Emerging Technologies in Digital Libraries: Net Interactive Document (NID) Experiences and Prospects

Zaka, Bilal

Abstract: ***Modern digital library systems face challenges related to data analytics, interoperability, access control, and institutional and user collaborations. This paper presents the work done to add value to the conventional digital library workflows in a next-generation library system. Inherent capabilities to foster collaboration, content co-creation and AI interventions are among the desired features guiding the development of modern library systems. The work also highlights aspects of data transformation and access models in a more connected information ecosystem.***

Index Terms: ***Digital Library, Content Curation, AI interventions, Indexing, Big Data, Interoperability***

1. *INTRODUCTION*

Digital libraries are modern versions of conventional libraries having the same primary objective of information dissemination. The main features of a library include the ability of information cataloguing, easy access, circulation, content update, and record maintenance. At a more abstract level libraries promote education and culture in society. The digital face of libraries has all the traditional processes and services, but it additionally has certain features that are attributed to the digital nature of hosted information contents. A good digital library must provide a coherent view of the various types of media content it contains. Indexing and search capabilities of digital libraries make information access more precise. Contrary to service provision to immediate communities by conventional libraries, digital libraries tend to reach out to a larger and more diverse user base.

In the last decade, we have seen tremendous growth in digital content and the use of information technology in almost every field of life. Digital libraries now are a lot more than simple storehouses for organized digital documents. They are multimedia and structured information spaces running various services for their users. Digital libraries now are also platforms for communication, collaboration, learning and research. Knowledge seekers with access to fast internet and mobile devices look for immediate and precise information almost all the time. Digital libraries are expected to provide this type of access to their users and, also, a place to be more productive. As a result, we see changes in the architecture of digital libraries and the increasing use of new and innovative technologies.

The NMC Horizon Report 2017 Library Edition [1] suggests that the trends related to user experience improvements, cross-institution collaborations, and rethinking of library spaces are forcing digital libraries to use the latest technology. The challenges facing modern digital libraries in this evolution phase include accessibility, technology literacy, adapting to existing organization designs, and ongoing integrations. More complicated issues being faced by libraries during this transition are interoperability, intellectual property, rights management, and economic and political pressures.

The enormous size of digital content in libraries that tends to increase exponentially is now being treated as big data [2]. To effectively use such large information spaces, it is essential that, in addition to traditional data processing software, advanced computational methods are also used. The use of modern analytic tools will help to consume information in digital libraries more effectively. Furthermore, it will help to reveal patterns, trends, and associations that are linked to the human behavior of users and their interactions.

The research on the topic of how big data and user preferences are changing digital library services [3][4] highlights the fact that user data plays an important role in the new digital library ecosystem. User interactions captured by digital libraries not only help in fulfilling the information needs of users but also continuously add resources to the library. User-contributed resources build a broader understanding of desirable library resources. In the current information overload era, digital library users look for a personalized experience and access to relevant content from diverse media spaces. They are also more interested in autonomous tagging of information contents for personal understanding and peer sharing. The ability to supplement library contents by adding annotations to source information is another desirable aspect of modern digital libraries.

Research work and studies [1][2][3][4] are highlighting futuristic trends and a way forward for the digital library systems but existing mainstream platforms are still using the conventional data processing, management, and publishing approach.

An experimental study evaluating digital libraries [5] reveals that users of digital libraries acknowledge the improvements in systems by modern technology interventions. The improvements in information retrieval mechanisms, usability, and access interfaces are acknowledged. However, at the same time users of digital library systems also feel the need for further improvements with cutting-edge data science and ICT usage. The commonly used claims of open-source digital library platforms [6] are to be extensible and increasingly allow the use of diverse media types, standard metadata, content management system, and multimodal access. There are also efforts made to cover challenging access management issues with copyrights and licensing implementations. Fast and reliable storage and hosting also complement the digital preservation agenda of digital libraries. Many digital publishing platforms started offering recommender systems based on content filtering or collaborative filtering approach. we also see increasing use of machine learning algorithms offering different analytics on library and user data. While all this development shows progression, there is still a lot of room for improvements in domains such as interoperability, user and institutional collaborations, and AI usage during content creation and consumption.

The author of this paper and his team started working on the development of a modular and extensible digital library platform that makes use of niche technologies [7][8]. The main objective of this development activity was to address the above-highlighted deficiencies in digital library systems. Over the past couple of years, our team deployed different instances of a next-generation digital library platform[1] at the Institute of Interactive Systems and Data Science, Graz University of Technology Austria.

