OPEN SCIENCE, AGILE,
& CULTURAL HERITAGE

JANUARY 2023

Trends and Issues in Library Technology

IFLA IT Section

International Federation

of Library Associations and Institutions

TRENDS & ISSUES IN LIBRARY TECHNOLOGY

IFLA Information Technology Section | Cultural Heritage, Open Science and Agile Issue

Departments

Featured Member Profile

Feature Articles

Conference Review

Editor's Notes

Cultural Heritage, Open Science & Agile

This new issue of Trends and Issues in Library Technology reflects the dynamism and complexity of the present state of library IT, and its continuing rapid development. An excellent suite of international articles range across areas of natural language processing, topic modeling and cultural heritage archives. New research overviews open science, metadata, AI, and agile IT management methodologies. Current ideas and methodologies are presented that are being rapidly applied in libraries globally. This extends from new project management technologies to new infrastructure possibilities and many new projects to reflect upon, possibly implement, or join. I welcome you to explore and read and review the amazing projects presented.

Our issue begins with our IT section chair, Edmund Balnaves, outlining the IT section's ambitious goals for 2023. Successful projects from the recent past are also discussed, notably our lauded AI Satellite Conference at the National University of Ireland in Galway. These successes continue to build with a further library AI satellite conference organized and sponsored by our AI special interest group and hosted by Singapore's national library. Please read more about these with a review of the previous Satellite by Dr. Lynn Kleinveldt and note about the current from organizers Patrick Cher and Dr Balnaves.

Our Information Coordinator, Francois-Xavier Boffy, advances a fascinating new project involving IT credentialling and open badges to better identify skillsets of our IT members for IFLA's different global regions. If you have interest, please review Francois Xavier's overview, and join this compelling group. This IT badging prototype project will be discussed further at our upcoming conference in Rotterdam (2023) for possible launch and then, hopefully IFLA application during the 2024 year.

We also have a fascinating self-portrait from Souleymane Sogoba of Mali regarding 'A Day in the Life of an Information Manager in Africa.' Souleymane is the Documentation and Communication Officer for the Mali Rural Development Sector Planning and Statistics Unit and outlines important developmental roles for his library and association with Mali's government, the United Nations Food and Agricultural Organization (FAO) and the European Union (EU). In his portrait, Souleymane focuses on his unit's fundamental work and their project in developing Mali's Digital Archives and the specific unit's role in these efforts.

Cultural Heritage, Open Science and Agile Methodologies

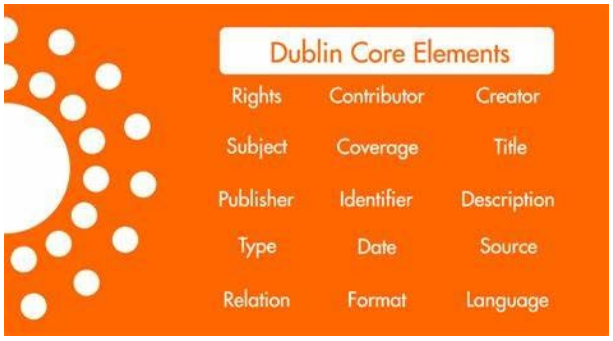
For our feature IT R&D articles, this issue opens with Sümeyye Akça from the Department of Information and Records Management, Marmara University, Istanbul, Turkey. Dr. Akça outlines an intriguing project involving natural language processing, topic modeling and the Kadi registers.

The Kadi registers are a large archive of legal administrative documents written in Turkish, Arabic and Persian that shed important light on the history and socio-economic life of the Ottoman empire from the 15th C. to the early 20th. Akça utilizes innovative natural language processing and a Latent Dirichlet Allocation model for the optimization of topic clusters and optimizing algorithms for search and retrieval.

In her article, Sümeyye outlines the fascinating optimization process and natural language/data visualization methodologies to create topic headings and map both cognitive distances between headings and topic clusters. It seems there is room for further application and generalization of the model's large, correlated text archives for primary cultural heritage research archives and new digital library primary research historical investigation. With parallel recent large natural language AI model progress such as Open AI's ChatGPT, there seems much fertile territory in this

area for further exploration for libraries and research in general.

Continuing this data and libraries' trend, Juan Miguel Palma Pena, of the National Autonomous University of Mexico, shifts from cultural heritage and language to data and Open Science. In his research, Dr. Palma Pena focuses upon metadata schema analysis and research data repositories for Open Science. In focusing upon Research Data Repositories, Dr. Palma Pena analyzes differences and commonalities of current metadata schemas ranging from Dublin Core to OpenAIRE to the Research Data Alliance (RDA). Juan Miguel's work looks further into how these metadata schemas logically organize and structure our data to make research more findable, accessible, interoperable, and reusable.



My own research continues these Open Science and data repository perspectives. I trace an example of current experimental data's workflows across global distances through IR and data repositories, foregrounding the progress of AI discovery through digital libraries. The research lifecycle illustrated here moves from datasets and metadata to online data research repositories and new AI discovery. Our academic libraries' digital scholarly research ecosystems are central to today's open science, AI needs and possibilities. These open science processes move forward daily through the enabling power of networked information, faster processing power and our libraries' online data repositories enabling capacities for powerful search and retrieval globally.

Following this, Wouter Klapwijk, Director of Library Technology Services from Stellenbosch University, South Africa, overviews agile project management methods in an academic library setting. Klapwijk focuses his explicatory work on the Agile Manifesto, Atlassian's Jira Service Desk System and the Japanese Kanban Agile conceptual framework. The conceptual division of projects into 'To Do, In Progress and Done' to streamline workflows and complete projects is crucial in today's complex library IT projects. Klapwijk, advocates Kanban as a Zen-type methodology that allows a library's IT manager to better oversee complex IT projects, open bottlenecks and reassign workflows to complete challenging projects in an 'agile', efficient manner.

In sum, we've got an excellent issue with provocative articles, applications and cutting-edge project management examples. Send us a note if you have been spurred to action by one of our authors' projects. We are always looking for innovation, new efficiencies and author's articles. Thanks also to Patrick Cher and Wouter Klapwijk for help with graphic design and editing.

Finally, to our members, I would like to extend best wishes for a productive and creative upcoming year. There are many new possibilities for our area of libraries in the year ahead. I wish all of you good fortune with all your projects, physical and digital.

Kind regards,

Ray

Ray Uzwyshyn, Ph.D. MBA MLIS
rju13@msstate.edu
Editor, Trends and Issues in Library Technology

IT Section Chair's Corner

Message from the Chair

The Information Technology section passed significant milestones in 2022 as Artificial intelligence has been a prevailing theme in information technology, and no less so in libraries. The emergence of AI in many aspects of government, work and social engagement made it important to establish a forum for this area in IFLA for discussion on the implications for libraries and their clients. Dr Lynn Kleinveldt continued our series of webinars throughout the year, including a focus on AI.

In 2022 we oversaw the formation of the new Artificial Intelligence Special Interest Group, chaired by Dr Andrew Cox from University of Sheffield, England. Artificial Intelligence has been a significant theme for the section, with the hosting of a satellite conference "New Horizons in AI" at the University of Ireland, Galway, coordinated by Edmund Balnaves and May Chang. The conference was well attended and had an exciting set of papers covering all aspects of artificial intelligence. Themes of ethics in AI as well as many practical examples of implementation were discussed. Presentations from this satellite conference can be found at https://www.universityofgalway.ie/ifla/abstracts/.
WLIC 2022 saw a continuation of this engagement with the first SIG meeting and the formal establishment of the SIG. We also had a town-hall style open discussion on the topic of AI that was standing-room only.

WLIC 2022 also contained other excellent sessions sponsored by the IT section. Elena Sanchez coordinated a session on "Agile in the Library: methods and tools for project management, collaboration and innovation" in conjunction with the Subject Analysis and Access section. The utility of Agile in the library environment made it likely we will pursue further sessions on this theme. The National Library of Finland has long had Agile development central to its library strategy.

In conjunction with the Big Data Section, we also had a session on libraries at scale: "Perspectives on data access and use at scale: Lessons from the field," coordinated by Cory Lampert, Patrick Cher and Alenka Kavčič-Čolić. This explored various aspects of scalable digital library implementations with a particular focus on metadata at scale.

This decade has seen many international challenges. The emergence of war in Europe, the challenges of climate change and the continuing struggle against Covid have put great stress on all aspects of international cooperation. Libraries continue to leverage their resources on the ground and online to support communities challenged by these international trends. Libraries have leveraged technology to be versatile in continued delivery of services despite severe difficulties. The IT section and this publication provide a valuable forum for discussion of the directions in information technology amidst the growing uncertainties that frame the international environment in which libraries operate.

Regards
Edmund

Edmund Balnaves, Ph.D.,
ejb@prosentient.com.au
Chair, IFLA IT Section

IT Related Skills Open Badges and an IFLA Prototype Project: IT and beyond

Francois Xavier Boffy, francois-xavier.boffy@univ-lyon1.fr
Head of IT Department, Claude Bernard - Academic Library
Université de Lyon, France

IFLA is a network of information services and library professionals with various large and wide ranges of skillsets. Within the different sections and workgroups, expertise, skills, and specific knowledge add value to our meetings organized under IFLA's main umbrella. The general organization may also increase chances to find various appropriate experts for a domain searching within its ranks or a specifical geographical range, but because of the number of members and skillsets we need a better sort of compass, or at least better guide points from both colleagues and references.

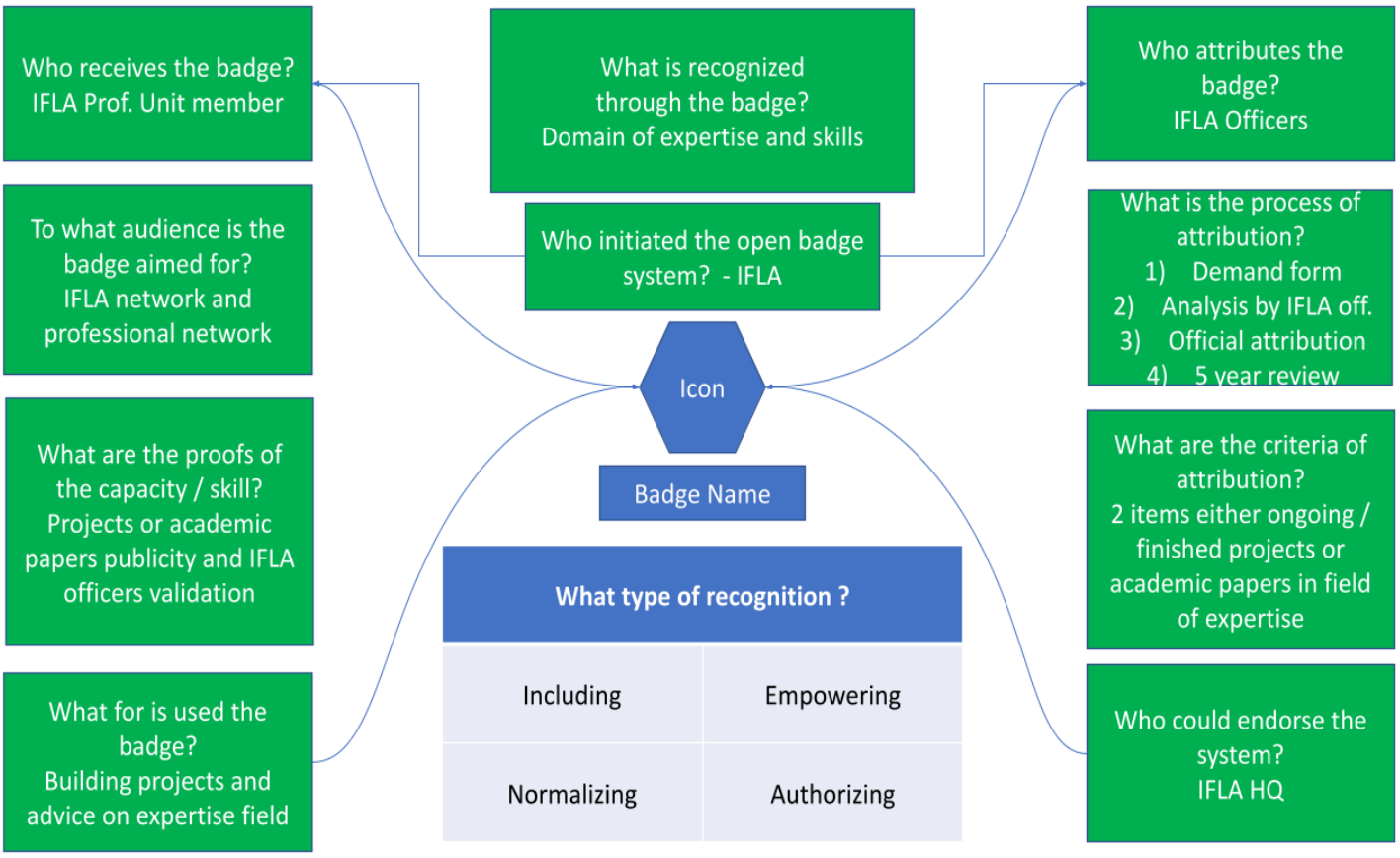
Specific IT topics and skills include by nature a certain level of technical expertise needed. IT activities, including implementation and development, commonly face the question of finding appropriate skill levels for projects or services ranging from programming and interface design to metadata application and project management. Accordingly, the need for skills visibility may be higher in our IFLA IT Section than in other professional units where soft skills dominate. As discussions with colleagues from IFLA Headquarters have currently been ramping up around the need for a general "skills registry," the IT Section is naturally interested because of the more specific technical skills and intrinsic nature of IT.

Even if no concrete online service has previously been built, the analysis of current new possibilities has opened discussions towards this type of project. One of the major ideas now is to create a new Open Badge system for IFLA. This relies on existing technical possibilities that could be ideally utilized and reused on the different IFLA communication channels.

On the public side, this open badge system could be implemented on the IFLA website. On the back office (intranet) side, this could be used on the IFLA section

workflow and project management system and within the current Basecamp project management system workspaces. Proof of concept for this type of new system could be made this year (2022-2023) within the purview of the IT professional units. This would extend to the IT Section and section sponsored Special Interest Groups including the AI Special Interest Group and Big Data Special Interest Group. If successful, the system prototype could be extended in 2023-2024 to the whole organization following its evaluation and assessment. Further collaboration with all professional units will be crucial to providing the most appropriate system. Early in the project several members of IFLA's Continuing Professional Development and Workplace Learning Section also expressed interest to participate in this project as a second phase of implementation and/or parallel prototype.

In the schema of an IFLA Open Badges system outlined in conceptual design below, we can already define many aspects (See Green Sections in Particular):



Inspired by Serge Ravet model - Reconnaitre ORA

Some solutions naturally give rise to the open badge skills model within the IFLA IT context. As a section example, the list of achievable IT skills can be maintained and developed by existing IFLA professional units' officers or by a newly created position role within specific sections of "Open Badges Directors." Of course, we can borrow from more well-established online skills models such as the UK's Chartered Institute of Library and Information Professionals (CILIP) and their Professional Knowledge and Skills Base. See here for their example: PKSB.

Beginning to develop and refine the open badge model prototype and discuss further details of the process and its viability and sustainability will be priorities for the next months for the Open Badge IFLA working group. Ideally, a proof concept will be presented to IFLA community during the next WLIC in Rotterdam, Netherlands. With these developments in mind, we welcome you to join us in this exciting project. Please consider lending your IT members your expertise and stay tuned for more on this exciting project at our upcoming conferences in 2023 and 2024.

If you have feedback, comments or wish to help and participate in the upcoming development open badges prototypes and projects dealing with information technology topics within the IFLA library professional context, please feel free to contact (francois-xavier.boffy 'at' univ-lyon1.fr).

AI Satellite Conference in Singapore

Edmund Balnaves, ejb@prosentient.com.au
Chair, IFLA IT Section

National Library Building, Singapore

The IFLA Information Technology Section is pleased to announce our satellite conference hosted by National Library of Singapore: AI in Focus: Artificial Intelligence and Libraries - 2 to 3 March 2023.

Following on from our successful satellite conference in Galway, the IFLA Information Technology Section is planning sessions and workshops to provide insight into the role of libraries in the responsible use of AI. We are witnessing the adoption of AI in all aspects of library operations, including public and reference services, automated classification, customer service through chatbots and discovery systems. Vendors are adding elements of AI in library products including discovery and research platforms. The role and impact of AI present both opportunities and challenges.

The program will include key note speakers, workshops and presentations in the practical application and ethical dimensions of AI in libraries. This is a great opportunity to discover more about a fast-developing area for libraries.

This will be a hybrid conference with in-person and streamed sessions, and virtual workshops.

You can find details on the call for papers and the venue at https://ifla.nlb.gov.sg/.

We will make a further announcement as the programme is released and registrations are opened.

Above all, save the date and join us on 2nd to 3rd March 2023!

A Day in the Life of a Document Information Manager

Souleymane Sogoba, sogoba.souleymane@gmail.com
Documentation and Communication Officer, Planning and Statistics Unit
Sustainable Development Sector, Mali

Souleymane Sogoba, D&C Officer, PSUSDC, Mali

Souleymane Sogoba is in charge of Documentation and Communication at the Planning and Statistics Unit of the Sustainable Development Sector in Mali, Africa. He is a holder of a Master I in Information and Library Sciences and since 2018 he has been working in library related activities ranging from archival documentation to the daily management of the Documentation Center. He is the only graduate in information sciences in his unit and hence is responsible for all the library-related activities of the Center.

Range of Professional Activities

In the world of information sciences, Souleymane is also involved at the international level. He is active in the Information Technology Section of **IFLA** and in charge of the activities of the sections of the Association International Francophone des Bibliothécaires et Documentalistes (**AIFBD**) being member of their Board of Directors.

aifbd
Association Internationale Francophone des Bibliothécaires et Documentalistes

At the regional level, Souleymane is an active member of the Board of Directors of the African Library and Information Institutions Association (**AFLIA**). On behalf of **AFLIA**, Souleymane leads the **ALVA** Project which aims to map all types of libraries in Africa through data collection.

African Library & Information Associations & Institutions

In Mali, Souleymane is Head of the Partnership Section of the Malian Network of Documentation for Development (**REMADOC**).

Over time, he has accumulated years of experience in the field of libraries, archives, documentation, communication and a passion for Open Access to be part today of the Associate Editors of Directory of Open Access Journals (**DOAJ**).

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

The Mali Rural Development Sector Planning and Statistics Unit

The Planning and Statistics Unit of the Rural Development Sector (CPS/SDR) was created by the Law N°07-020 of February 27, 2007 and the decree N°07-166/P-RM of May 28, 2007. Its missions are as follows:

- coordinate the preparation of plans, programs and projects as well as policy and strategy analysis
- monitor and evaluate sectoral development plans, programs and projects and ensure intra-sectoral and spatial consistency;
- prepare forecasts and monitor environments and the economic situations
- monitor files relating to financing and technical cooperation;
- coordinate, in conjunction with services

responsible for human resources training programs in planning and statistics;
- coordinate the production of statistical information and the realization of studies.

The CPS/SDR is headed by a Director assisted by a Deputy Director. It includes four (4) Units and one Documentation and Communication Center. Each of the units has its own objectives to achieve, in line with the missions of the structure. Thus, the Documentation and Communication Center contributes to the achievement of the assigned missions.

