confidence: 0.719)
[article] --> [digital] (confidence: 0.737)
[data, library] --> [digital] (confidence: 0.917)
[paper, libraries] --> [digital] (confidence: 0.931)
[paper, library] --> [digital] (confidence: 0.932)
[library] --> [digital] (confidence: 0.941)
[information, library] --> [digital] (confidence: 0.957)
[libraries] --> [digital] (confidence: 0.958)
[data, libraries] --> [digital] (confidence: 0.962)
[information, libraries] --> [digital] (confidence: 0.964)
[libraries, library] --> [digital] (confidence: 0.968)
[based, libraries] --> [digital] (confidence: 1.000)
```

Association Rule Mining for the Author Keywords Attribute:



No.	Premises	Conclusion	Support	Confidence
2	cultural	heritage	0.036	0.789
3	digital, retrieval	information	0.024	0.833
4	digital, archiving	preservation	0.021	0.900
5	preservation, archiving	digital	0.021	0.900
6	digital, heritage	cultural	0.021	0.900
7	digital, system	library	0.024	0.909
8	information, library	digital	0.026	0.917
9	heritage	cultural	0.036	0.938
10	libraries	digital	0.121	0.944
11	library	digital	0.093	0.951
12	linked	data	0.045	1
13	information, libraries	digital	0.033	1
14	library, system	digital	0.024	1
15	digital, cultural	heritage	0.021	1

```
Association Rules
[visualization] --> [information] (confidence: 0.733)
[cultural] --> [heritage] (confidence: 0.789)
[digital, retrieval] --> [information] (confidence: 0.833)
[digital, archiving] --> [preservation] (confidence: 0.900)
[preservation, archiving] --> [digital] (confidence: 0.900)
[digital, heritage] --> [cultural] (confidence: 0.900)
[digital, system] --> [library] (confidence: 0.909)
[information, library] --> [digital] (confidence: 0.917)
[heritage] --> [cultural] (confidence: 0.938)
[libraries] --> [digital] (confidence: 0.944)
[library] --> [digital] (confidence: 0.951)
[linked] --> [data] (confidence: 1.000)
[information, libraries] --> [digital] (confidence: 1.000)
[library, system] --> [digital] (confidence: 1.000)
[digital, cultural] --> [heritage] (confidence: 1.000)
```

6. References

1. Pomerantz, J., Abbas, J. & Mostafa, J. Teaching digital library concepts using digital library applications. *International Journal on Digital Libraries* 10:1–13 (2009).
2. Agosti, M., Crivellari, F., Di Nunzio, G.M. et al. Understanding user requirements and preferences for a digital library Web portal. *International Journal on Digital Libraries* 11, 225–238 (2010).
3. Judy, J. (2005). "What is usability in the context of the digital library and how can it be measured?" *Information Technology and Libraries* (pp. 3-12).
4. Ojsteršek, M. (2011). Advanced features of Digital library of University of Maribor. *International Journal of Education*.
5. Nagu, Bansode. (2019). Impact of New Technologies in the Digital Libraries. *STM Journal-Journal of Advancements in Library Sciences*.
6. Rafi, Muhammad & JianMing, Zheng & Ahmad, Khurshid. (2018). Evaluating the impact of digital library database resources on the productivity of academic research. *Information Discovery and Delivery*, Volume 47.
7. Nabi, Somaira. (2012). World Digital Library: A Case Study. *Trends in Information Management* 8 (1), pp. 23-31.
8. Khot, Namita & Chavan, Yashvant. (2015). Digital Libraries: Challenges and Problems.
9. Gobinda G. Chowdhury, (2002). Digital libraries and reference services: present and future. *Journal of Documentation*, Vol. 58 Issue 3 pp. 258 – 283.
10. Mahadeva Gowda, Rajashekhar and S, Pavithrabai M., "Recent Trends in Digital Library Publications: A Scientometric Analysis'' (2022). Library Philosophy and Practice. 7058.
11. Smeaton, A., Callan, J. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries* 5, 299–308 (2005).
12. Borgman, C.L., Wallis, J.C. & Enyedy, N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries* 7, 17–30 (2007).

7. Contributions:

Kaushal Sen – Literature review of 2 articles. Took charge of organizing all the review works done by the team and preparing the document for both first & second project submission. Collection of dataset.

Venkata Suryasatya Kakarla – Responsible for organizing team calls for discussion. Assisted in the document preparation. Literature review of 2 articles. Assisted in Excel Analysis and visualizations

Karthik Rayapati – Literature review of 2 articles. Checking of duplicates in the cleaned dataset using RapidMiner Studio.

Kundana Lanka - Literature review of 2 articles. Data visualization in RapidMiner for second submission.

Likith Mallipeddi - Literature review of 2 articles. Data visualization in RapidMiner for second submission.

Sandeep Prasad Owk - Literature review of 2 articles. Assisted in Excel Analysis and visualizations

8. Future Work

- The team will be working on generating more analysis to insights from the collected dataset. Next target to achieve to provide trends or change in the interest of the authors in every 5 years of slots between 2002 to 2022.
- Target 2 will be to implement some other machine learning models in RapidMiner Studio to get more insights on the dataset.

The final outcome of the project will provide a detailed case study on the International Journal on Digital Libraries, analyzing the articles and creating useful visuals.