This paper discusses the results of our experiments done to improve content creation, co-creation, management, and interoperability features in a digital library system. In the following sections of the paper, we will identify the focus area of the library administration portal, i.e. - content preprocessing before publishing and its delivery to users. The system's ability to engage users for content co-creation and digital access right management is also presented. We will also highlight measures taken to make the system more interoperable. We will conclude the paper by pointing out ways forward and the needed next steps toward a comprehensive and modern digital publishing platform.

2. Content Curation in Digital Libraries

In modern times most of the general information and intellectual content created by individuals and institutions are in digital format. The term data or content curation is used for the creation, organization, integration, and value addition of the information in digital libraries. Due to the exponential growth of digital information, the managers of information stores and librarians that host this information look for innovative ways to preserve this knowledge.

Data curation is becoming an increasingly important domain because of its importance in information discovery, delivery, and re-use. A generic and comprehensive data curation life cycle model by Digital Curation Center [9] provides a high-level graphical overview of the steps involved in the complete process.

The actions presented in the data curation lifecycle (Fig. 1) are used by different digital library systems, starting from data object collection. Actions performed range from adding descriptions and metadata, preservation planning, engaging the user community, setting access mechanisms and finally transforming data into suitable formats for storage and use.

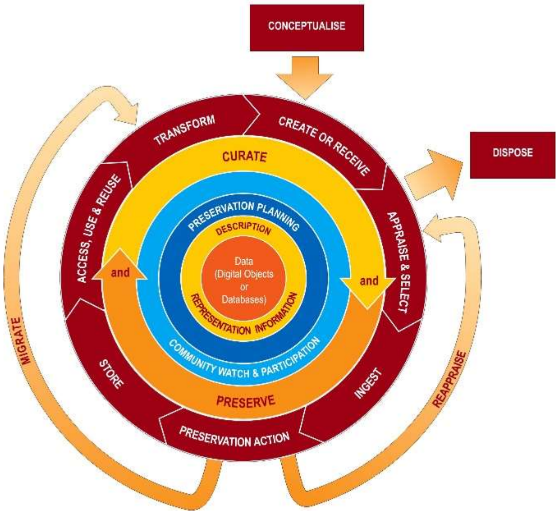


Fig.1.: DCC lifecycle model [source: https://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf]

[1] https://nid.iicm.tugraz.at/

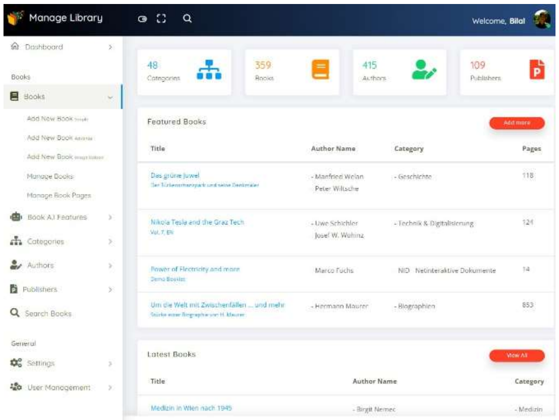
There are additional steps involving data reappraisal, migration, conceptualization, and removal. Led by this holistic lifecycle and generic action categories, the management portal of the NID library system is designed to perform the following data curation steps using administrative accounts:

2.1 Data Addition

NID system allows uploading standard PDF documents using a simple interface with minimal description data and default access and use models. The advanced document addition allows detailed description entries with options to select access rights and settings related to community use, data transformation, and views. NID also allows the creation of library objects using high-resolution images. Support for documents in proprietary formats e.g. MS word can also be added for user-specific needs.

2.2 Additional Descriptors and Preservations Planning

Using the Manage Book options of the NID library system administrative users can edit book preservation and descriptor settings and also set view menu options and annotation addition behavior for a particular document.