The Center was created in 1992 with the support of Technical and Financial Partners (**TFPs**) such as the **FAO** and the European Union (**EU**), and is responsible for collecting, processing and disseminating documentation and information on the development of the Rural Development Sector (**RDS**). It is also in charge of the elaboration and implementation of a Communication Strategy on the performances and the major constraints of the development of the Sector.

Food and Agriculture Organization of the United Nations

Finally, the Center creates and animates the documentary network of the Sector.

In the Documentation and Communication Center, the tasks are routine and classic for information and library science, namely: physical and material treatment of documents, intellectual treatment of documents, service to users, daily management of mail and information dissemination of information through corresponding channels. These points concern both the management of the Archives and the Documentation Center. In most Malian library related administrations, the archives and the documentation centers (libraries) are grouped together.

Larger Projects – Digital Archives, Digitization and Archiving

Since 2011 the Mali Rural Development Sector Planning and Statistics Unit has initiated a project of physical and digital archiving of documents.

The overall objective of this project is to develop an archiving and digitization plan for the archives of the Planning and Statistics Unit of the Rural Development Sector in order to provide it with efficient physical and electronic archive management tools that guarantee better document management.

Mali: Food and Nutrition Enhancement, Koulikoro Region Food and Nutrition Security Project

National Archives and Library of Mali

The management of archives is a sensitive issue that requires the implementation of a policy ensuring the reliability and durability of the administrative memory necessary to facilitate and support the public service mission. To this end, it is therefore necessary to design and implement a regulatory framework defining the methods and procedures for processing, preserving and communicating each type of document according to its own characteristics (type, nature, medium, frequency of use, etc.).

The elaboration of such a framework is an obligation prescribed by the Decree N°02-424/P-RM of 9 September 2002 fixing the modalities of application of the Law N°02-052 of 22 July 2002 relating to the Archives. It is in this context that a study for the betterment of an archiving and digitization plan for the archives of the Documentation and Communication Center, and a support to the archiving and digitization of documents, has been financed by the European Union (EU) through its technical assistance to the CPS/SDR.

This digital archiving project is based on the Mali Rural Development Sector Planning and Statistics Unit's project to set up a Data Center. This digital documentation and communication center involve the integration of electronic document management (**EDM**) and digital archiving that will help in the endorsement, safeguarding and capitalization of the RDC's information heritage. To put in perspective, the Documentation and Communication Center will set up a digital library to facilitate access to free and paying electronic scientific and technical resources. A documentary database will be set up through the Data Center for the storage and dissemination of data. Services and products provided by the Documentation and Communication Center include:

- On-site consultation;
- Loan of documents;
- Inter-unit loan of information;
- Referral to units and other structures;
- Selective dissemination of information;
- Information monitoring;
- Animation of professional groups on social media;
- Question and answer service.

The Mali Rural Development Sector Planning and Statistics Unit is hopeful with all of this new innovation, infrastructure and digital information, that the future will be better for the documentation and communication center through the multiplication of innovative initiatives and improvement of the IT master plan and the IT system in place. All of this together bodes well for the next generation of development and information services for Mali.

Topic Modeling on the Kadi Registers Database

Dr. Sümeyye Akça, sumeyyesakca@gmail.com
Assistant Professor, Department of Information and Records Management
Marmara University, Turkey

Image 1 – Kadi Registers, Istanbul

Kadi legal registers shed light on social life in the Ottoman Empire. They are one of the most important sources for today's historians working in these areas (Akça, 2005). These legal and administrative documents were written in Turkish, Arabic, and Persian and maintained in book format in different courts operating in Istanbul from the 15th century onwards. The books were maintained and updated from the second half of the 15th century through the first quarter of the 20th century. They are a key source of Turkish culture and history, closely linked to Turkish economic and political life. With the understanding of the wealth of content contained in the Registers, over the last 50 years many researchers in Turkey have been investigating and transcribing them to the Latin alphabet. Each book contains records for every unique event in this history. The registers give us clues to the legacy structure of the Empire. Although to date there has been significant effort to transcribe them, some books still await transcription into the Latin alphabet.

Introduction to Topic Modeling of the Registers

Creating subject authority files in this field has been considered valuable in terms of making these notebooks available to wider audiences through an effective database. It is important to define and catalog resources according to subject authority files in order to provide effective results during information retrieval. Subject authority files' purpose is to reach information that the user is seeking in various ways (Gültekin, 2020, p. 47). This study therefore aims to create the authority records that will be used to define these sources to facilitate access.

A topic is assigned by the transcriber to each record (register) in a book transcribed into the Latin alphabet. The subject search in the indexed books is presented to researchers through the relevant record. A record may contain more than one topic or, in some cases, a single topic may be covered in more than one record. Topics can be used to better understand and read the Kadi registers. This ranges from which topic each record is to which words are frequently used in interrelated topic clusters. Topic clusters also offer researchers a broader

set of topics, with access to records on related topics for researchers searching in related subject areas. It is important that these resources are used effectively as they have a significant historical value.

Literature Review

Studies in which computer or computational methods are used on historical data and documents have recently been carried out in order to make data easier to understand and interpret. Grant et al. (2021) studied topic modeling of the archive on refugees from the south within the UK and US governments and the United Nations High Commissioner for Refugees (UNHCR) between 1972 and 1979. In this study, researchers analyzed the role of various state and non-state actors on this material and their discourses, as well as the results of resettlement programmes.

In another study, Yang, Torget, and Mihalcea (2011) applied topic modeling on newspapers published in Texas, USA, between 1829 and 2008. This study brought to the fore the difficulty encountered by historians working on large numbers of newspapers transferred to the digital environment. Topics discussed in the country's agenda and the changes in these issues over the years could be more easily followed with the addition of topic modeling within digitization projects.

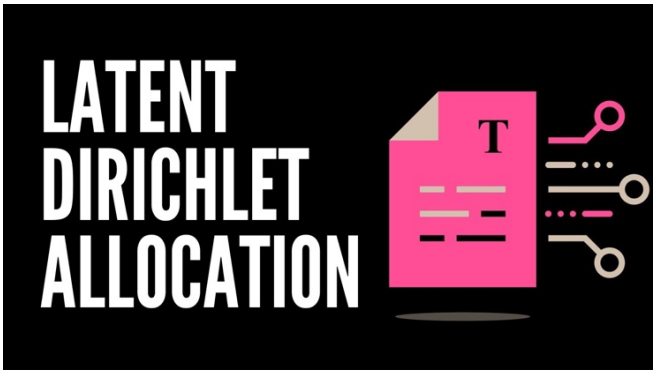
Schöch (2021) worked on topic modeling on French drama texts between the years 1610-1810. He examined whether different dramatic genres have distinctive dominant themes or dramatic sub-genres have their own specific plot patterns. The study also looked at the extent to which topic clustering and classification methods based on subject scores produced results consistent with traditional classification distinctions.

Image 2 – Library of Seyhulislam made by Sultan Abudhamid II for Kadi Registers

This present study on Kadi registers aims to automatically assign the topic of each record in the Kadi registers using a topic modeling method, and thereby revealing the distinctive dominant word groups within each topic. The study examines the common word groups used for each topic, concomitantly comparing the surfaced words groups with the topics tagged during transcription.

Dataset and Research Method

Topic Modeling is an unsupervised machine learning method to find groups of words called "topics" in a text document. These topics are made up of words that often occur together. They often share a common theme. Therefore, the topics with a predefined set of words can be used as phrases to best describe the entire document. In order to improve the identification of the subject of each record, Latent Dirichlet Allocation (LDA) is run on the text. LDA is an example of topic modeling where each document is considered a collection of topics, and each word in the document corresponds to one of the topics. For this purpose, the application of LDA to the text is done utilizing Python programming software.



In order to run the topic modeling algorithm, the transcribed Kadi register book of Üsküdar (1561-1563) was used. This book contains important information about the social life of Üsküdar in those years. The register consists of court records (huccets) such as purchase and sale transactions, fugitive slaves, and murder cases (Akça, 2015, p. 9). There is a total of 648 individual court records in the book. In the transcribed version of the book, a classification has been made about the topic of each court record. There is an index consisting of subject, person, event and place information. The topic modeling algorithm was run on the text in order to machine-classify the topics of the court records in the text and to compare them with the topics tagged during transcription.

The study began with importing the data and applying pre-processing tools on the text. The `initial_clean` function is used to clean the punctuation marks on the text and change uppercase letters to lowercase. In addition, `word_tokenize` is used to set words and symbols according to spaces. The `remove_stop_words` function is used to remove the most commonly used words. The open-source Gensim library is used to create the corpus and a dictionary. Gensim is an open-source library for unsupervised topic modeling and Natural Language Processing (NLP) using statistics-based machine learning.

In order to apply topic modeling, the number of topics must first be determined. There are many approaches currently available for this task. These approaches generally look at the distributions of LDA such as subject-term, document-subject, calculating the distances between pairs of topics and determining the most appropriate number of topics (Akbulut, 2022, p. 29).

Four metrics are used to determine the most appropriate number of topics for the LDA algorithm to be applied on the Kadi registers. These metrics include statistical calculations based on topic, and document calculations on the text. Each metric is run in a R-project statistical analysis environment over a single code. As a result, the most suitable number of topics for the model to work on the data set was determined to be 30 (See *Figure 1*).

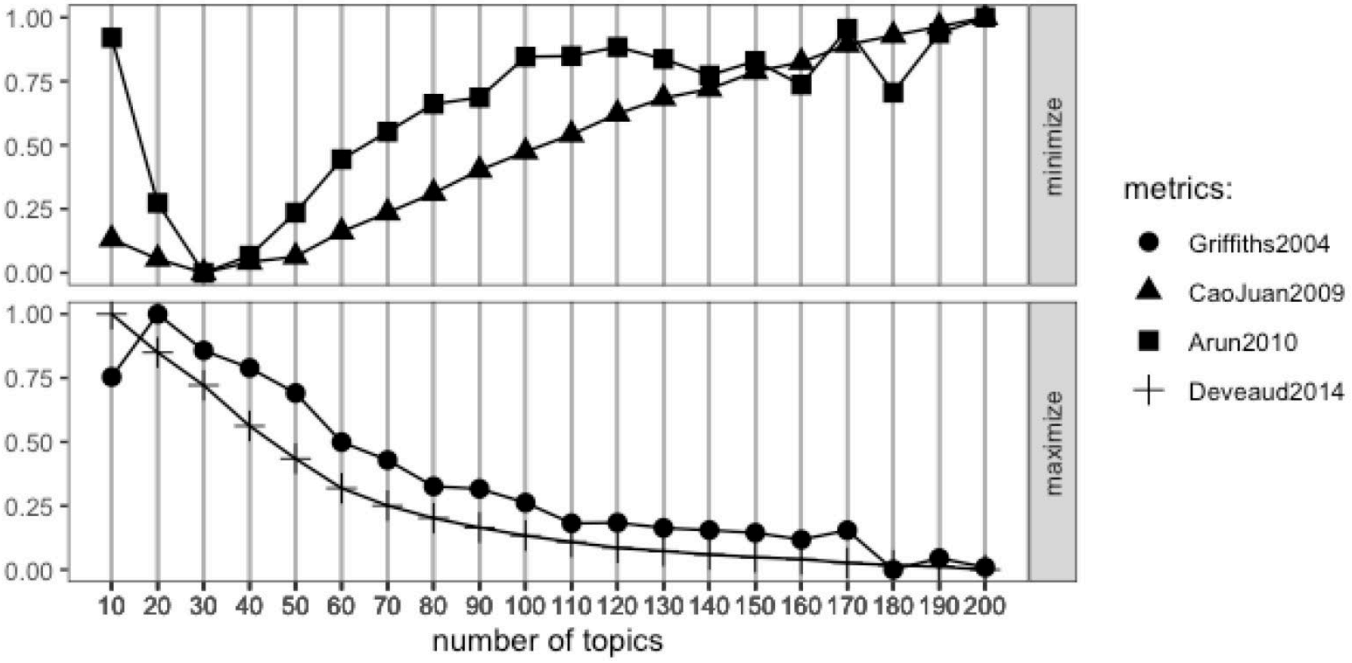
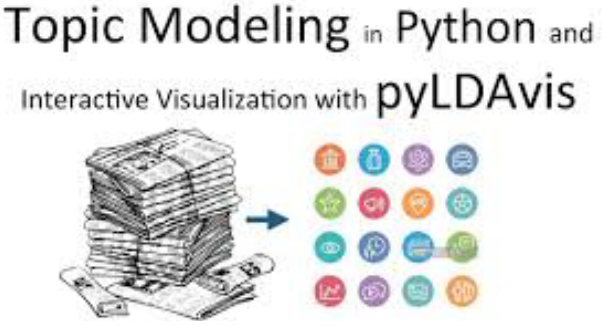


Figure 1 – Four metrics used to determine the optimal number of topics for the LDA model

Then the topic models were created and the results analyzed. PyLDAvis is used to visualize the results. PyLDAvis is designed to help users interpret topics in a topic model as appropriate to a collection of text data. It provides an interactive web-based visualization with information from the LDA topic model.

Finally, an evaluation was made to determine which topic is dominant for each record. To get a better idea and to validate the results, it created an output to give the most appropriate topic for each text data and we can easily see the topic ratios with their keywords. Through this model, a range of "topics" – recurring themes – and the extent to which each document addresses these issues, has been explored on the Kadi registers.



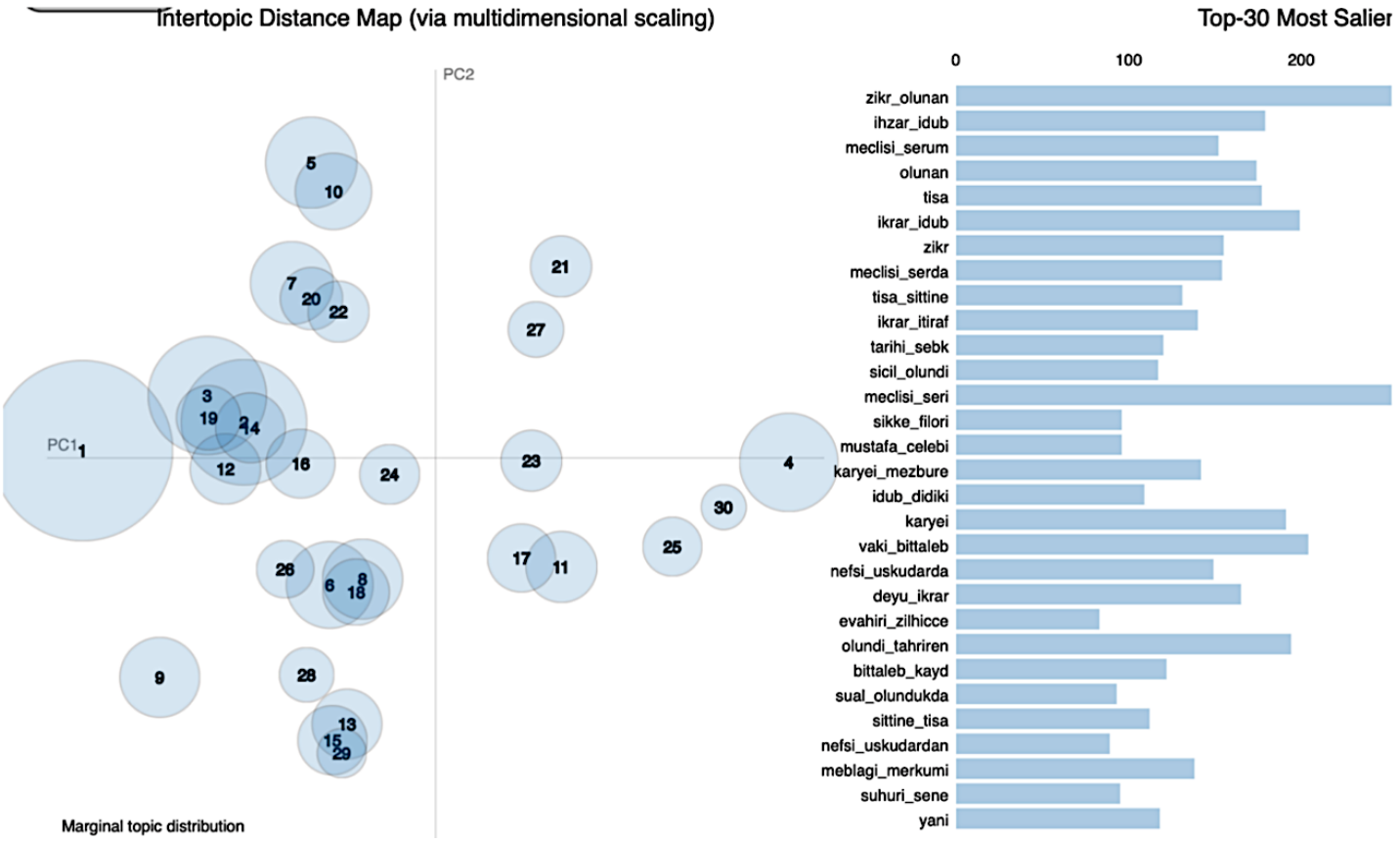


Figure 2 – LDA algorithm on the Kadi registers representing all clusters

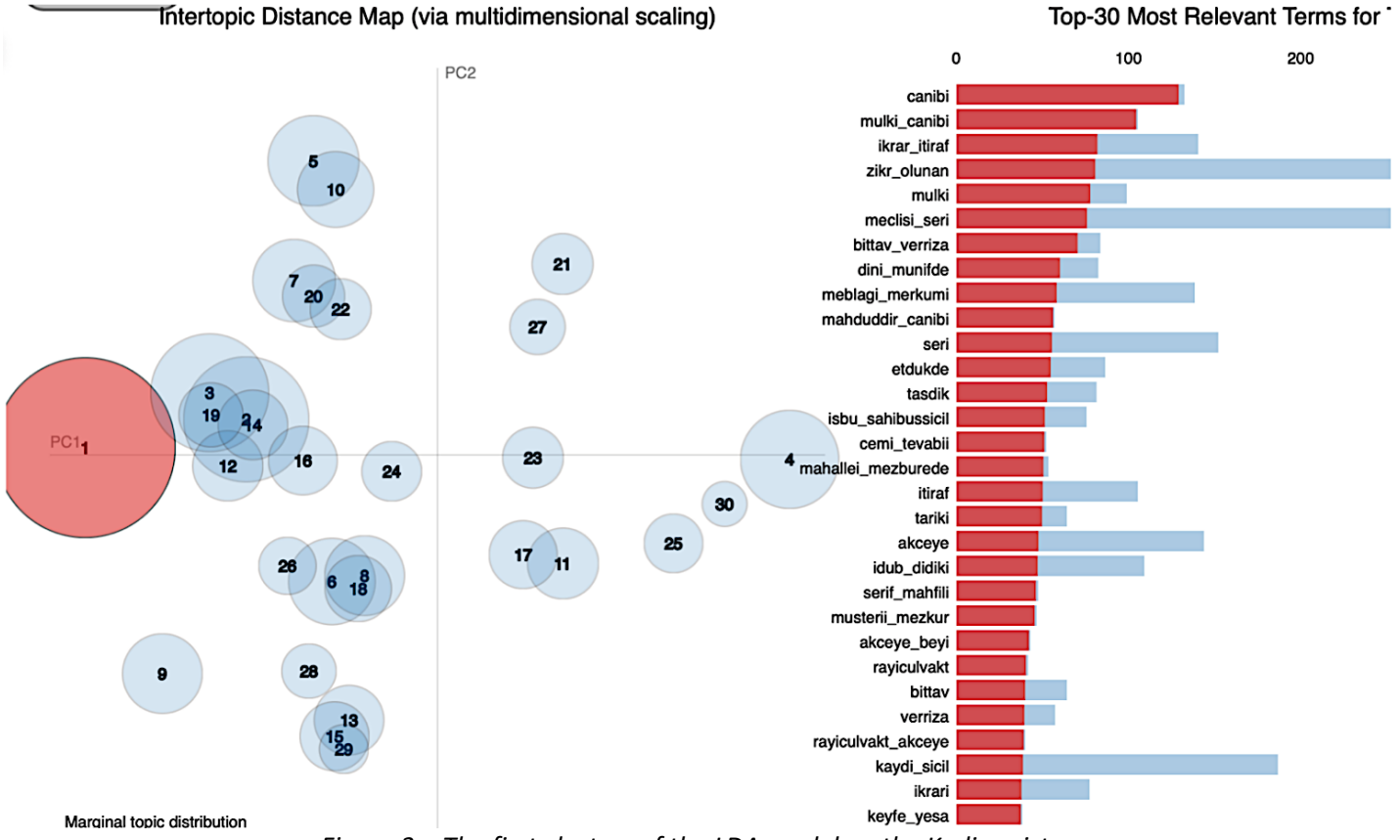


Figure 3 – The first clusters of the LDA model on the Kadi registers

Findings and Discussion

The phrases used in the Kadi registers and the most frequently used words is easily viewable in the graphic created by the LDA model. Indicators of time and place names, and the name of the office where the court is seen, are among the most used words in the data set (See *Figure 2*).

The first set of the LDA model shows that there are the most buying-selling (bey') cases in the data set. The frequency of the words used in the first cluster in the whole document (red and blue) can be easily observed in *Figure 3*. Looking at the previous topic tagging, buying-selling cases constitute approximately 16% (101 numbers) of the total individual cases.

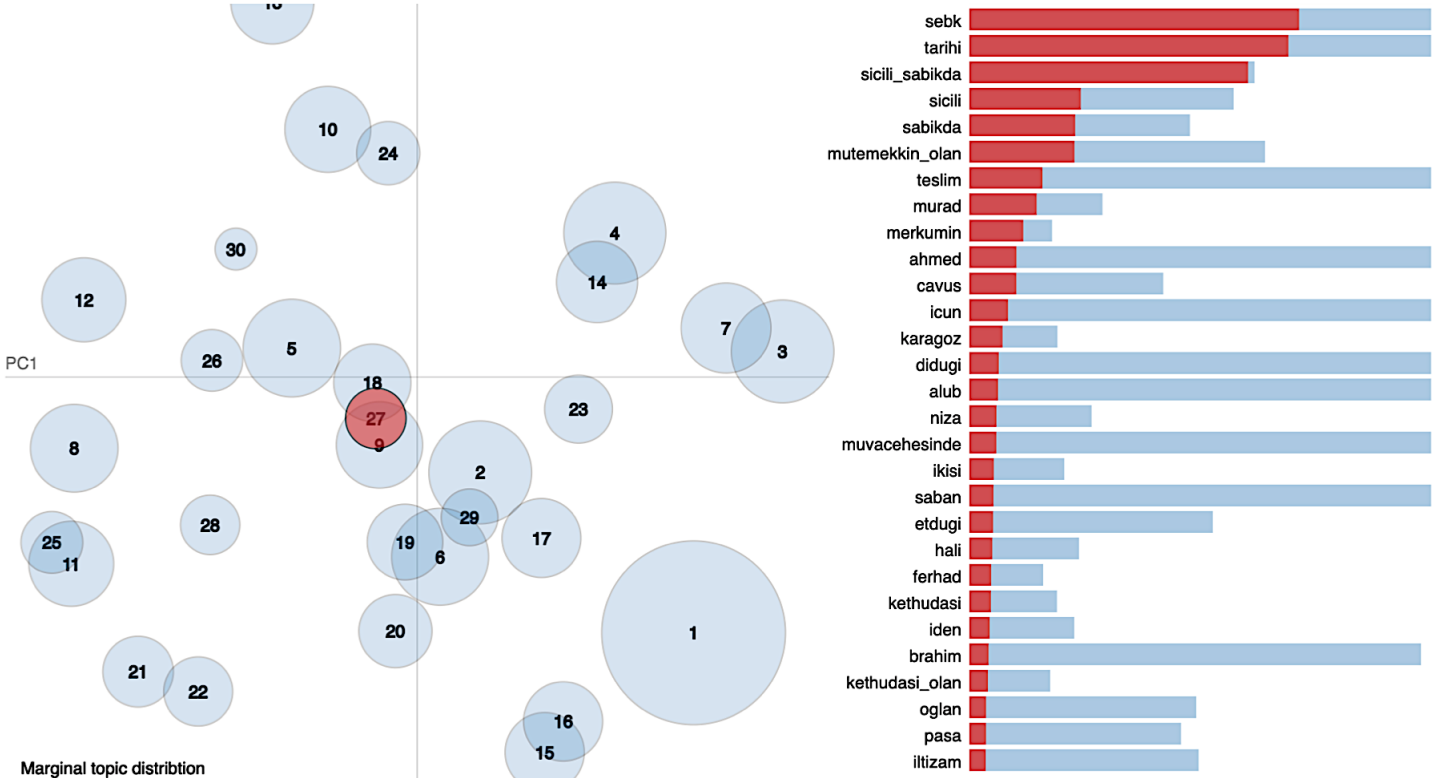


Figure 4. The 27th cluster of the LDA model on the Kadi registers.

In *Figure 4*, the words used in the topic cluster, which are rarely mentioned in the whole data set, are seen. In a detailed reading, the topics of tax farming, fugitive slaves and hostility *(niza')* in the records of the Kadi registers were gathered in a cluster. This result shows that the words used in these topics are related or these topics can be included together under one record. In the LDA model figure, it can be seen that as the cluster circles get smaller, the frequency of use of the words under this cluster in the entire data set has a distinctive feature (See *Figure 4*). The topic labeling made by the author on the kadi register also confirms these results. According to this, tax farming is mentioned 21 times and the issue of fugitive slaves 40 times in the Book (Akça, 2005). The most distinctive issue in the data set is seen as the surety *(kefil)* cases. Since the words used in surety cases are generally names (foreign names - non-Muslim), they have formed a separate cluster as words that are rarely used in the entire document universe (See *Figure 5*). Also, for a closer look at the LDA model and to see the general situation of each case, the TSN-E algorithm was applied on the data set. As can be seen in *Figure 6*, unlike the LDA, the status of each case is visualized in two dimensions according to its topic. For this, it used the topic information of each case which was previously determined. One of the important points here is that we can pursue which subjects are different, which have a unique meaning in the data set, and which subjects are similar to each other. In this image, it is seen which subjects converge with each other on the basis of easily used words. The words used in buying-selling *(bey')*, debt *(deyn)*, pricing *(narh)*, surety-deputation and alcohol cases are seen as related words.

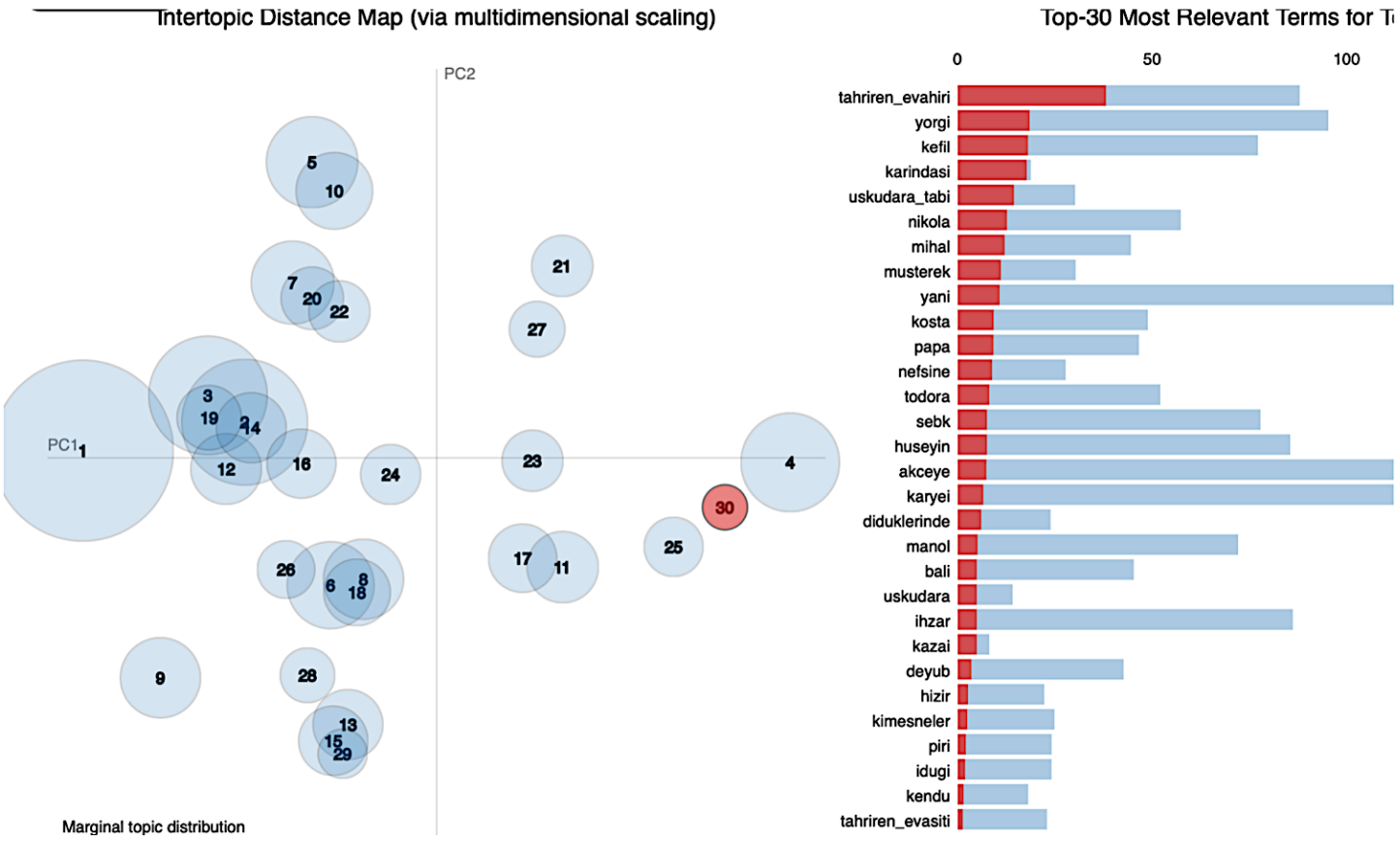


Figure 5 – The 30th cluster of the LDA model on the Kadi registers

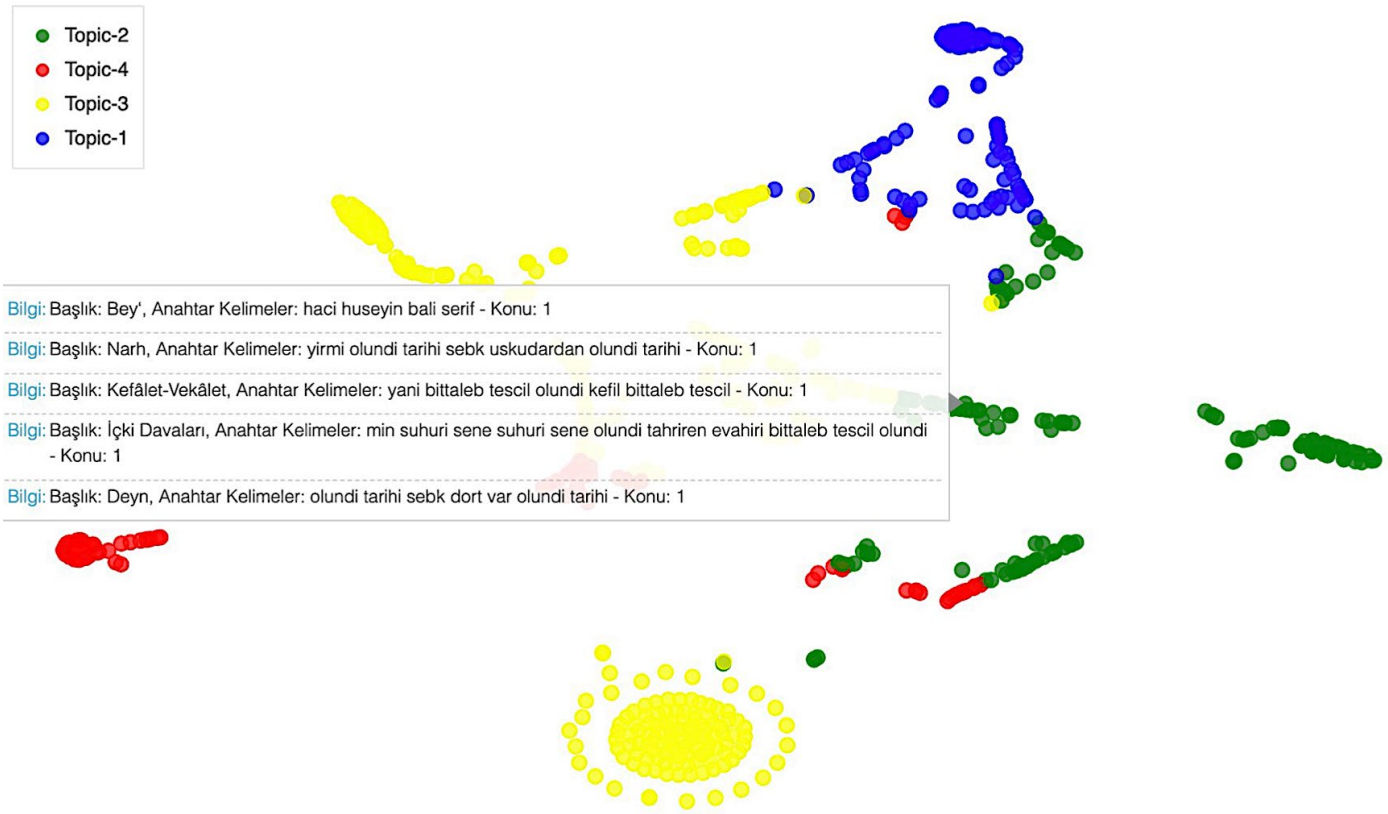


Figure 6 – TSN-E algorithm on the dataset

Conclusion

Applying Latent Dirichlet Allocation (LDA) topic modeling on historical texts clearly is a method that contributes to users seeing results in a more effective and contextual way for information retrieval. Growing the context and access to historical records through computer methods is important in scientific circles as well as socially and culturally. The results of this study can be used for effective search-result retrieval in the Kadi Registers database.

Acknowledgement

I would like to thank Dr Müge Akbulut for her support in the study.

References

Akbulut, M. (2022). Bilgi Erişimde İlgi Sıralamalarının Artırımlı Olarak Geliştirilmesi (Incremental Refinement of Relevance Rankings in Information Retrieval). Unpublished doctoral dissertation, Hacettepe Üniversitesi, Ankara.

Akça, S. (2005). Üsküdar Kadılığı 23 nolu ve h. 968-970 tarihli sicilin diplomatik yönden incelenmesi: Metin ve inceleme. Unpublished Master's thesis, Marmara University, İstanbul.

Grant, P., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic modelling on archive documents from the 1970s: global policies on refugees. Digital Scholarship in the Humanities, 36(4), 886-904.

Gültekin, V. (2020). Konu otorite dizini Nedir? Nasıl oluşturulur? Türk Kütüphaneciliği (Turkish Librarianship), 34(1), 46-64.

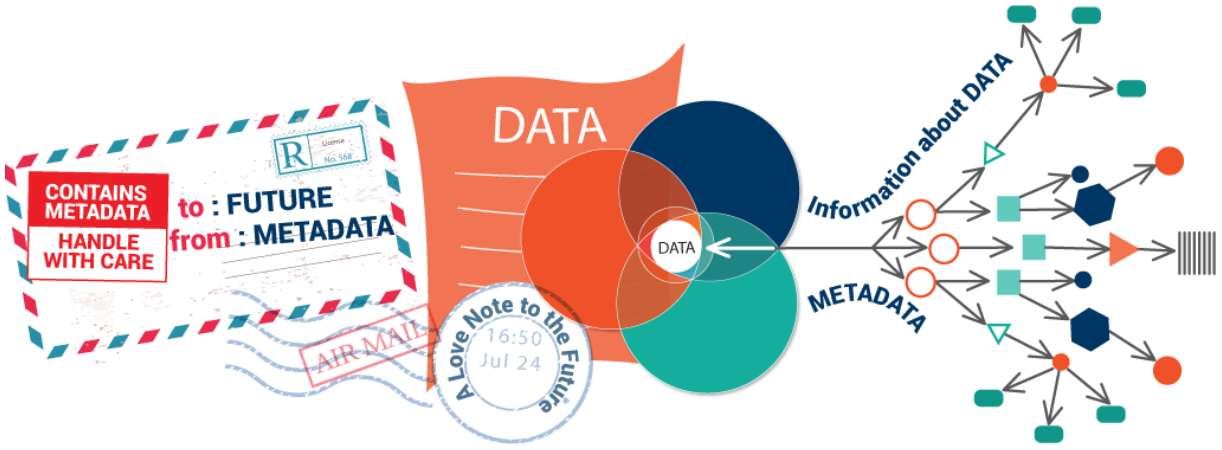
Schöch, C. (2021). Topic modeling genre: an exploration of french classical and enlightenment drama. arXiv preprint arXiv:2103.13019.

Yang, T. I., Torget, A., & Mihalcea, R. (2011, June). Topic modeling on historical newspapers. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 96-104). https://aclanthology.org/W11-1513.pdf

Image 3 – Istanbul Kadi Registers Published in Print and CD (2010)

Research Data Repositories for Open Science: Metadata Schema Analysis

Dr. Juan Miguel Palma Peña, jemajumi@hotmail.com
Humanities Coordination
National Autonomous University of Mexico, Mexico



What is Metadata? (Source: OntoText)

Introduction

The open access movement has promoted the development and implementation of diversity in institutional repositories. Repositories are the main channels that enable academics towards free access to different academic publications and research outputs such as articles, books, book chapters and conference proceedings.

Currently, there is a need to promote open science through open access to academic publications, as well as research data. Both are used to generate research and subsequently a publication. Openness and unrestricted exchange of research outputs calls for the increased development and implementation of both publication repositories and data repositories. Data repositories are platforms that promote FAIR principles of findable, accessible, interoperable, and reusable datasets and research methods.

Like publications, data in data repositories must be organized with standards and metadata schemas. Various metadata schemas exist for this purpose, such as: OpenAIRE, Research Data Alliance, and Dublin Core. The aim of this research is to present and analyze metadata schemes that are used to organize research data repositories logically and in a structured way, particularly those linked to academic publications to facilitate their findability, accessibility, reusability, and interoperability. This study also includes massive data sets used to generate research, to facilitate its identification, access, retrieval, interoperability, and usability.

Academic communication is in an era of transformation as research societies and communities have highlighted the need to develop and implement unrestricted communication channels to data in the open science environment that is financed with public funds from Higher Education Institutions (HEI) and Research Centers (RC).