48 Categories | 359 Books | 415 Authors | 109 Publishers

Featured Books

Title	Author Name	Category	Pages
Das grüne Juwel Der Türkenschanzpark und seine Denkmäler	- Manfried Welan Peter Wiltsche	- Geschichte	118
Nikola Tesla and the Graz Tech Vol. 7, EN	- Uwe Schichler Josef W. Wohinz	- Technik & Digitalisierung	124
Power of Electricity and more Demo Booklet	Marco Fuchs	NID Netinteraktive Dokumente	14
Um die Welt mit Zwischenfällen ... und mehr Stücke einer Biographie von H. Maurer	- Hermann Maurer	- Biographien	853

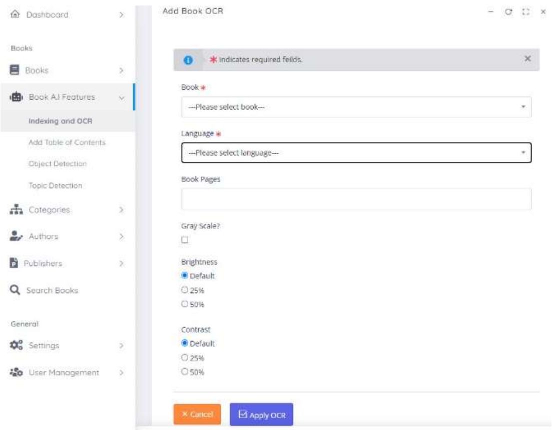
Latest Books

Title	Author Name	Category
Medizin in Wien nach 1945	- Birgit Nemec	- Medizin

Fig. 2.: Overview of NID library administrative portal

2.3 Creation of Representation Information

A NID library offers, in addition to the specification of standard document metadata, the creation of an inverted index of documents. This allows fast full-text search of text contents of documents with the specification of the exact location in documents. The full-text search facility is complemented by various enhancements that include the detection of text and objects using Optical Character Recognition (OCR) technology, Natural Language Processing, and Computer Vision algorithms, more details on this are given in section 4.



Add Book OCR

* Indicates required feilds.

Book *
---Please select book---

Language *
---Please select language---

Book Pages

Gray Scale?
□

Brightness
● Default
○ 25%
○ 50%

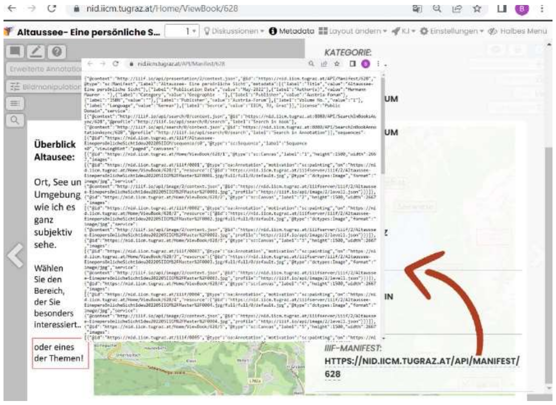
Contrast
● Default
○ 25%
○ 50%

Fig. 3.: NID OCR, with image preprocessing options

2.4 Data Transformation

The information content added in the NID library is transformed into an image format for end-user delivery and consumption by other systems. The International Image Interoperability Framework (IIIF) standard[2] is selected for the online delivery of high-quality, attributed digital objects at scale. The image-based transformation of conventional documents is done by keeping the text lookup feature intact while eliminating the easy possibilities of text copying without permission. The manifest metadata descriptor of IIIF helps in the specifications of page-level objects of the system. This transformation greatly helps in device-independent, cross-platform delivery of information contents in a library.

NID library offers to set display properties of a document at a page level. Librarians (administrators and editors) can set the display of pages in a particular order and choose to hide the pages from the display as well.



Überblick Altausee:

Ort, See un Umgebung wie ich es ganz subjektiv sehe.

Wählen Sie den Bereich, der Sie besonders interessiert..

oder eines der Themen!

{"@context":"http://iiif.io/api/presentation/2/context.json","@id":"https://nid.iicm.tugraz.at/API/Manifest/628","@type":"sc:Manifest","label":"Altaussee- Eine persönliche Sicht","metadata":[{"label":"Title","value":"Altaussee- Eine persönliche Sicht"},{"label":"Publication Date","value":"May-2022"},{"label":"Author(s)","value":"Hermann Maurer - "},{"label":"Category","value":"Geographie - "},{"label":"Publisher","value":"Austria-Forum"},