Publication Repositories and Data Repositories: Elements and Differences

Publication Repositories

Publication repositories have as a reference the "Berlin Open Access Declaration" (Berlin, 2003), which points towards the main routes for free access to literature financed with public funds. It suggests that open institutional repositories should promote communication, distribution, exchange, circulation, availability, and visibility of resources, with the purpose of preserving content and eradicating access barriers. (Callicot, 2016)

The publication repositories concept is shown in *Image 1*.

Image 1 Publications Repositories Concept

According to new trends in academic communication, publication repositories have a fundamental role for opening and sharing research outputs, which, in turn, need collaboration with data repositories.

Data Repositories

The "Berlin Open Access Declaration" (Berlin, 2003) states that "Not only articles, but also "data and metadata, original materials, digital representations of pictorial and graphic materials, and academic multimedia material should be freely accessible and usable". To that purpose, data repositories encourage barrier-free availability, storage, sharing, and use of research data. (McNeill, 2016)

"platforms that support[s] the preservation, discovery, use, reuse, and manipulation of scientific data objects that support published research. They often provide added value to data through quality assurance and metadata enhancement, and have an operating model based on data harmonization under a common schema." (University of Minnesota, 2020)

Image 2 Data Repositories Concept

Specialized literature suggests that due to trends in academic communication for opening of research outputs in the medium term, HEIs and RCs must develop and implement a repository for data and/or collaborate in one. (Greenberg, 2012)

Based on concepts of publication repositories and data repositories, the premise of this document is to integrate data records and records of academic publications to store and harvest them in institutional repositories, and in this way, enrich the research cycle.

Likewise, it is significant that research data have specific attributes and particularities that must be considered when recording, describing, and relating to attributes of publications in repositories. Since they are independent repositories does not mean that there is no relationship between them. (Hernandez, 2013)

Hence it can be said that creating repositories for open science aims to meet the demands that the funders, actors, and communities of this movement place on the storage and dynamic and free exchange of data, publications, and research outputs. Similarly, giving attention to the recording and standardized description of research data for its indexing and location, as well as advancing the ease and intuitiveness of data recovery by users, is also of importance.

It is therefore pertinent to identify elements that will be treated in data repositories, such as concepts, types, formats, and principles that are needed to organize research data.

Research Data: Concepts, Types, Formats, Principles

In its current context, research data produced in HEIs and RCs are diverse. These are administrative, statistical, governmental, cultural and research data sets. Generally, this requires treating the data in accordance with the needs and trends of scientific fields in which they are produced. The present study addresses such research data that was used to generate an investigation and later a publication in Humanities and Social Sciences; therefore, it is significant to review the concepts of these resources.

Concepts

The main concepts about research data used for this document are presented in *Image 3*.

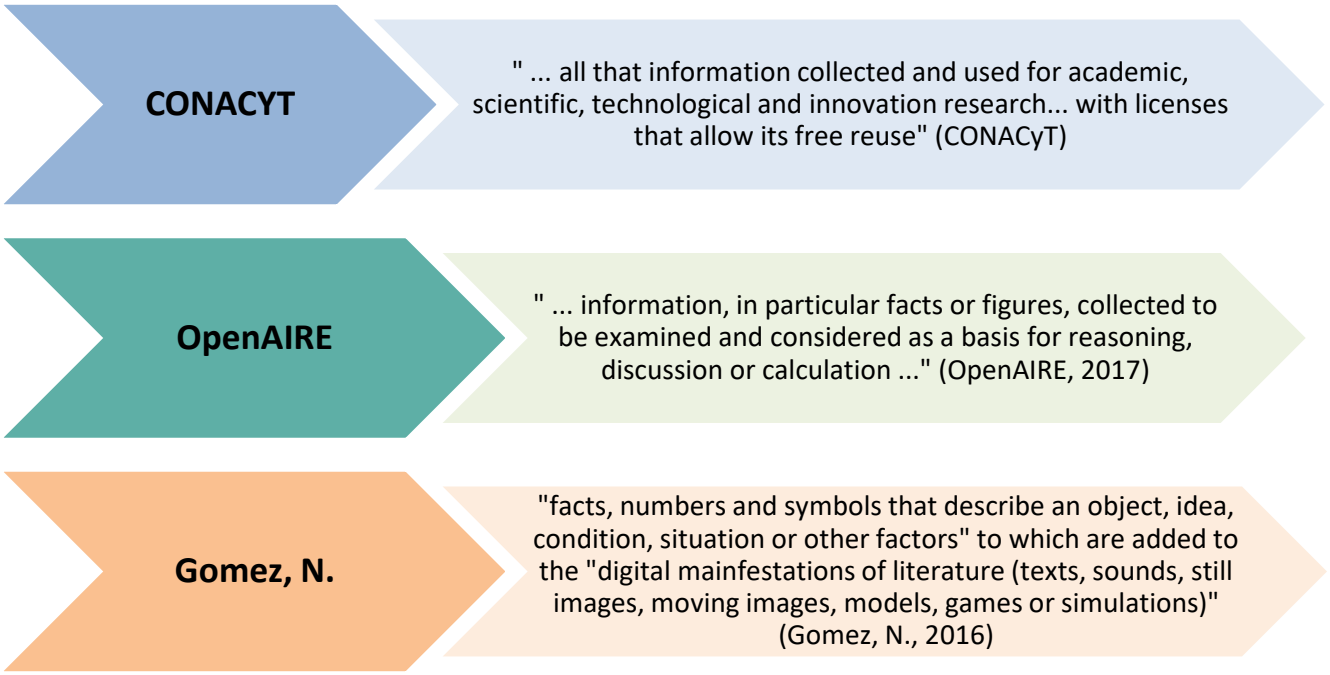


Image 3 – Concepts of Research Data

The concepts listed above provide common elements about research data, such as that they are manifestations in text, sound, images, among others, as well as need to implement licenses and acknowledgment of authorship so that they are available, accessible, interoperable, compatible, and freely reusable.

In this sense, we can define research data as those data that support and validate outputs of original research, which have probably been reviewed by peers for publication, and which must be organized in repositories, this will facilitate access to them, describe them, exchange, reuse, guarantee their transparency and reproducibility in environments in which they are stored, as well as provide information services according to needs and contexts in which they are implemented.

At the same time, a relevant element to identify about research data is its typology.

Typology

Typology of research data is related to different processes of research activity and particular attributes. In this regard, OpenAIRE points out that:

"In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from field work, survey results, interviews, recordings, and images. The focus is on research data that is available in digital form." (European Commission, 2017)

Melero (2014) proposes different types of research data:

1. Numeric data. They are mostly obtained from statistical and mathematical studies.
2. Observational-descriptive data. They include data from sensors, surveys, samples, images, and historical records.
3. Computational data, such as programs and output data, which are reproducible through technological means.
4. Experimental data which accompany experiments from planning, preparation, and implementation, to obtaining results.
5. Derived or compiled data. They include reproducible information for interoperability through text and data mining, compiled databases, among others.
6. Originating and/or reference data. They are obtained from peer-reviewed, published, and selected data, banks, or databases.

It can be defined that types of research data are produced based on purposes and fields of knowledge for which they are used. This is same reason why originating and/or reference data are relevant.
Likewise, a significant particularity of research data to be considered for its recording in repositories are their formats.

Formats

A large part of research data formats is *de facto* objectified in digital form. Some research data formats are shown in *Image 4*.

• Documents (Text, MS Word), spreadsheets • Field notebooks, diaries • Questionnaires, surveys • Experimental data • Movies, audio or video files • Photographs, image files • Collection of digital objects	• Information files • Database content (video, audio, text) • Methodologies and workflows • Content analysis • Artifacts • Models, algorithms • Recordings, interview notes

Image 4 – Research Data Formats
Source: Dewitt Wallace Library; University of Leicester

The analysis of formats of research data are see specifically within the context of HEIs and RCs. Examples are source and/or reference data, which are resources frequently produced and used to support academic publications, such as: books, maps, newspapers, photographs, linguistic studies, movies, object analysis, among other resources. (Gomez, N., 2016).

Furthermore, the contexts of current academic communication demand that research data be aligned with the FAIR principles (Findable, Accessible, Interoperable, Reusable).

FAIR Principles

One of premises of research data repositories is to implement FAIR principles to store, harvest and reproduce data and processes derived from research.

According to GoFAIR and Guajardo, M. (2020), the FAIR principles for research data are showed in *Image 5*.

Linking FAIR principles to the organization processes are based on implementing standards for description and establishing access points and relationships in open access data repositories.

The set of elements analyzed above constitute a theoretical basis to develop and implement research data repositories, and in turn, the need for relevant recording, description, and organization process with standards of attributes and particularities of such resources is highlighted according to their typology and formats.

Likewise, this situation needs to be treated with metadata schemes for research data to integrate functional and useful records that maximize accuracy of location, access, interoperability, and use of such resources.

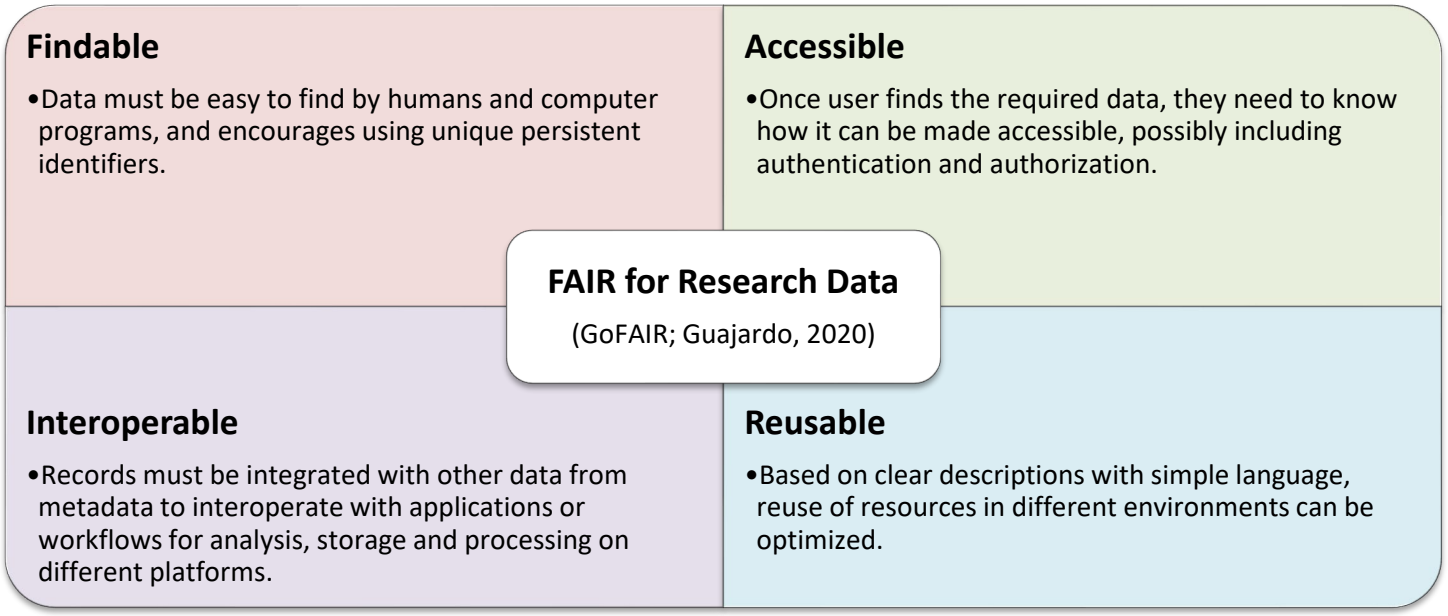


Image 5 – FAIR Principles for Research Data

Metadata Schemas for Research Data: Analysis

In accordance with the transition of academic communication, the organization of research data also requires the implementation of metadata schemas. The schemas are key elements that provide logical and organized structure for the compilation, access, collaboration, and reuse of information resources in repositories, based on principles of organization and FAIR.

This section presents the analysis of different metadata schemes for research data developed by different organizations. A brief description of the schemes is presented, as well as their main characteristics for organizing research data in repositories, as shown by various metadata definition tables.

OpenAIRE Metadata Schema

Open Access Infrastructure for Research in Europe (OpenAIRE) is an instrument of the Horizon 2020 Program, in which protocols and standards are developed for exchange and access to research outputs for a system of open global investigation.

The OpenAIRE scheme is implemented by various international research organizations because it provides an updated matrix of elements to organize information resources, mainly academic publications.

A set of mandatory basic metadata elements for research data (OpenAIRE) recommended by OpenAIRE, are presented in *Image 6*.

OpenAIRE Metadata for Research Data		
	OpenAIRE Metadata	**Metadata Description**
1.	Identifier	Resource Identifier
1.1	IdentifierType	Identifier type (DOI, Handle, URN, URL)
2.	Creator	Author
2.1	CreatorName	Author name
3.	Title	Title
4.	Publisher	Publisher
5.	PublisherYear	Year of publication
6.	Subject	Subject
7.	Contributor	Institution or person that financed development of resource
8.	Date	Date
8.1	DateType	Type of date (publication-distribution; embargo; public date; embargo end)
9.	Language	Language
10.	ResourceType	Resource type
10.1	ResourceTypeGeneral	General type of resource
12.	RelatedIdentifier	Related identifier
12.1	relationedIdentifierType	Related identifier type (Arxiv, ISBN, ISSN)
12.2	RelationType	Relation type (IsCitedBy, Cites, IsPartof, HasPart, IsDocumentedBy, IsCompiledBy)
13.	Size	Size (pages, Megabytes, etc.)
14.	Format	Format (PDF, XML, etc.)
15.	Version	Version
16.	Rights	Rights (level of access: open; closed; embargo)
16.1	RightsURI	URL or URI (of license)
17.	Description	Description of resource (free text, indexed, etc)
17.1	DescriptionType	Type of description (abstract, table of contents, methodology, etc.)

Image 6 – OpenAIRE Metadata for Research Data
Source: OpenAIRE Guidelines for Data Archives

OpenAIRE proposes seventeen basic metadata fields to describe research data, to allow recording persistent identifiers, not only DOIs, in 1.1 "`identifierType`" metadata. Likewise, it recommends using metadata: 7. "`Contributor`" to link funding information; in addition, it recommends using metadata 8.1 "`DateType`"; also, encourages exporting links for posts and setting related data (e.g., properties and mappings 12. "`RelatedIdentifier`"); as well as encourages encoding in metadata 16. "`Rights`", to declare types of access and ownership; and recommends using metadata 17. "`Description`", to deepen the description of attributes of resources.

Research Data Alliance Metadata Schema

Research Data Alliance (RDA) is an initiative promoted by the European Commission and other organizations, with the aim of building a social and technical infrastructure for open exchange and reuse of data.

RDA has an inclusive approach that covers all stages of data lifecycles, engaging producers, users and managers of data, processing, and storage for an open and interoperable exchange of global research data.

RDA Metadata Working Group (RDA, Metadata) develops and works on a core metadata set for research data, which is shown in *Image 7* (RDA. Metadata Element)

RDA Metadata for Research Data		
	RDA Metadata	**Metadata Description**
1.	Unique Identifier	Unique identifier (DOI, URI)
2.	Originator	Creator (organizations / people)
3.	Project	Project (Full and/or abbreviated project name)
4.	Provenance	Provenance (origin, location or source of resource)
5.	Spatial coordinates	Geographic location of resource
6.	Temporal coordinates	Temporal coordinates (data production date)
7.	Related publications	Related publications (IsCitedBy, Cites, IsPartof, HasPart, IsDocumentedBy, IsCompiledBy)
8.	Citations	Citations (how to cite resource)
9.	Descriptions	Description of resource (free text, indexed, etc.) It includes type of description (abstract, table of contents, methodology, etc.)
10.	Schema	Schema (Format)
11.	Medium	Medium
12.	Related software	Related software (related to data and how it is transmitted)
13.	Facility / equipment	Facilities and equipment used to generate data/s
14.	Keywords (terms)	Keywords
15.	Availability (license, persistence)	Availability (license, persistence) (restrictions, conditions)
16.	Location	Location (URL)
17.	Quality	Quality (Metrics)

Image 7 – RDA Metadata for Research Data
Source: RDA Metadata Elements

RDA provides seventeen metadata fields which have an internal syntax (structure) and use of terms that require semantics. This list of metadata is recommended, it allows location of data; supports contextualization (relevance and value) and facilitates action (interoperation). A significant contribution of RDA is that it develops an open and collaborative directory of metadata standards for scientific data. (RDA. Metadata Standards)

Dublin Core Metadata Schema

Dublin Core is developed by the Dublin Core Metadata Initiative (DCMI) to establish a set of basic metadata fields to describe various information resources available on Web.

The Dublin Core metadata schema is widely used (Gómez, N., 2016) due to the level of adaptability of its basic format, as well as the high degree of adaptation and enrichment of its extended format to in describing different information resources. (Greenberg and White, 2012)

The Dublin Core metadata schema is adopted by several international organizations for the organization of information resources, mainly academic publications, and gradually, for the development of research data repositories. The basic Dublin Core metadata schema for research data is shown in *Image 8*.

Basic Dublin Core Metadata for Research Data		
	RDC Metadata	**Metadata Description**
1.	dc.identifier	Resource identifier (DOI)
2.	dc.identifierType	Identifier type (DOI, Handle, URL)
3.	dc.creator	Author
4.	dc.creatorName	Author name
5.	dc.title	Title
6.	dc.Publisher	Publisher
7.	dc.PublisherYear	Year of publication
8.	dc.date	Date
9.	dc.dateType	Type of date (publication-distribution; embargo; public date; embargo end)
10.	dc.ResourceType	Resource type
11.	dc.resourceTypeGeneral	General type of resource
12.	dc.RelatedIdentifier	Related identifier (arxiv, ISBN, ISSN, URL, Handle)
13.	dc.relationType	Relationship type (IsCitedBy, Cites, IsPartOf, HasPart, IsDocumentedBy, IsCompiledBy)
14.	dc.rights	Rights (access; embargo)
15.	dc.rightsURI	URL or URI (of license)
16.	dc.subject	Subject

Image 8 – Basic Dublin Core Metadata for Research Data
Source: Dublin Core. Metadata Element Set

The sixteen Dublin Core basic metadata fields allow elements of a research data set to be recorded and described. This set of metadata is mandatory. Likewise, in a basic metadata table for DC research data, "`dc.identifier`" metadata to use of persistent identifiers stands out; metadata "`dc.identifierType`" to assign the type of identifier; "`dc.dateType`" metadata record allow designate publication-distribution dates; embargo; public availability; and end of embargo; "`dc.resourceType`" to register type of data that is related to "`dc.resourceTypeGeneral`" to specify format of data; "`dc.relationType`" metadata is essential to establish a relationship with some other resource, be it publication and/or data.

Final Considerations

The transition of academic communication, represented by the development of data repositories in a framework of open science, highlights the need to study elements that will make up these communication channels.

Therefore, we can conclude that research data repositories face challenges of establishing relationships with publication repositories to increase reproducibility of research processes and results, either to support publications and/or support the development of research processes.

Likewise, research data that is produced in the context of scientific research is diverse, so it is concluded that it is necessary to be clear about the concepts, types, and formats to make up data repositories.

Also, given that a trend in open science is the dynamic flow for the exchange and opening of research outputs, it is relevant to harmonize FAIR principles with information organization processes so that the data is findable, accessible, interoperable, and reusable

without barriers.

In conclusion, the organization of research data for repositories with collaboration of library science standards will be a useful element that will help create the right conditions for data exchange to become a standard in academic and research contexts.

The analysis of different metadata schemes for research data is a theoretical and methodological approach that allow us to identify basic descriptive elements of research data to relate them to metadata of academic publications to contribute to openness and reproducibility of such resources in repositories.

References

Berlín. (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities / Max Planck Society and the European Cultural Heritage Online.

Callicot, Burton, Scherer, David, Wesolek, Andrew. (2016). "Introduction". En *Making institutional repositories work*. United States of America: West Lafayette, Indiana: Purdue University Press. p. XV.

CONACyT. (s.f) "*Lineamientos específicos para repositorios".* México: Consejo Nacional de Ciencia y Tecnología. [s.a.] Available in: https://www.siicyt.gob.mx/index.php/normatividad/2-conacyt/1-programas-vigentes-normatividad/lineamientos/lineamientos-especificos-para-repositorios/4704-lineamientos-especificos-para-repositorios/file

European Commission. (2017) "H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2, 21 march". p. 4. Available in: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Gómez, Nancy-Diana; Méndez, Eva; Hernández-Pérez, Tony (2016). "Social sciences and humanities research data and metadata: A perspective from thematic data repositories". *El profesional de la información, v. 25, n. 4,* pp. 545-555. Available in: http://dx.doi.org/10.3145/epi.2016.jul.04

GoFAIR. (s.f) "*FAIR Principles"*. Disponible en: https://www.go-fair.org/fair-principles/

Greenberg, Jane (2012). "Bibliotecas digitales para datos de investigación". Máster en Bibliotecas y Servicios de Información Digital, Universidad Carlos III, Madrid, 13 febrero. Available in: http://163.117.69.23/mediasite/Viewer/?peid=1c8617e82d83456c940479af3c3f368b

Guajardo, M. (2020). "Factores determinantes para la implementación del esquema de metadatos para repositorios de datos de investigación de la Política de Ciencia Abierta en México" En: Rodríguez, A. Coord. (2020). *La revolución de los datos bibliográficos, científicos y culturales*. México: UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información. pp. 143-160. Available in: http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/8

Hernández-Pérez, Tony; García-Moreno, María-Antonia (2013). "Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios". *El profesional de la información, 2013, mayo-junio, v. 22, n. 3*, pp. 259-263. Available in: http://dx.doi.org/10.3145/epi.2013.may.10

McNeill, Katherine. (2016). "Repository Options for Research Data". En *Making institutional repositories work*. United States of America: West Lafayette, Indiana: Purdue University Press. p. 25.

OpenAIRE. "Application Profile Overview". En *OpenAIRE Guidelines for Data Archives*. Available in: https://guidelines.openaire.eu/en/latest/data/application_profile.html

RDA. "*Metadata IG*". Available in: https://rd-alliance.org/groups/metadata-ig.html

RDA. "*Metadata Element Set*". Available in: https://www.rd-alliance.org/groups/metadata-ig.html

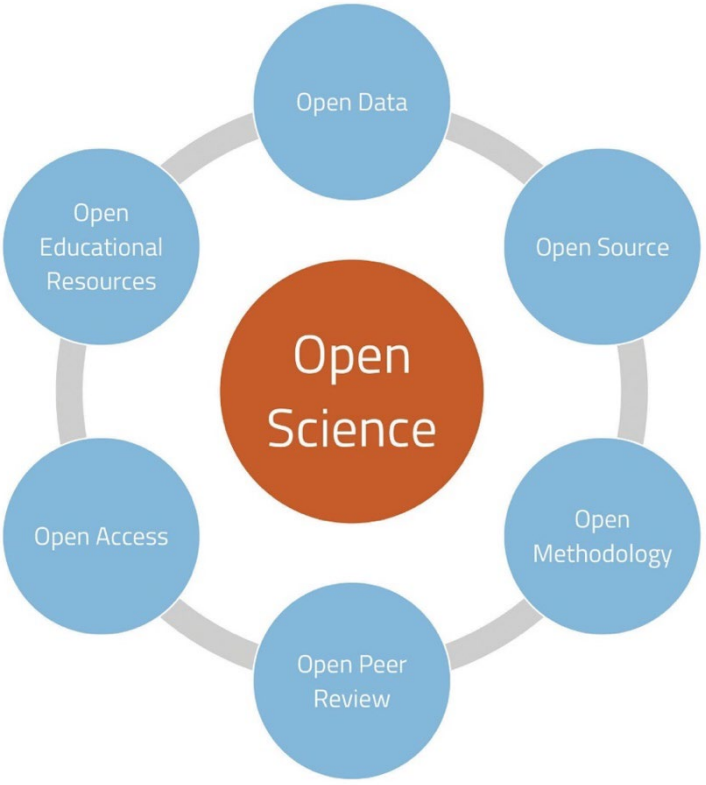
RDA. "*Metadata Standards Directory Working Group"*. Available in: http://rd-alliance.github.io/metadata-directory/

University of Leicester. (2020). "*Research Data Definitions".* Available in: https://www2.le.ac.uk/services/research-data/old-2019-12-11/documents/UoL_ReserchDataDefinitions_20120904.pdf

University of Minnesota, (2020). "*What is data*?". Available in: https://www.lib.umn.edu/datamanagement/whatdata#rep

From Open Science and Datasets to AI and Discovery

Dr. Raymond Uzwyshyn, rju13@msstate.edu
Associate Dean, Research Collections
Mississippi State University Libraries, US

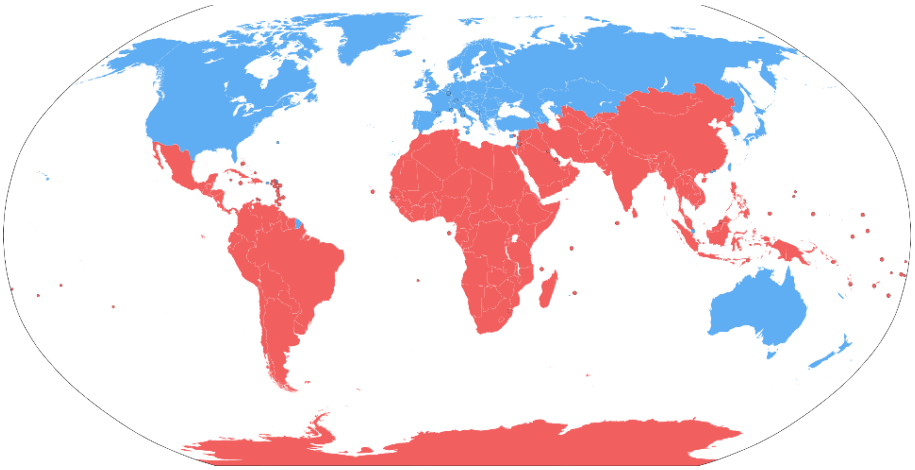


The Many Facets of Open Science Digital Library Ecosystems

Introduction

Exciting new online library infrastructures are now available for open science and experimental research data. These possibilities are being enabled by online data repositories, research ecosystems and artificial intelligence discovery. This research overviews and outlines these recommended new infrastructures focusing upon online research data repositories and open science scholarly ecosystems. Together, these better orient scientific research towards discovery and new possibilities.

This paper utilizes a university library research data ecosystem to overview the utility of placing an online research data repository within an open science research ecosystem framework. Specific examples of online data and open discovery within Artificial Intelligence from Deep Learning are provided. Two online medical related image dataset library examples foreground the importance of online data repository ecosystems towards open science within digital library ecosystems, and particularly, Artificial Intelligence and new discovery. The first example is from a recent discovery (2017) from scientific deep learning neural net towards cancer detection. The example utilizes object recognition and big data for machine learning and neural net training from a US university (Stanford). The second example (2022) builds on the first model's methodology utilizing an undergraduate student thesis from BRAC University from Dhaka, Bangladesh. Stanford's earlier AI neural net model and enabling affinities with geographically dispersed ecosystem methodologies open AI neural net research with online available datasets. Both examples give compelling evidence illustrating the value of open science. Online open data research repositories within data-centered scholarly ecosystems enable the future progress of science and discovery in our new millennia. They bridge collaboration possibilities among traditional global north and south divides.

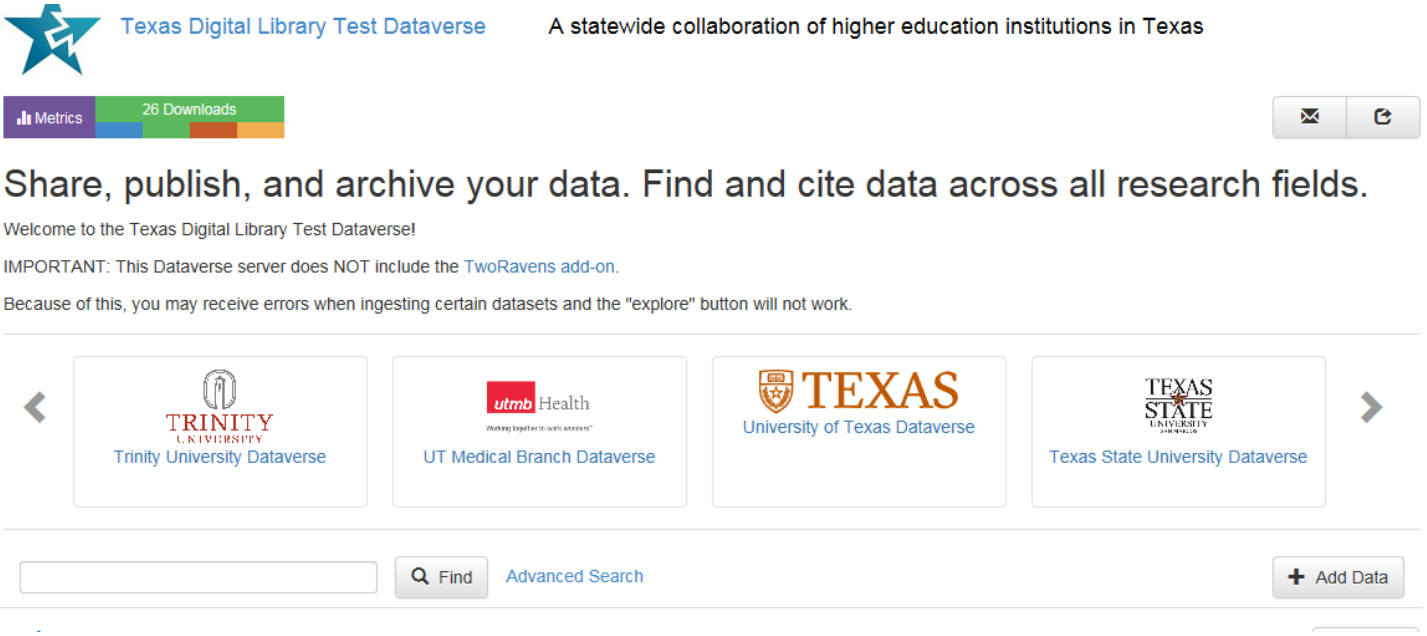


World Map Showing Traditional North-South Global Divide

These new potentials for open science are enabled by: recent constellations of global networks, powerful new algorithmic AI Neural Network Deep Learning models and online storage and retrieval capacity of data research repositories. This increasing computing processing power pragmatically enables these paths. Rudiments of an online data research repository and open science research ecosystem are outlined. Easily implementable digital library examples are utilized. These examples show how these new infrastructures may be used to enable future AI methodologies for scientific discovery in the 21st century. While this is not the only utility of these ecosystems and repositories, it is an important pathway which shows large promise.

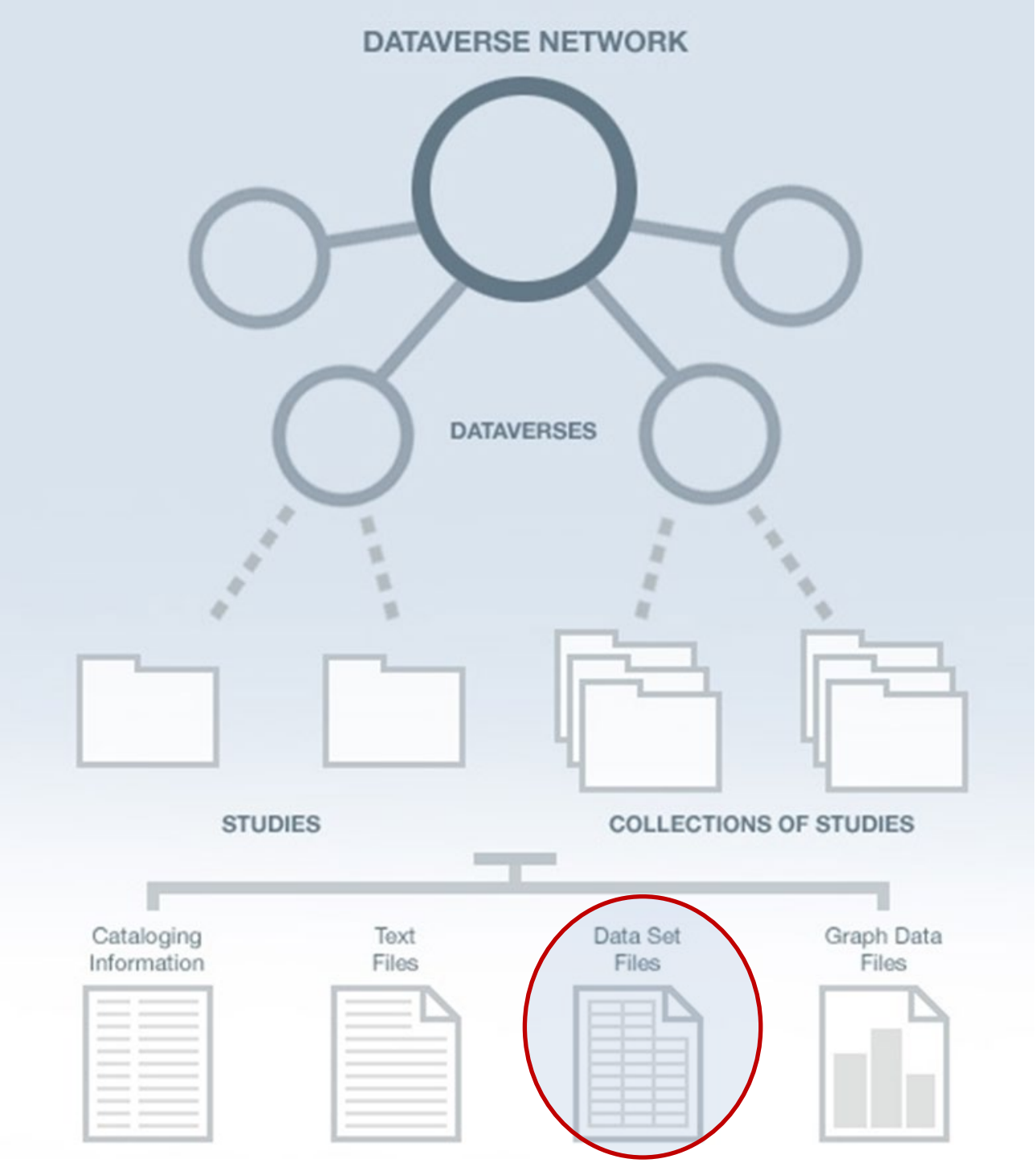
What is an Online Data Research Repository?

An online data research repository allows one to share, publish and archive a researcher's data. It is at once a platform to manage a researcher's and institution's data and metadata. It is also a perma-linking strategy for Data Citation, a way to manage mandated large grant compliance, and an efficacious global data archiving and sharing strategy.



Texas Data Research Repository: https://dataverse.tdl.org

The Texas Data Repository is a good example of an online data repository. It utilizes Harvard's open source Dataverse software, customized towards a multi-university strategy.[1] This Data Repository aggregates various individual university's data for search and retrieval. It can be configured as a single instance for searching or to search across an entire group of institutions. The repository can also easily be configured on consortial, state, or international levels.



Texas Data Repository Consortial Architecture

[1] See Uzwyshyn, Online Data Repositories (2016). https://www.researchgate.net/publication/304780954_Online_Research_Data_Repositories_the_What_When_Why_and_How

Digital Scholarship Ecosystems

A digital repository may also be placed within a larger digital scholarship ecosystem. This enables a horizon of content and further global network communication. The prototypical digital scholarship ecosystem utilizes well-known open-source digital repository software (i.e., Dspace, etc.), for the university's digital collections repository. Four other tertiary components are then utilized by researchers to better enable online global communication and network possibilities. These four applications are an online electronic theses and dissertation management system, ETD System (VIREO), identity management system (ORCID), open academic journal system software (OJS3) and user interface content management software (OMEKA). Together, these function as a unified digital scholarship ecosystem comprising larger thematic elements. Synergistically, these enable technologies of content and communication. [2]

This ecosystem allows for great facility in enabling data-centered methodologies. It continues to build on strong foundations. It provides foundational training data for later AI pathways that may be needed. The general common characteristics for such a digital system are open-source software, active developer communities, communication, and content repository components. The open-source software allows customizability and connection between components. Active developer communities for the software enable an exchange of new possibilities with regards to continuing innovation. The open-source code allows bridges among systems. The sum of the system's capabilities exceeds the separate parts. Collocating open-source digital components in a networked research ecosystem enables large connections, network effects and untapped possibility.

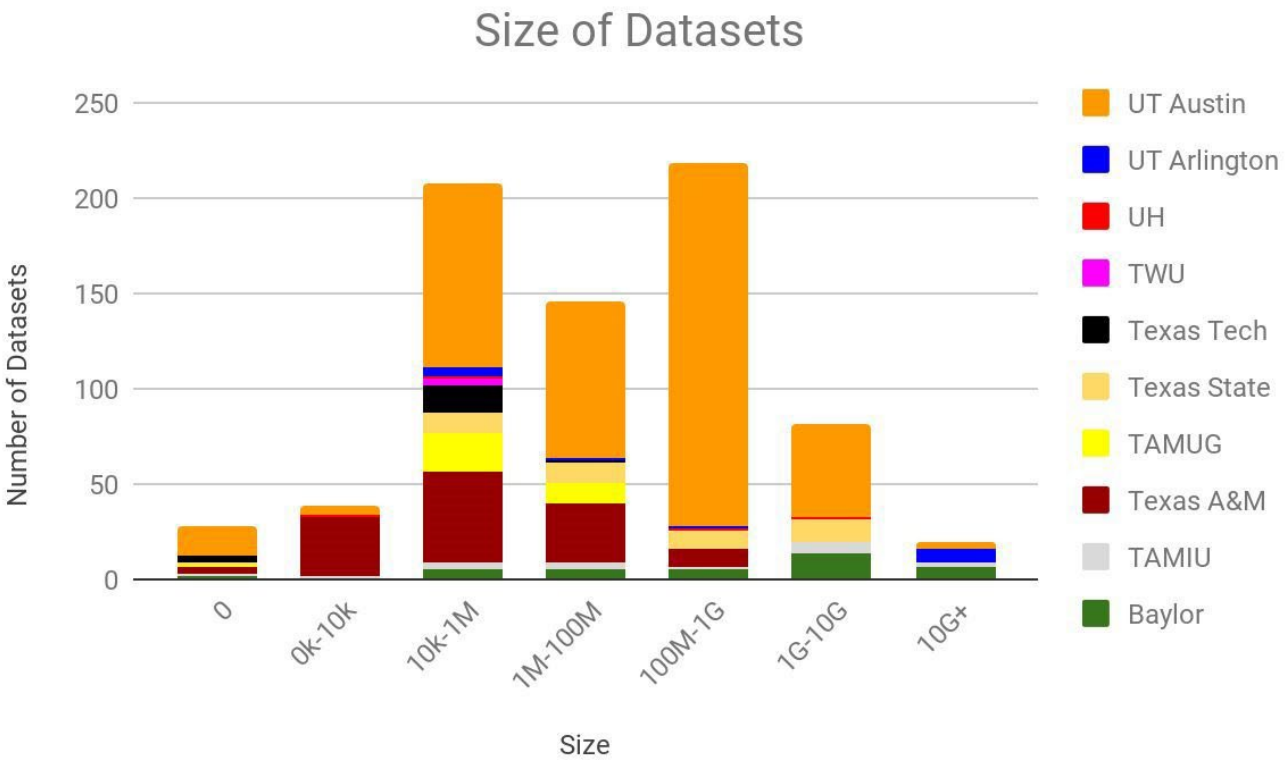
Cooperatively, these digital ecosystem components enable a larger open science research cycle. This cycle moves from original search and retrieval of data and content to gathering and analysis of data, to later writing, publishing, sharing and further discovery.