IIIF-MANIFEST:
HTTPS://NID.IICM.TUGRAZ.AT/API/MANIFEST/628

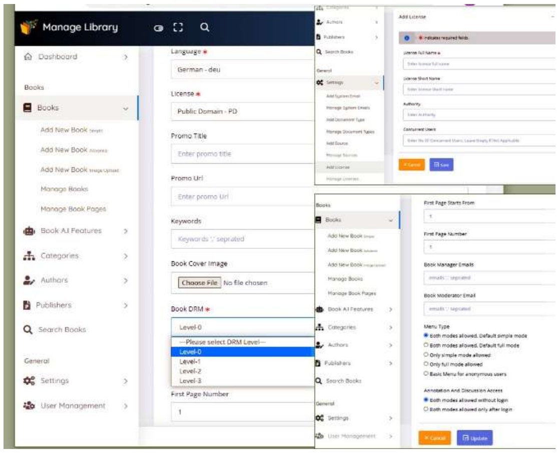
Fig. 4.: Extended metadata of library objects in NID

2.5 Access, Use and Re-use

Access, Use and Re-use: Content access management in NID administrative portal is done

[2] https://iiif.io/

using a unique three-tier approach. The first access management mechanism is based on Digital Rights Management (DRM) levels. The default NID systems come with pre-defined three DRM levels i.e. 0/1/2. The DRM-0 selection means rights management is disabled and access is granted to all users including anonymous ones. DRM-1 access allows only registered/privileged users to access the library objects. The DRM-2 assigned objects can only be viewed by the editors and system administrators. These levels correspond to the pre-defined user categories in NID distribution i.e., anonymous, registered, editors and administrators. The NID library system can have more user categories and more corresponding DRM levels. The second tier of access management is based on license types. The NID system allows the addition and management of license types which in turn are assigned to objects being added to the library. The default system license types include different variants of Creative Commons[3] license type, Public Domain, and Copy Right/All rights reserved. The license descriptor not only identifies the use, re-use, and sharing characteristics, but also compliments the NID user access layer mechanism. A license type in the NID system can be created with specific concurrent user access to limit its online use according to the allowed distribution rights of the library. A document in the library added with a Public Domain license type provides a download link to its source file. The third layer of access is invoked by privileged users, editors and administrators, with the console section of the main library portal. Access control at the user group level can be defined by creating a group of users and assigning exclusive access rights to selected library objects to a certain user group.



Language *
German - deu

License *
Public Domain - PD

Promo Title
Enter promo title

Promo Url
Enter promo Url

Keywords
Keywords ',' seprated

Book Cover Image
Choose File No file chosen

Book DRM *
Level-0
--Please select DRM Level--
Level-0
Level-1
Level-2
Level-3

First Page Number
1

Add License
* indicates required fields.
License Full Name *
Enter license full name
License Short Name
Enter license Short name
Authority
Enter Authority
Concurrent Users
Enter No Of Concurrent Users, Leave Empty If Not Applicable

First Page Starts From
1
First Page Number
1
Book Manager Emails
emails ',' seprated
Book Moderator Email
emails ',' seprated
Menu Type
● Both modes allowed, Default simple mode
○ Both modes allowed, Default full mode
○ Only simple mode allowed
○ Only full mode allowed
○ Basic Menu for anonymous users
Annotation And Discussion Access
● Both modes allowed without login
○ Both modes allowed only after login

Fig. 5.: NID Library Access control options

3. Collaborations and Interoperability

Digital Libraries in more interactive and interlinked online environments face the long-term challenge of user and institutional collaborations. While librarians and technologists both agree on the fact that digital content publishing platforms must have intuitive means of promoting user interactivity at both system level and across different systems, we see limited inherent capabilities of user collaborations and data interoperability across different platforms. The surveys done to access the capabilities of common digital library systems reveal that there are efforts made to promote standardized metadata generation and consumption for interoperability and syndication services [10]. The prevailing digital library systems are also increasingly making use of various external social media plugins to promote user collaborations. These efforts need to be complemented by adding built-in user interaction features in digital libraries and standard application programming interfaces (APIs) for sharing and consuming data.

To support the development of a digital library system in this positive direction we added the following collaborative and interoperability capabilities to the NID platform.

3.1. *Annotations, Discussions, Quizzes at Public and Group Levels*

Going beyond passive consumer of information, NID library users can add comments, or contribute information in form of annotations, comments, feedback, quizzes, and even start a discussion at the document page level. The content co-creation by the user community must also include necessary moderation controls. NID implements this control by specifying the user interaction behavior at the base level where a librarian can specify whether the library object/document allows anonymous interactions or only logged-in registered users can contribute to discussions and annotations. The added contents are made available to general users of the library system unless the library objects' content moderation is not enforced. The content moderation is enforced by specifying the moderator email of any document in a NID library. This moderation control can be applied at the page level but also allows adding multiple content moderators for specific sections of a document. This additional control allows visibility of user contribution only after approval/ activations by the moderators or editors of the library system. The NID system also provides the facility to create user groups. The library users can create user groups and add document interactions (annotations, quizzes, and

[3] https://creativecommons.org/

discussions) that are only visible to users of the intended group.