Open Science Research Cycle.

Data, Datasets, Big Data

Data comes in a variety of file types, formats, media, and sizes. For AI and, particularly recent Deep Learning, labelled and unlabeled datasets become important for machine learning and AI model training. Within open science frameworks, metadata (labelling) is key. One size also does not fit all for various open science data research repository project needs. There are many types of sizes for data projects and repositories. Repositories utilizing Dataverse can typically upload currently up to 4GB data for individual files and 10GB Datasets. This may not seem particularly large currently, considering recent examples. There are now mammoth level natural language processing datasets, or image/video modelling datasets, used for training Google's DeepMind or Microsoft's Open AI (see Mitchell, 2022). These models utilize Terabytes and Exabytes of data. Smaller datasets though, serve the needs of still many academic researchers and have served researchers well for the last six years (2017-2023).

[2] See Uzwyshyn, 2020. Available at: https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_Scholarship_Ecosystem

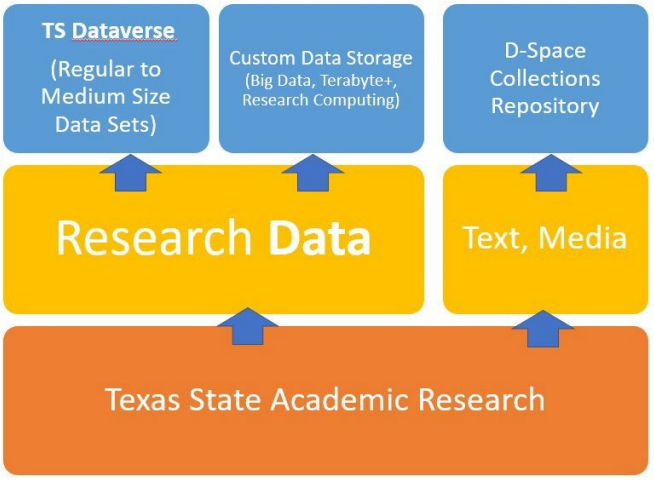


General Size Preferences for Data Research Repository Datasets Example (See Waugh, 2020)

Most experimental researchers' open science datasets for upload have been between the 1< <1000 MB range. Currently, there is the growing recognition by researchers that 'bigger' data repositories are needed. These begin in the GB and TB ranges, though preparation must be taken now for the next phase. Many researchers are also working extensively with specialized media or GIS datasets. In these cases, for larger and custom data storage, it is still not yet feasible to place these huge datasets online, especially those in the Exabyte range. These are typically placed with university research computing data centers, or the local area supercomputing center for custom data storage, should these needs arise. This type of storage is usually worked out by researchers in preliminary grant applications expecting this level of data storage needed for research work.

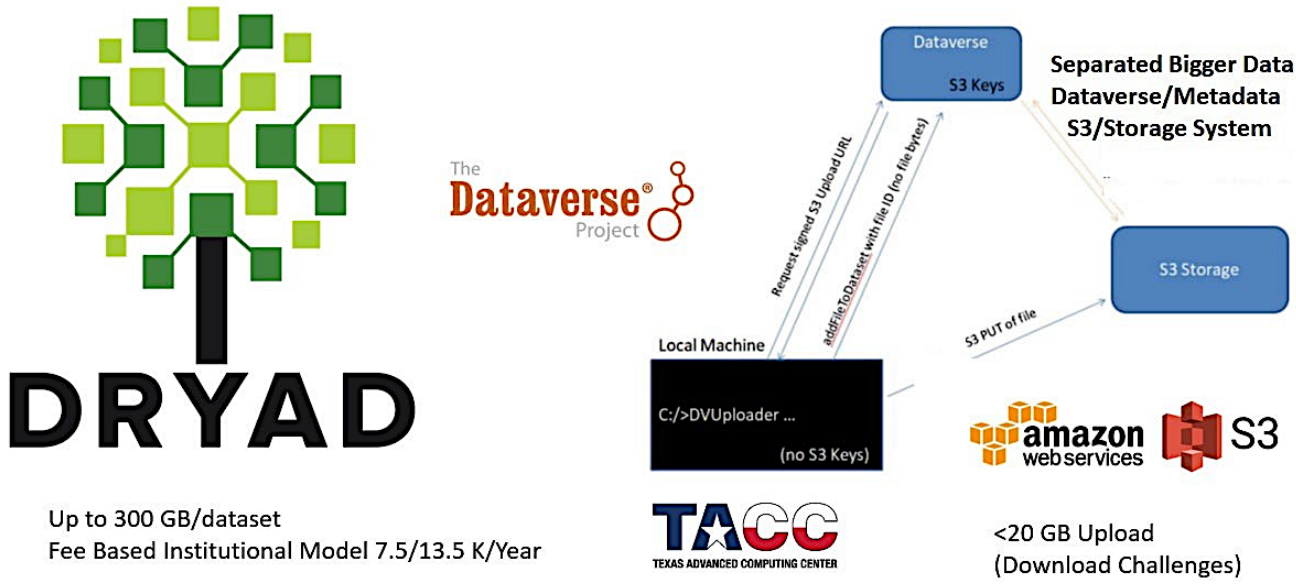
Beyond custom big data storage needs, the requests for very 'big data' (Terabytes, Exabyte storage) are still few, but these requests are increasing.

In this regard, libraries have been exploring various 'bigger data options and beta prototypes (2020-2022) with partnerships.



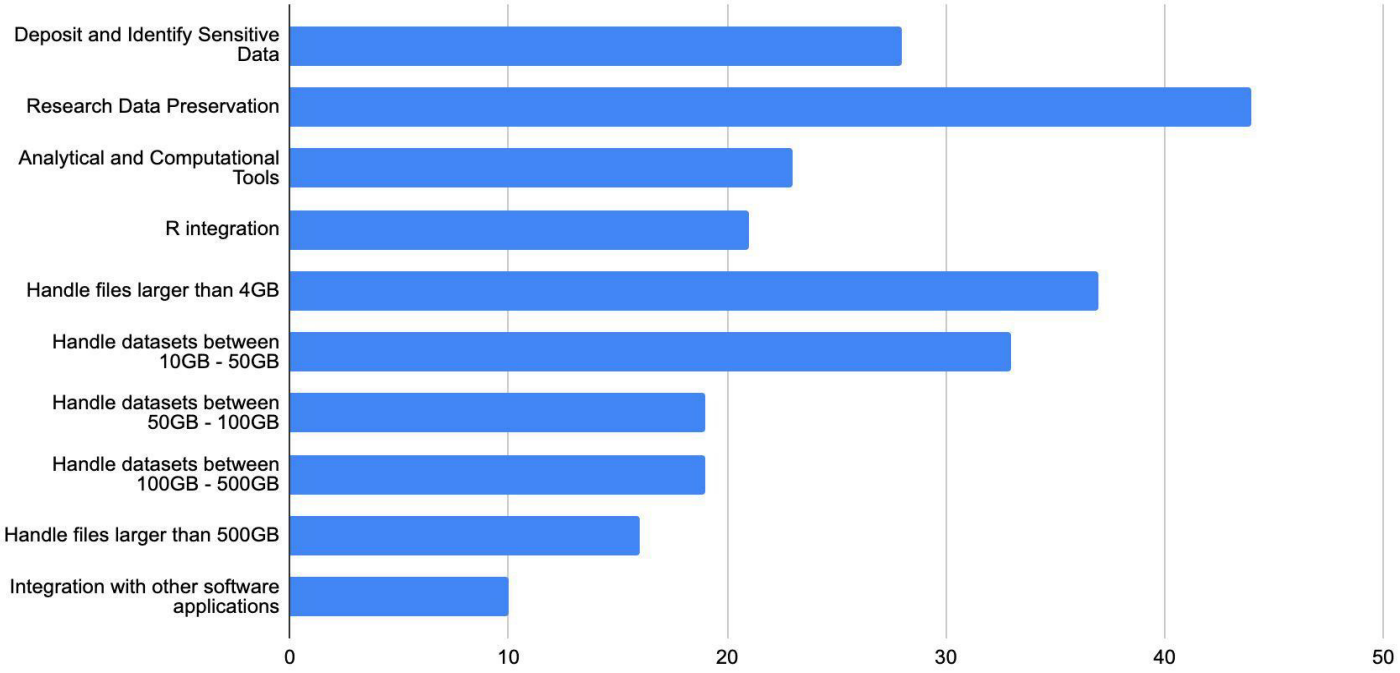
General University Big Data Storage Model

This ranges from commercial partnerships (Amazon Web Services S3 storage) to state efforts (i.e., Texas Advanced Computing Center, TACC) to hybrid metadata/storage pointer systems and more fee-based institutional models filling middle ground space needs (i.e. 300GB/dataset, Data Dryad).



Beta Prototyping Bigger Data Online Texas Data Repository Architectures. 2020-2022, TACC: https://www.tacc.utexas.edu/, Data Dryad https://datadryad.org/stash

Currently, 'Big Data' (Exabyte, Terabyte) is among, but not at the top of, the list of new data research repositories feature-set requests that most researchers would like to see. Higher on this list of new features is long-term research digital data preservation[3]. Also ranking high, is managing middle ground data files (4-10 GB range) and datasets in the 10-50 GB range as well as being able to safely deposit and clean sensitive data (i.e., medical related, etc., see data survey below). Support for analytical and computational tools also comes high on the list. Ranging from data analytics and visualization, these tool and data literacy requests help to enable researchers from non-computer science disciplines towards new AI methodologies such as those being forwarded through neural net and deep learning. [4]

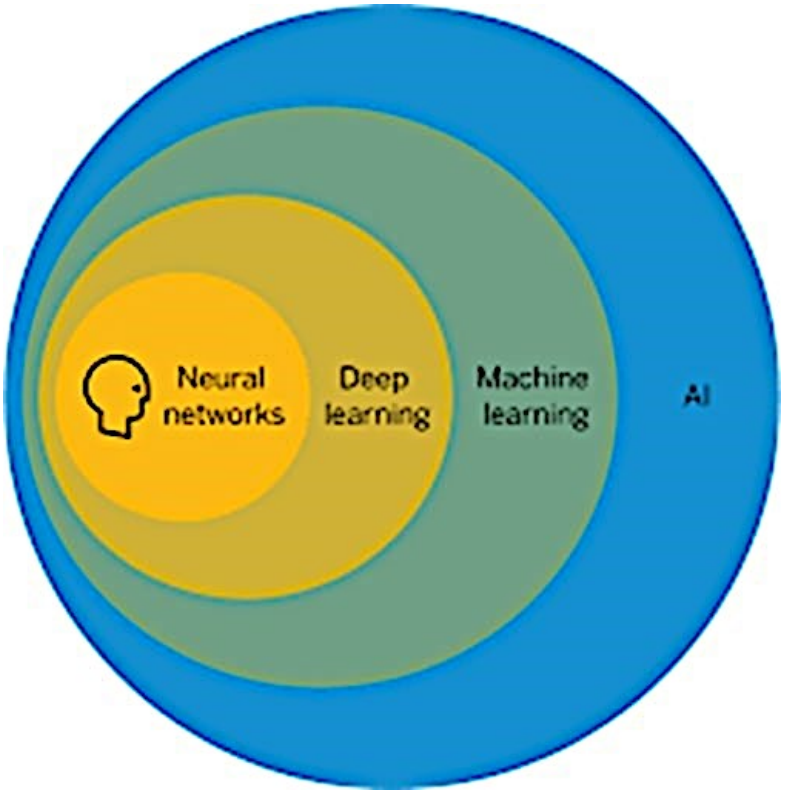


What New Data Research Repository Features Would Users Like to See? (Chan-Park, Sare and Waugh, 2022)

[3] See Uzwyshyn, 2021. Frameworks for Long Term Digital Preservation. Computers in Libraries.

[4] An important learning infrastructure for these new library/university researcher algorithmic literacy needs is filled by international communities arising such as The Carpentries https://carpentries.org/

Data Research Repositories, Digital Ecosystems and AI



Venn Diagram of relationships among AI and its subdomains of Machine Learning, Deep Learning and Neural Nets

The last five years (2017-2022) have shown incredible progress and gains in analytic computational tools and discovery. This is particularly true with methodologies associated with new domains of Artificial Intelligence. Machine learning, deep learning and neural net research has shown incredible potential for open science paradigm breakthroughs (Mitchell, 2022). These breakthroughs range from Computer Vision (Facial/Object Recognition) to Natural Language Processing (speech to text recognition and translation), to Cybersecurity (Fraud Detection). These advances also include Conversational Chatbots, Robotic Agents and Strategic Reasoning (AlphaGo, Game Theory).

Breakthroughs have been enabled through a fortuitous combination of better algorithms, greater computing processing power (Compute), more precise metadata schemas, online datasets and, increasingly, open science research data repositories and ecosystems.

The following section utilizes recent discoveries from Neural Net object identification to illustrate how online data research repositories and online data research ecosystems are facilitating the next generation of global collaboration with networked ecosystems research, discovery, and open science.

Cancer Detection, Library Image Data Repositories, AI

In 2017, an innovative new cancer detection methodology was published in Nature by a Stanford University group. They proposed the use of Neural Nets to train an AI neural network (Esteva, Nature, 2017). This training utilized big data and a dataset of 129,460 images of 2,032 diseases. The dataset of images (1.41 million) classified skin cancer lesions utilizing deep neural networks. After comparison, the neural net machine learning AI model did equal to or better than thirty board certified dermatologists with decades of experience.

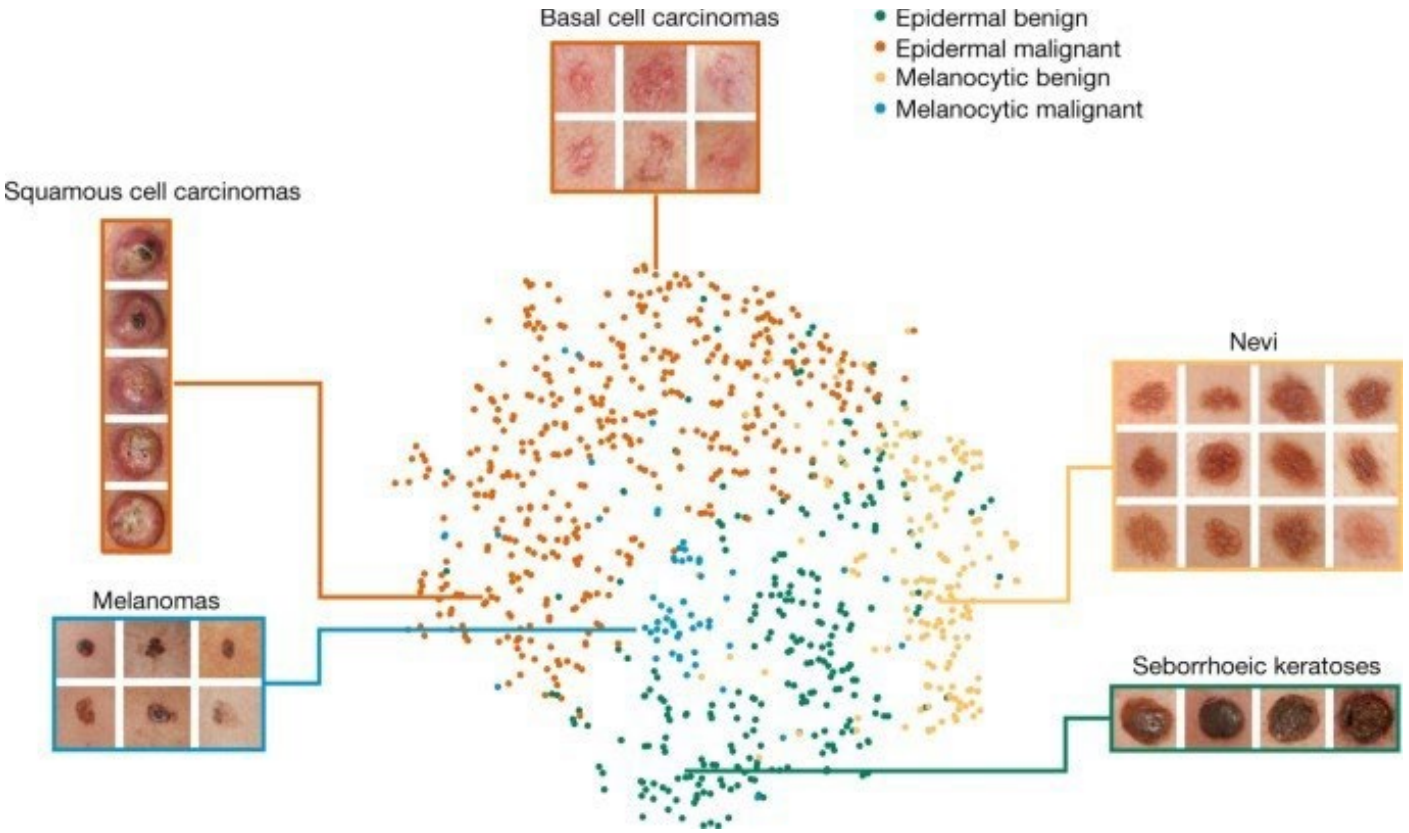


Image from Dermatologist Level Classification of Skin Cancer with Deep Neural Nets (Esteva et al, 2017) [5]

The neural net model was able to successfully classify epidermal lesions on mobile phones for early cancer detection into benign and cancerous (malignant) lesions better than credentialled experts. This method involved pixel-level differentiation and training through a multilevel neural net AI model.