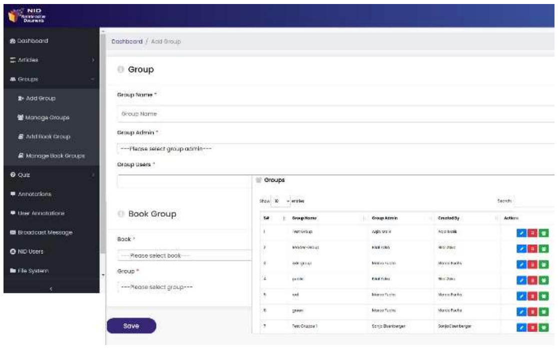


Fig. 6.: User Group features in NID Library

3.2. *Interoperability*

Seamless data sharing at the system level is very critical to maintaining a robust and extensive information ecosystem. Digital libraries must be capable of sharing information contents in addition to exchanging metadata. The NID library system makes use of the IIIF standard to extend the seamless content delivery and syndication services. NID data representation is based on a simple design principle reducing the use of any proprietary data format and specific technologies. It uses image formats that can be used on any platform and device. The IIIF image representation is complemented by link data in JSON-LD format. This web standard link data uses an open standard file format and data interchange format. The context of original data and use of link data is presented in human-readable text. The JSON manifest files store and transmit data object references consisting of attribute-value pairs and arrays.

NID system provides ways to access, view, search, and share digital library objects in form of images, audio, and video. To support digital library synergy the NID system provides multiple APIs that allow sharing of information and content aggregation. This includes search API, Presentation API, Image API, and data services APIs.

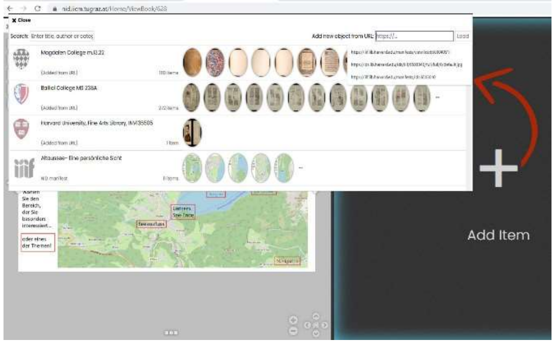

Close
Search: Enter title, author or collection
Add new object from URL: https://...
Magdalen College ms.0.23
Balliol College MS 238A
Harvard University, Fine Arts Library
Altaussee - Eine persönliche Sicht
Add Item

Fig. 7.: Access to objects in other libraries

The front-end operational example of content sharing and digestion APIs include a "transclusion function" for URL shared contents and import of IIIF-compliant information objects from other libraries.

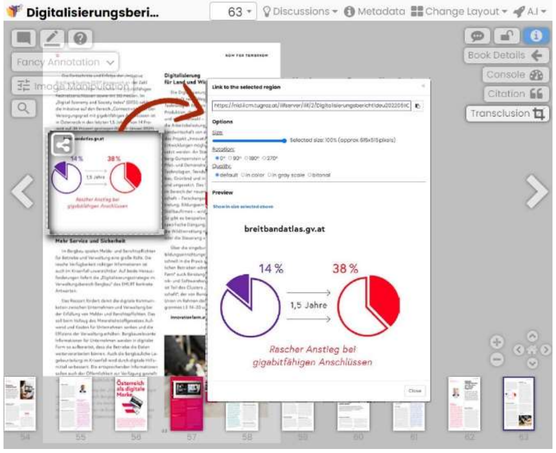


Fig. 8.: URL-based content sharing

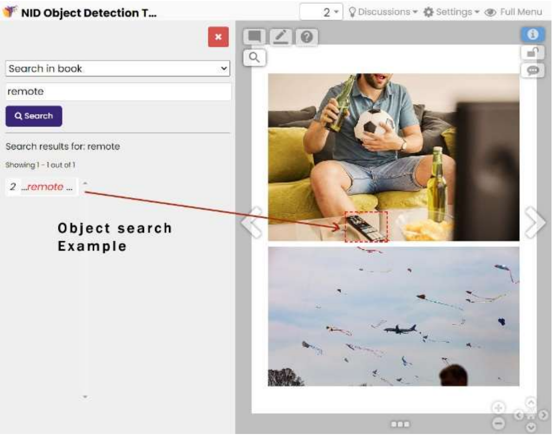
4. *Big Data and AI Applications in Digital Library Systems*