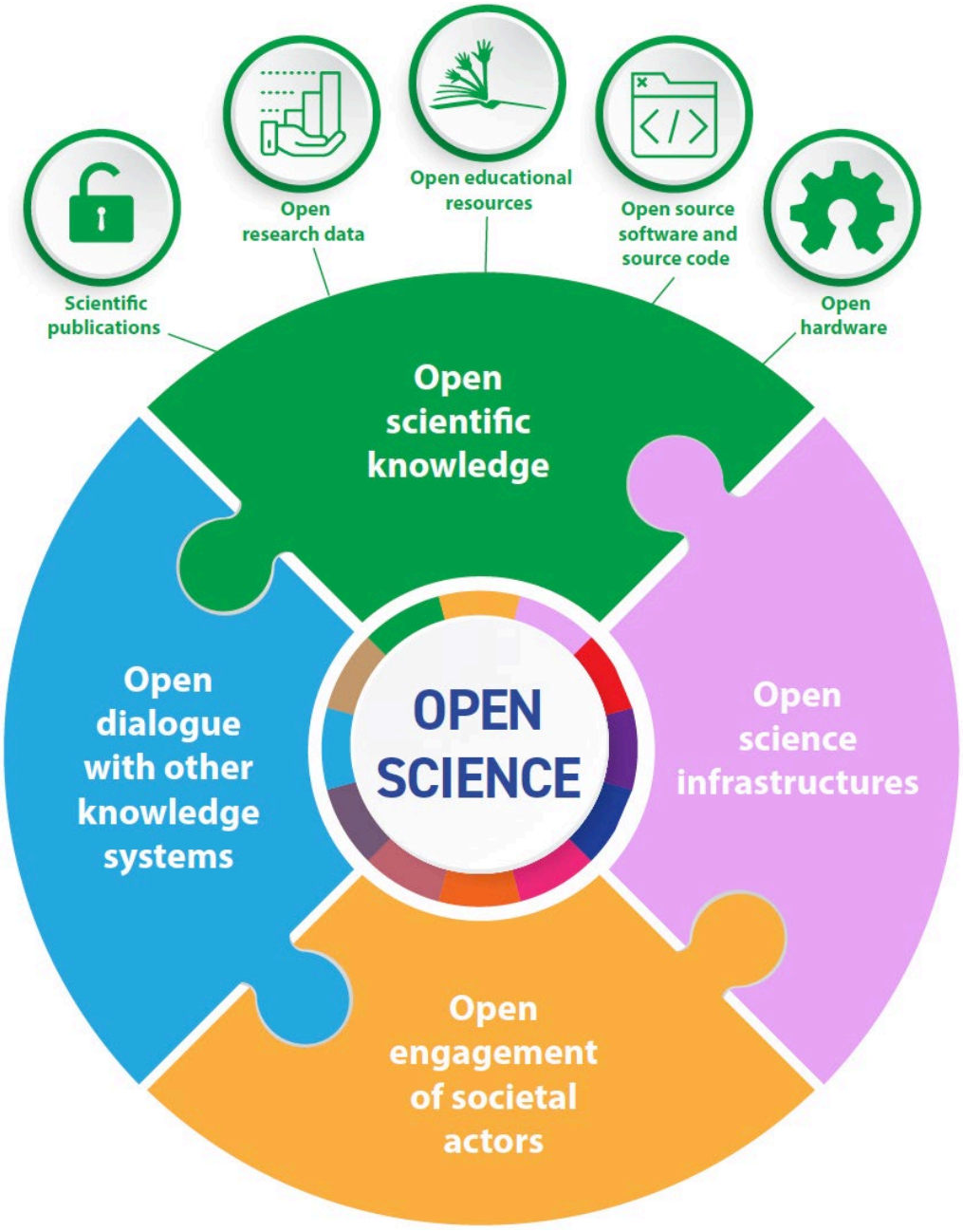
The large relevance of the digital image data repositories for initial training and metadata labelling should not be underestimated for researchers. In a recent article on *Deep Learning in Cancer Pathology Surrounding a New Generation of Clinical Biomarker* (Echle, 2020), the authors emphasize the need for: organized digital data repositories, metadata preprocessing for later accuracy in training and external validation.

Open Science, AI, and Data-Centered Ecosystems

Huge data sets like the Stanford example are not the only and most recent of those able to be utilized through AI and neural net methodologies. Innovative global open science infrastructures are being assertively forwarded. (see UNESCO Figure below). AI machine learning possibilities are also being leveraged efficiently through previous algorithmic training and the application of new regular sized datasets. New affordances are enabled through a confluence of data research repositories and researchers' willingness to share their research and data sets through open science.

Research data libraries open search and retrieval. These allow other researchers globally to apply algorithmic machine learning and building on previous models to available new online research data.

[5] See also, the original article from Nature. Esteva, A, Thrun, S. et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature, Volume 542 (February 2, 2017). pp. 115-119. doi:10.1038/nature21056 and Eschle, 2020. Summary Video: https://youtu.be/IvmLEq9piJ4



UNESCO FACETS for Open Science, See: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

If a university or research institution does not possess a Data Repository and the researcher is conducting valid academic research, they can utilize the Harvard repository. Harvard's open source Dataverse software may be utilized for the uploading of datasets from other universities globally. Appropriate research datasets may be uploaded for sharing later or use by researchers anywhere. Dataverse is also open-source software. This means any research level library, institution and university should can set up their own instances of data repository and digital ecosystems.

To further trace these innovative discovery example pathways, the HAM10000 image dataset is a diverse collection of multi-source dermatoscopic images of cancerous skin lesions. This dataset was uploaded to Dataverse by Viennese Dermatologist, Dr. Philip Tschandl, in 2018, a year after the Stanford Nature Neural Net algorithmic methodology article appeared.

As can be seen, the images, data and metadata can be easily downloaded, unzipped, and used by researchers for neural net training purposes.

The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions

Version 3.0

Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", https://doi.org/10.7910/DVN/DBW86T, Harvard Dataverse, V3, UNF:6:/APKSsDGVDhwPBWzsStU5A== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Access D

Contact Owner

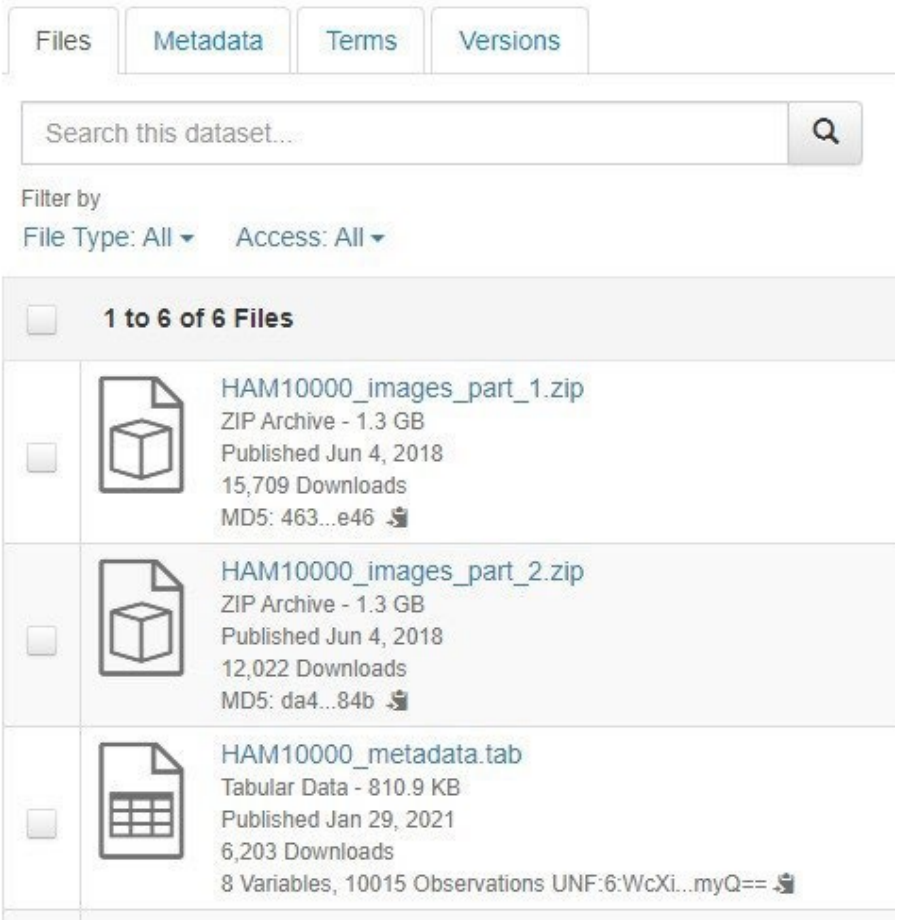
Dataset Metrics

58,334 Downloads

Description

Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (`akiec`), basal cell carcinoma (`bcc`), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, `bkl`), dermatofibroma (`df`), melanoma (`mel`), melanocytic nevi (`nv`) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, `vasc`).

HAM10000 Dataset in Dataverse Data Research Repository,
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

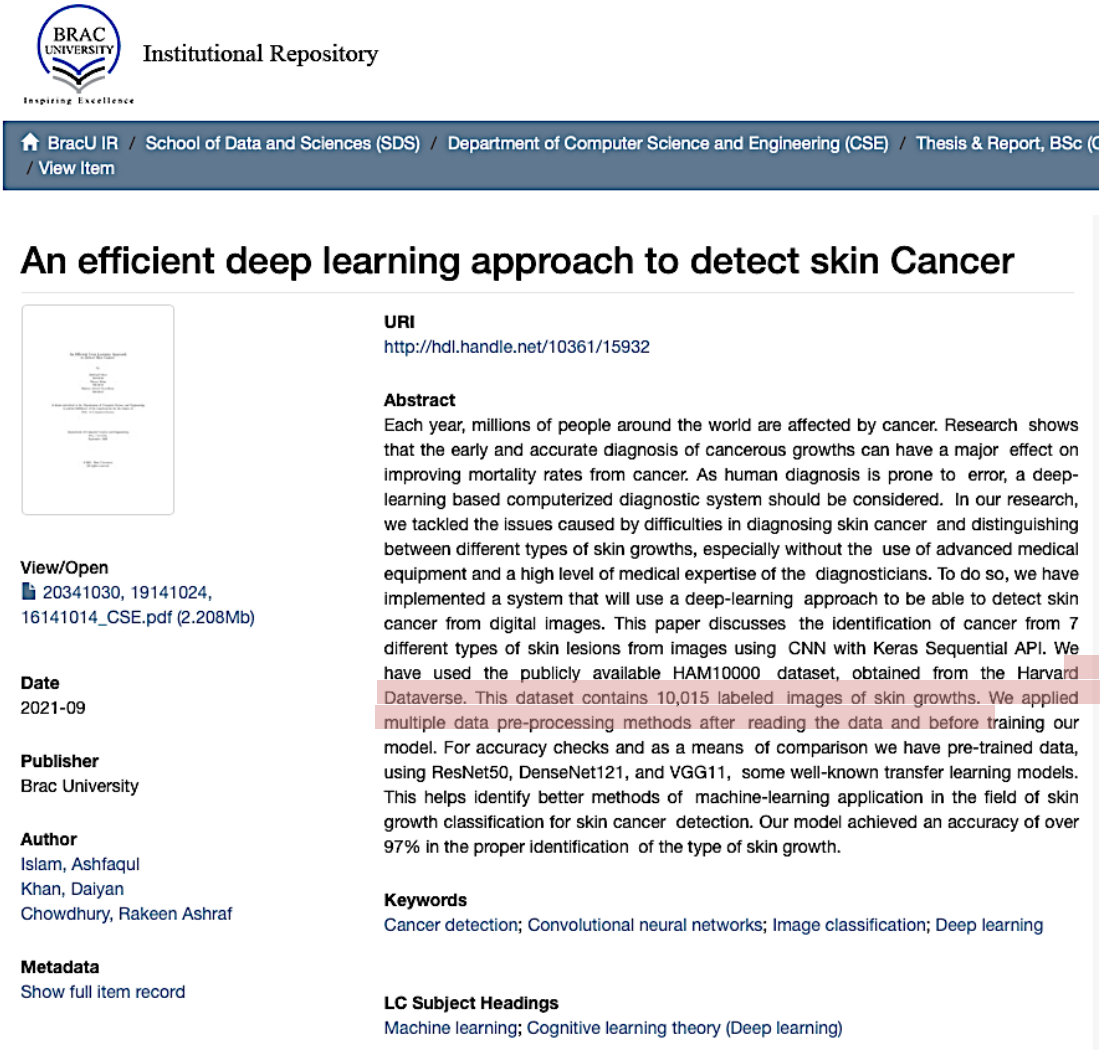


HAM10000 Dermascopic Cancer Images, Harvard Dataverse Repository,
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

One may use data research repositories to facilitate open science around the globe through online dataset reuse. This occurs through researchers in other areas of the globe training and further developing previous deep learning and neural net models. This includes customization of variables, weights and new gathered datasets. This also becomes remarkably interesting with regards to possibilities for open science, globally dispersed academic researchers, and new possibilities for discovery and innovation.

Below is a cover page from BRAC University from Dhaka Bangladesh. They are using DSpace, a well known open source institutional repository software to house theses and dissertations from the School of Data and Sciences, Dept. of Computer Science and Engineering. Here, the computer science and engineering students had earlier downloaded Dr. Tschandl's dermatological cancer training images, metadata, and datasets. They utilized the labelled image data as training material to train a deep learning neural net algorithm with further parameterization to recognize cancer growths with efficiency and accuracy to build on previous results. The example becomes interesting for global possibilities of telemedicine, mobile possibilities, data and global populations which may not have as quick access to trained specialists as those in the West.

This is also a particularly good example of open science and AI possibilities operating on global levels through the enabling power of digital scholarship ecosystems and data repositories' aggregation abilities. Content and specialized image data sets with specialized labelled metadata can be aggregated online that would otherwise be unavailable. This data can now be easily brought together utilized, reviewed, and improved upon with new machine learning algorithmic techniques. An example of new research, and an exceptionally good thesis has been produced below by undergraduates from the global south. Globally dispersed content and data, from three different continents, has been aggregated to advance the pursuit of knowledge and science. There is a speed, dispersion and utility here that would be unimaginable in previous centuries.



BRAC UNIVERSITY Institutional Repository

Inspiring Excellence

BracU IR / School of Data and Sciences (SDS) / Department of Computer Science and Engineering (CSE) / Thesis & Report, BSc (C / View Item

An efficient deep learning approach to detect skin Cancer

URI
http://hdl.handle.net/10361/15932

Abstract
Each year, millions of people around the world are affected by cancer. Research shows that the early and accurate diagnosis of cancerous growths can have a major effect on improving mortality rates from cancer. As human diagnosis is prone to error, a deep-learning based computerized diagnostic system should be considered. In our research, we tackled the issues caused by difficulties in diagnosing skin cancer and distinguishing between different types of skin growths, especially without the use of advanced medical equipment and a high level of medical expertise of the diagnosticians. To do so, we have implemented a system that will use a deep-learning approach to be able to detect skin cancer from digital images. This paper discusses the identification of cancer from 7 different types of skin lesions from images using CNN with Keras Sequential API. We have used the publicly available HAM10000 dataset, obtained from the Harvard Dataverse. This dataset contains 10,015 labeled images of skin growths. We applied multiple data pre-processing methods after reading the data and before training our model. For accuracy checks and as a means of comparison we have pre-trained data, using ResNet50, DenseNet121, and VGG11, some well-known transfer learning models. This helps identify better methods of machine-learning application in the field of skin growth classification for skin cancer detection. Our model achieved an accuracy of over 97% in the proper identification of the type of skin growth.

Keywords
Cancer detection; Convolutional neural networks; Image classification; Deep learning

LC Subject Headings
Machine learning; Cognitive learning theory (Deep learning)

View/Open
20341030, 19141024, 16141014_CSE.pdf (2.208Mb)

Date
2021-09

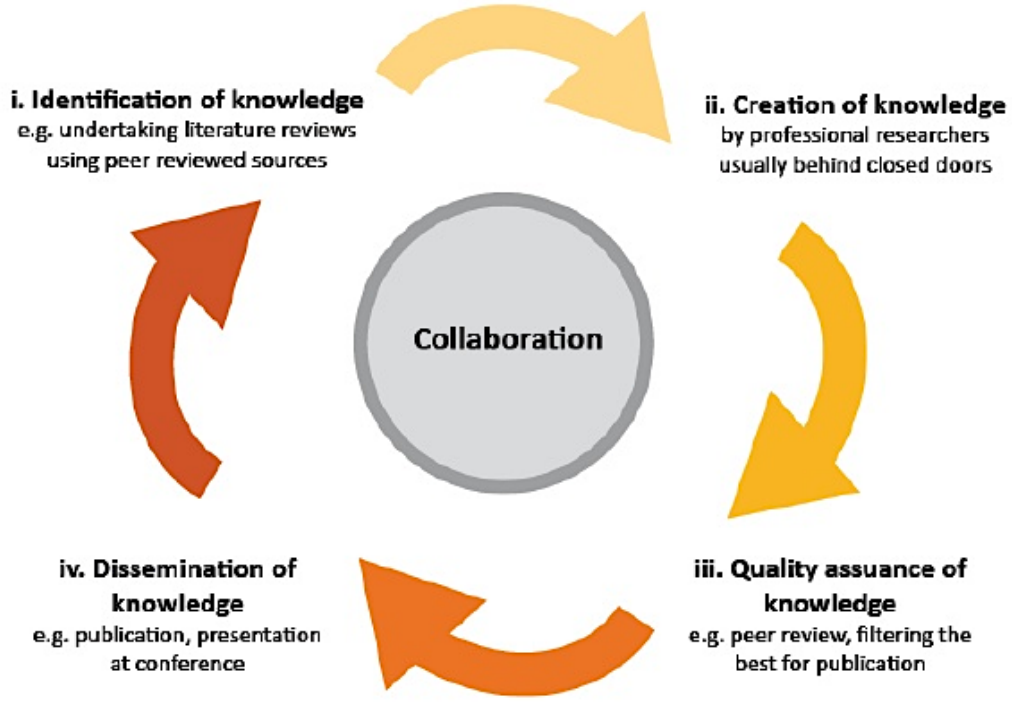
Publisher
Brac University

Author
Islam, Ashfaqul
Khan, Daiyan
Chowdhury, Rakeen Ashraf

Metadata
Show full item record

BRAC University Dspace Repository 2021 Deep Learning/AI Thesis
http://dspace.bracu.ac.bd/xmlui/handle/10361/15932

Conclusion – AI, Data and the Open Science Research Cycle



The Academic Research Cycle, Cann, Dimitrious and Hooley, 2011.

New data repository and digital scholarly ecosystem possibilities are enabling the academic research cycle, progress of knowledge and discovery in our new millennia in amazing ways. Open science possibilities empower a new global networked generation towards incredible new science and knowledge discovery. This is directly through the enabling power of digital libraries, data research repositories and open science scholarly ecosystems.

The creation of data and knowledge usually occurs hidden away in research labs, file cabinets and computer hard drives. Data sharing has now been enabled through possibilities of networked communication and content technologies. This sharing by researchers on a global stage allows transparency towards the quality assurance of knowledge ranging from online peer review to availability of data and research for citation, discovery, download and pragmatic use.

Paired with other ecosystem possibilities such as open online academic research journals, theses, and dissertation (VIREO) and online identity management systems (ORCID), and new multimedia user interface possibilities, these tools facilitate global collaboration and intrinsically human creative activities of discovery, invention, innovation, and progress from previous generations of researchers and scholars.

References

Artificial Intelligence. Machine Learning. Neural Networks. Future Technology. Bloomberg Businessweek Canada. 2022. https://www.youtube.com/watch?v=ypVHymY715M

Cann, A., Dimitriou, K. Hooley, T. 2011. social media: A Guide for Researchers. Research Information Network. University of Derby, UK.

Chan-Park, C. and Sare, L. Waugh, S. 2022. *Results of the Texas Data Repository User Survey*, 2022. Texas Conference on Digital Libraries Presentation.