Big data is mainly characterized by Volume, Variety, Velocity, Variability and Value [11]. The data in modern digital libraries fit this characterization to a great extent. Increasing use of conventional and mobile computing devices, improved ICT infrastructure, internet applications, and convenient digitization of legacy data are what cause exponential and speedy growth of data in libraries. The variety of sources and content types add to the variability in library data. Contrary to general user data available on social platforms and common information systems, the data available in digital libraries is of greater intellectual value as it generally comes from credible sources. Libraries in the big data era need to shift focus from conventional data processing applications and employ tools more suitable and designed exclusively for big data. The challenges associated with massive datasets are capturing and storage of data, searching for the desired information, selective sharing and transfer, visualization, querying, privacy etc. Artificial Intelligence (AI) works very well in data analytics, thus making AI and big data seemingly inseparable. To exploit the full potential of data volumes in digital libraries, the NID library makes extended use of machine learning algorithms in several areas as follows.

4.1. Computer Vision (CV)

NID uses Tesseract[4] - one of the most accurate open-source OCR engines. It helps in indexing processes by extracting machine-readable text from image-based documents. This feature is very practical for libraries that archive historical documents through a scanning process. The seamless integration of the inherent Optical Character Recognition (OCR) engine automates the indexing tasks. It saves time and eliminates the use of expensive external software. Long short-term memory (LSTM) based techniques are used by the OCR engine supporting up to 116 languages including experimental support for old manuscripts. The NID platform has supplemented the OCR processing with additional image pre-processing features. Our experiments yielded greater text detection accuracy when scanned document images were pre-processed with grey scaling and brightness and contrast adjustments. This temporary intermediate backend processing is useful for documents with text having colored or image backgrounds.

Another AI application tested in the NID library system is in the computer vision domain using Object Detection techniques to supplement a conventional text index of a document. The NID library added the YOLO [12] algorithm to its data analytics toolbox. The Object Detection function applied to an object in the NID library detects and recognizes various image objects on a document page. It stores the location and label of the detected object to the standard text search index of that document. At present NID system uses the COCO[5] training dataset capable of detecting 80 object classes.



Search in book

remote

Search results for: remote

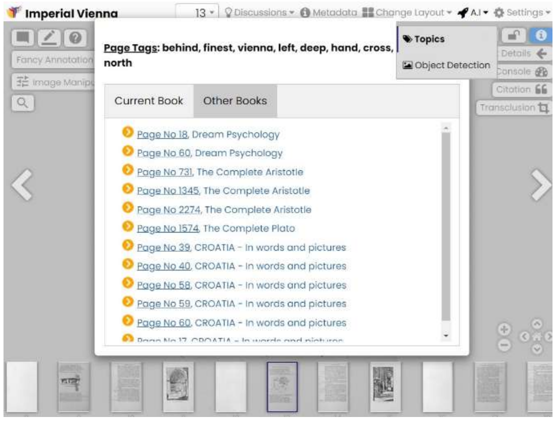
Showing 1 - 1 out of 1

2 ...remote ...

Fig. 9.: Search facility of objects without any text description

4.2. Natural Language Processing (NLP)

NLP-based AI algorithms help machines understand human language. Its use gives insights into text available in various documents of a digital library. A NID library contains the experimental feature of "Topic Detection". Topic modelling in NID is done using Latent Dirichlet Allocation (LDA), an unsupervised approach by extracting the patterns of word clusters and frequencies of words in the document [13]. The output of this AI application gives us the gist of contents available in a document. The NID topic detection features make use of detected topics to find similarities among pages of the same document and pages of other documents. The topic modeling can also be used for the automated classification of document clusters in large libraries.



Page Tags: behind, finest, vienna, left, deep, hand, cross, north

- Page No 18, Dream Psychology
- Page No 60, Dream Psychology
- Page No 731, The Complete Aristotle
- Page No 1345, The Complete Aristotle
- Page No 2274, The Complete Aristotle
- Page No 1574, The Complete Plato
- Page No 39, CROATIA - In words and pictures
- Page No 40, CROATIA - In words and pictures
- Page No 58, CROATIA - In words and pictures
- Page No 59, CROATIA - In words and pictures
- Page No 60, CROATIA - In words and pictures
- Page No 17, CROATIA - In words and pictures

Fig. 10.: Experimental feature of topic modeling and detection of similar pages.

5. CONCLUSION

In this paper, we discussed how digital libraries must evolve in the wake of mass digitization and the prolific use of the internet. Our initial efforts to introduce inherent collaborations and content co-creation features improve user involvement with the system. The experiments related to data transformation and contemporary as well as unorthodox access mechanism also adds value to interoperability and intellectual property rights areas. We also established that present-day digital libraries are seen as sources of big data. The use of artificial intelligence techniques for seeking meaningful insights from this data is inevitable. We see great potential for further experiments for the applications of cutting-edge natural language processing and computer vision algorithms on conventional library data.

[4] https://en.wikipedia.org/wiki/Tesseract_(software)

[5] https://cocodataset.org/

REFERENCES

[1] Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Giesinger Hall, C., Ananthanarayanan, V., Langley, K., and Wolfson, N. (2017). “NMC Horizon Report: 2017 Library Edition”, Austin, Texas: The New Media Consortium.

[2] Washington Kamupunga and Yang Chunting. “Application of Big Data in Libraries”, International Journal of Computer Applications 178(16):34-38, June 2019.

[3] Shuqing Li, Fusen Jiao, Yong Zhang, Xia Xu, “Problems and Changes in Digital Libraries in the Age of Big Data From the Perspective of User Services”, The Journal of Academic Librarianship, Volume 45, Issue 1, 2019, Pages 22-30, ISSN 0099-1333, https://doi.org/10.1016/j.acalib.2018.11.012.

[4] Alotaibi, Saqar Moisan F, “Big data analysis role in advancing the various activities of digital libraries: Taibah University Case Study- Saudi Arabia”, International Journal of Computer Science & Network Security, Volume 21 Issue 8, Pages 297-307, 2021 ISSN 1738-7906.

[5] Alokluk, J. A., & Al-Amri, A. “Evaluation of a Digital Library: An Experimental Study”, Journal of Service Science and Management, 14, 96-114. https://doi.org/10.4236/jssm.2021.141007. 2021

[6] Digital library. Wikipedia, Wikimedia Foundation, Access: May 22, 2022, https://en.wikipedia.org/wiki/Digital_library.

[7] Zaka, B., Maurer, H., and Delilovic, N. “Investigating Interaction Activities in Digital Libraries: The Networked Interactive Digital Books Project”, IPSI BgD Transactions on Internet Research Journal, Volume 16 (1) ISSN 1820 - 4503, January 2020.

[8] Maurer, Hermann; Zaka, Bilal¸ Eisenberger, Sonja: “Passively Acquiring Information Must End”, Proc.Eurospi 2021, Springer CCIS 1442, 151-163 (2021).

[9] Higgins, Sarah. “The DCC curation lifecycle model”, International Journal of Digital Curation. 3. 453. 10.1145/1378889.1378998. (2008)

[10] Verma, L., & Kumar, N. “Comparative Analysis of Open Source Digital Library Softwares: A Case Study”, DESIDOC Journal of Library & Information Technology, 38(5), 361-368. https://doi.org/10.14429/djlit.38.5.12425. (2018)

[11] Jain. A. “The 5 V's of big data”, Watson Health Perspectives, 17 September 2016. https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/. Retrieved 04 June 2022.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, DOI: 10.1109/CVPR.2016.91.

[13] Jelodar, H., Wang, Y., Yuan, C. et al. “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey”, Multimed Tools Appl 78, 15169–15211 (2019). https://doi.org/10.1007/s11042-018-6894-4.

Bilal Zaka is an experienced IT professional, academic manager and researcher, presently Head of IT Services at COMSATS University Islamabad Pakistan. He also provides consultancy services to the Higher Education Commission of Pakistan as a member of various technical committees. Besides IT-Telecom consultancy, Bilal worked on software development, network design and security auditing projects at various levels. He did his PhD in Informatics from the Graz University of Technology - Austria, and an MSc in Electronics from Quaid-e-Azam University Islamabad Pakistan. Bilal's professional career spans over 20 years and his work has been acknowledged and appreciated at national and international levels. (e-mail: bilal.zaka@gmail.com)