The Carpentries. Data Science Skills for Libraries and Researchers. 2022. https://carpentries.org/

ColdFusion (2018). *Why Deep Learning Now?* (Documentary Overview). https://www.youtube.com/watch?v=b3IyDNB_cIl

Echle et al. Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. *British Journal of Cancer*. November 2020. https://www.nature.com/articles/s41416-02001122-x

Esteva, A, Thrun, S. et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, Volume 542 (February 2, 2017). pp. 115-119. doi:10.1038/nature21056

Fridman, Lev. *MIT Deep Learning and Artificial Intelligence Lectures*. https://deeplearning.mit.edu/ 2022.

Islam, A., Khan, D. and Chowdhury, R. 2021. *An Efficient Deep Learning Approach to Detect Skin Cancer Undergraduate Thesis*. BRAC University DSpace Institutional Repository, 2021. Available: http://dspace.bracu.ac.bd/xmlui/handle/10361/15932

Kleinveldt, Lynn. Smarter high education learning environments through AI: What this means for academic libraries. *Trends and Issues in Library Technology: Special Issue on AI*: June 2022. pp. 12-15. https://repository.ifla.org/handle/123456789/1940

Marcum C and Donohue, R. *New Guidance to Ensure Federally Funded Research Data Equitably Benefits America*. Office of Science and Technology Policy: Washington DC, The White House. 2022. https://www.whitehouse.gov/ostp/news-updates/2022/05/26/new-guidance-to-ensure-federally-funded-research-data-equitably-benefits-all-of-america/

Mitchell, Tom. 2022 *Where on Earth is AI Headed?* Carnegie Mellon. https://www.youtube.com/watch?v=ij9vqTb8Rjc

NASA. Open Science Overview. 2022. https://science.nasa.gov/open-science-overview

Peters, T. and Waugh, L. Larger Data Storage Report: Research Data Management Initiatives and Planning, January 2022. Texas State University Libraries (Unpublished White Paper).

Texas Data Repository 2022. https://dataverse.tdl.org/

Tschandl, Phillip et al. Human-computer Collaboration for Skin Cancer Recognition. Nature Medicine, 22 June 2020, 1229-1234. See: https://www.nature.com/articles/s41591020-0942-0.

UNESCO. *Recommendations on Open Science*. 2021. Paris: General Conference. 41/C. Resolution 24. https://www.unesco.org/en/natural-sciences/open-science Final Report: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Uzwyshyn, R. 2022. Steps Towards Building Library AI Infrastructures: Research Data Repositories, Scholarly Research Ecosystems and AI Scaffolding. *New Horizons in Artificial Intelligence in Libraries* (IFLA Satellite Conference), National University of Ireland, Galway, IR.

Uzwyshyn, R. 2021. Frameworks for Long Term Digital Preservation Infrastructures. *Computers in Libraries*. September 2021. pp.4-8.

Uzwyshyn, R. 2020. *Developing an Open-Source Digital Scholarship Ecosystem*. ICEIT2020. St. Anne's College Oxford, United Kingdom. February 2020. Available at: https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_ Scholarship_Ecosystem.

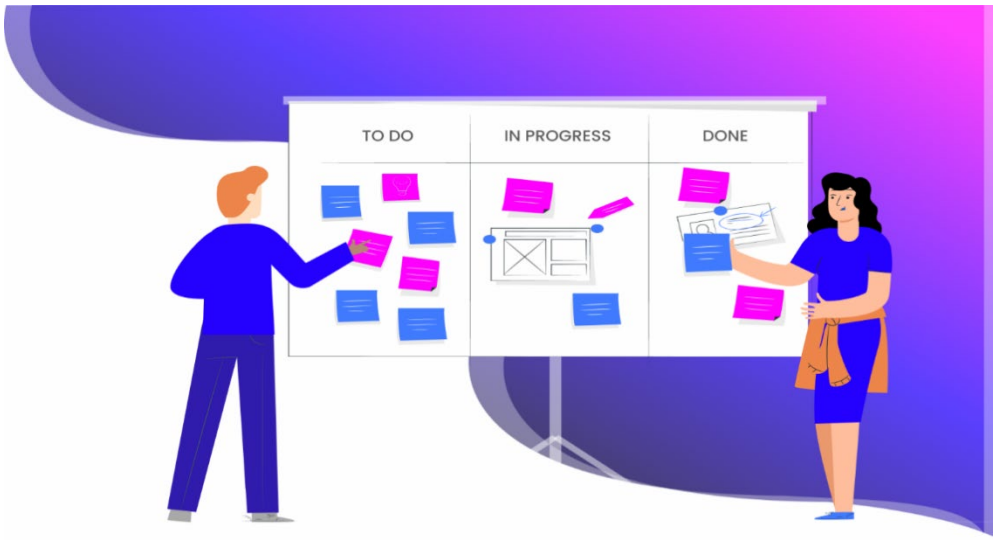
- - -. *Open Digital Research Ecosystems: How to Build Them and Why*. Computers in Libraries, (40) 8. November 2020. https://www.researchgate.net/publication/345956074_Online_Digital_Research_Ecosystems_How_to_Build_Them_and_Why

---. Online Research Data Repositories: The What, When Why and How. Computers in Libraries. 36:3, April 2016. pp. 18-21. http://rayuzwyshyn.net/TXU2016/OnlineDataResearchRepositoriesUzwyshyn.pdf

Waugh, L. *Texas State University Annual Usage Report 2020.* TXST Dataverse Repository. Texas Conference on Digital Libraries Presentation. Texas State University.

Applying Agile Principles for ICT Operations Management in an Academic Library Setting

Wouter Klapwijk, wklap@sun.ac.za
Director, Library Information Technology Services
Stellenbosch University, South Africa



Kanban Board: A Unique Agile Approach to Software Project Management

Introduction

Software project management deals with the peculiar problems associated with creating software while incorporating key elements of generic project management. It therefore addresses the important consideration that how systems are implemented is as vital a consideration as what a system does. In response to this, the publication of the Agile Manifesto in 2001 (Beck et al., Manifesto for Agile Software Development) marks the birth of agile as a methodology to improve the way past systems were implemented using heavyweight implementation methods. Since then, many agile frameworks have emerged such as Scrum, Kanban, Lean and Extreme Programming (XP), some which existed well before the writing of the Manifesto and became more popular thereafter.

Agile Principles

Agile is a group of methodologies that demonstrate a commitment to tight feedback cycles and continuous improvement. From more practical perspectives in the work environment of the ITS team of Stellenbosch University Library and Information Service, agile means more the assignment of greater autonomy and enablement to members of the team and a lesser focus on the detailed planning and specification of a product or service's criteria. Openness, trust, and autonomy are emerging as the cultural currency for institutions who want to attract the best information workers and get the most out of them. Agile is a cultural value. Teams should be empowered to work how they best see fit. This work culture has lately been advanced by the COVID-19 pandemic necessitating information workers to work in physically separate, but organizationally coherent fashions. In the case of hybrid library environments, Agile has also become one of the solutions to be more responsive to the demands placed on Information and Communication Technology (ICT) teams to deliver on digital library services.

Agile Manifesto

We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

Individuals and interactions	over processes and tools
Working software	over comprehensive documentation
Customer collaboration	over contract negotiation
Responding to change	over following a plan

While there is value in the items on the right, **we value the items on the left more.**

The 12 Principles of Agile

1. Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.
2. Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
3. Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
4. Business people and developers must work together daily throughout the project.
5. Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
6. The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
7. Working software is the primary measure of progress.
8. Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. Simplicity – the art of maximizing the amount of work not done – is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly.

Advancing the principles of Agile

Agile Alliance

Learn more at AgileAlliance.org

THE MANIFESTO AUTHORS

Kent Beck	Alistair Cockburn	Robert C. Martin	James Grenning	Ron Jeffries	Ken Schwaber
Mike Beedle	Ward Cunningham	Steve Mellor	Jim Highsmith	Jon Kern	Jeff Sutherland
Arie van Bennekum	Martin Fowler	Dave Thomas	Andrew Hunt	Brian Marick	

©2001-2019 The Agile Manifesto Authors. This declaration may be freely copied in any form, but only in its entirety through this notice.

The Agile Manifesto. https://www.agilealliance.org/manifesto-download/

The Agile manifesto starts off by saying: "we are uncovering better ways of developing software by doing it and helping others do it." but there are applications of agile that do not have anything to do with software development. Indeed, embracing agile outside the realm of software has caught on. In the pre-Manifesto times of 1998, Lorraine Haricombe and T.J. Lusher edited "Creating the Agile Library: A Management Guide for Librarians". At the time the book was meant to be a response to the rapid changes in the library environment due to the impact of technology in society and education – an environment that has been stable for decades. The book was well received. Reviewers acknowledged the need of an "agility of mind, innovation and public entrepreneurship" to cope with continual change by building on the core principles of librarianship (Hendry). Likewise, the Library 2.0 movement focused on librarianship and the need to adapt to change and interact more with users. This eventually culminated in a Manifesto for Librarians as defined by Laura Cohen (Cohen, 2006). Today many libraries practice agile principles when implementing software projects. A few have even adopted them in managerial practices.

Agile values and principles in organizational context at SU

In the two decades since its creation, the Agile manifesto's four values and twelve principles have been embraced to varying degrees by countless individuals, institutions, and professions. Agile project management is deeply rooted in these principles but slightly modified to make sense in the project management, rather than software development, environment.

SU in organizational context

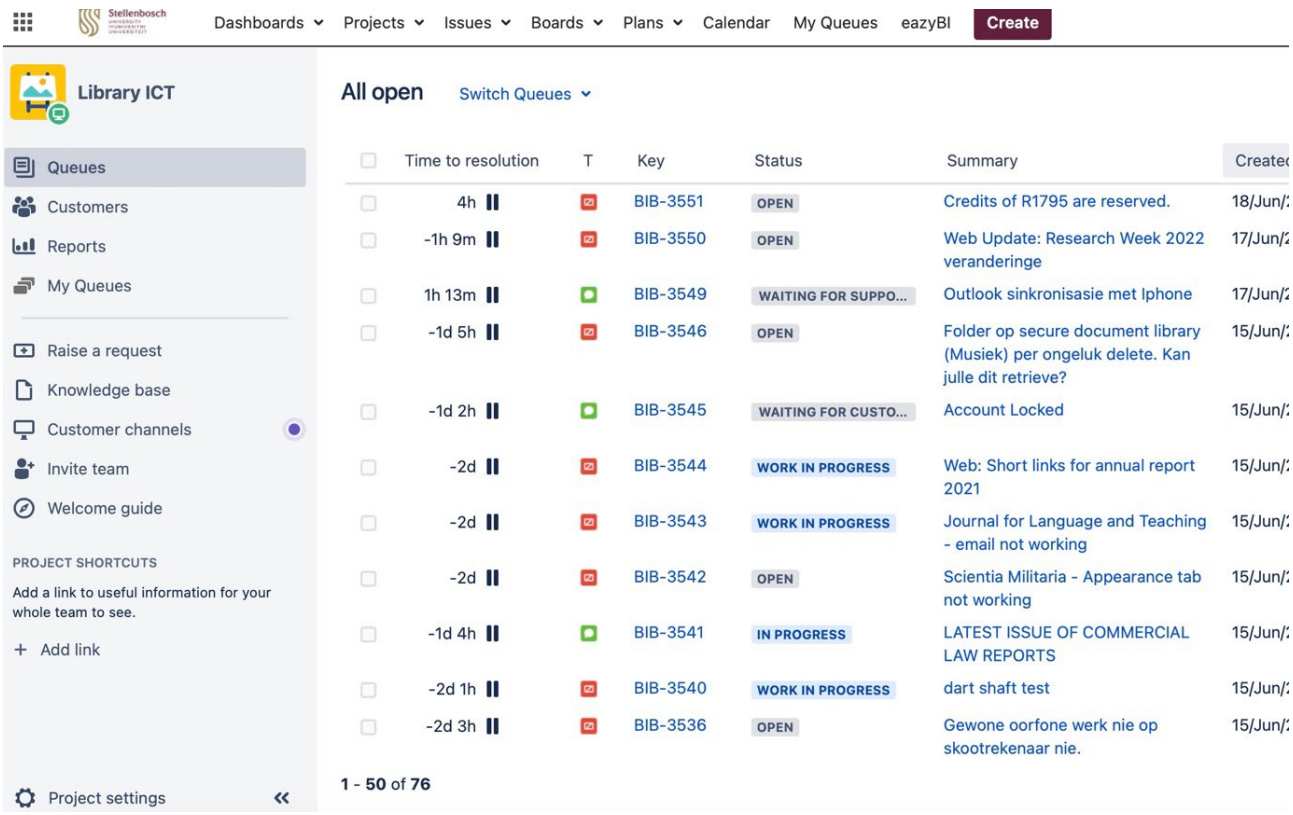
When considering how agile principles are applied within the organization context of SU, it would be fitting to first view the institution as an image of an organization. As an academic institution, SU can metaphorically be described in the image of an organization as an organism (Morgan, 1998). Universities do not operate in isolation of their environments. They are open systems that continually adapt themselves to their environments. A university as organization is made up of a set of subsystems. Each have their own characteristics, goals, and management styles functioning as part of an overarching global parent organization. This open systems approach is underlined by the contingency theory in which different approaches to management may be necessary to perform different tasks within the same organization. The Lawrence and Lorsch study (Lawrence, 1967) refine the contingency approach. Styles of organization may need to vary between organizational subunits because of the detailed characteristics of their sub-environments. The study shows that the degree of required differentiation in managerial and organizational styles between departments varies according to the nature of the industry and its environment. An appropriate degree of integration is needed to tie the differentiated parts together again. It therefore follows that there will be no single agreed on approach of how agile practices are applied to software development, or organization management. Each subsystem applies agile practices to fit their unique needs if they are guided by the correct principles.

Agile project management practices at SU

Agile project management practices were introduced at SU through the Division of Information Technology's (IT) software development team. The team is responsible for the project management of software development projects. For this they need a software project management tool. To this end the Jira system from Atlassian[6] has for the past 10 years been used for project and issue tracking. When launched by Atlassian in 2002, Jira was purely issue tracking software targeted at software developers. Later the application was adopted by non-IT organizations as a project management tool. Over time the application evolved from being a generic project management tool to also include agile project management features. In 2013, Atlassian announced a Jira service desk product. This possessed full service-level agreement support integrated with the agile-enabled workflows and tools. In 2018 Jira became the de facto solution for ICT service desk management at SU. This included the service desk support for the ITS team in the library as well.

https://www.atlassian.com/software/jira

[6] Atlassian website: Retrieved from https://www.atlassian.com



The Jira Service Desk system at the SU library

Fulfilling ICT support services with the Jira Service Desk

By doing service desk management with an agile-enabled tool such as Jira, the ITS team started practicing agile principles even before they knew there was a name for it. In retrospect, the team gradually started to identify with at least two of the four values of the manifesto. The ITS team started to change its way of working by practicing agile rituals such as holding regular meetings in a stand-up format, approaching tasks assigned to the team from a Jira backlog perspective, and maintaining documentation with a minimal-specification approach within Confluence, a Jira wiki package. Working collaboratively on one service desk and software project management solution with other ICT stakeholders at SU, the team also started to view itself from a cross-functional team perspective.

Improving ICT operations management through agile principles, methods, and tools

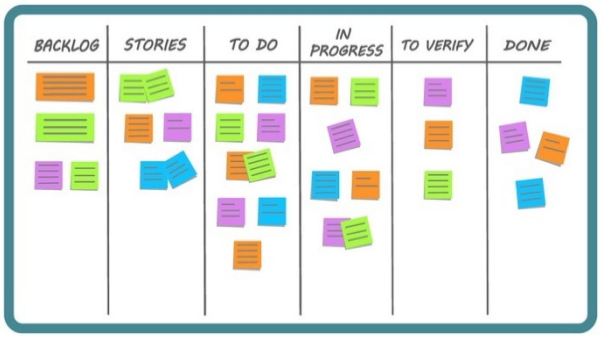
The ITS team was practicing agile principles out of due diligence to adhere to good ITIL[7] (Information Technology Infrastructure Library) standards when the COVID-19 pandemic forced the library's workforce to start working from home in early 2020. Pivoting to an online-only format forced the team to become completely reliant on the Jira service management system to field support calls and manage tasks and projects remotely. With this came the need to improve how task assignment was managed in an online environment, prompting the ITS team leader to explore further functionality on offer in Jira. In doing so, the team discovered the versatile task management tool, Kanban.

Using Kanban for ICT task management

Kanban is the Japanese word for "visual signal". It is an action performed through the physical passing of a card. The name Kanban comes from two Japanese words, "Kan" (看) meaning *sign*, and "Ban" (板) meaning a *board* [8]. Often our work is invisible or intangible. Kanban helps us to visualize our work, maximize flow (or efficiency), and limit Work in Progress (WIP) items. Kanban project management has been around since the late 1940s. It was studied by Toyota to use the rate of demand to control the rate of production of its vehicles. The car company applied it to their Lean manufacturing model, known as the Toyota production system.

[7] Axelos. ITIL® 4: the framework for the management of IT-enabled services. Retrieved from https://www.axelos.com/certifications/itil-service-management

[8] Toyota Production System. Retrieved from https://kanbanzone.com/resources/lean/toyota-production-system/



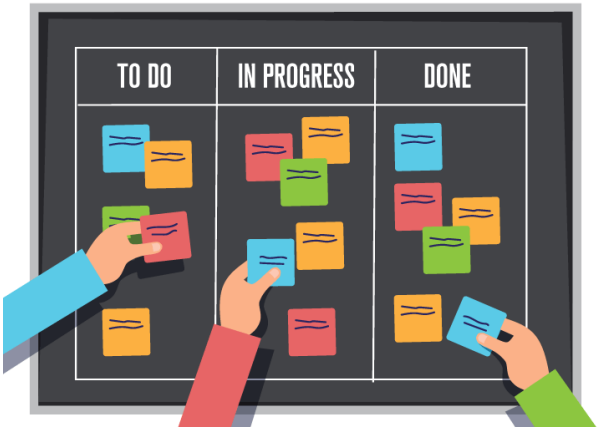
A simplistic Kanban zone board.
https://www.techtarget.com/whatis/definition/kanban

Today, Kanban software is a versatile task management tool used in many different industries. The core principles of the framework are timeless and applicable to almost any industry. Software development teams have found success with the agile practice and can begin practicing Kanban methods with little to no overhead once they understand the basic principles. Unlike implementing Kanban on a factory floor – which would involve changes to physical processes and substantial materials – the only physical things software teams need are a board and cards. Even those can be virtual. Kanban boards can be built on windows or walls, or with digital tools such as Trello and Jira. Moving the task assignment boards from the traditional office space's whiteboard format to the online environment, is what the ITS team found to be the most wanting when they started to work from home.

Starting a Kanban board

One of the core principles of Kanban is that it starts with what you do now. Kanban respects roles and responsibilities so you simply apply the Kanban methodology on how you currently work. Its purpose is to categorize all the stages a work item flows through from something you have not started to something that is done. This is referred to as a workflow. Each stage in the workflow has its own column. A basic Kanban board has a three-step workflow: To Do, In Progress, and Done. A work item (or card) flows through the columns on the board from the To Do stage to the Done.

Kanban cards should be small enough that a team can make progress on them in a reasonable amount of time, i.e., it should not take weeks to move the card forward. The task should also not be so small that the card represents every task a team member is working on. Creating the right balance is the first challenge the ITS team encountered when it started to use Kanban for ICT task management. Each service request logged on the Jira service desk can represent a Kanban work item. Care had to be taken to populate the Kanban boards in such a way that it remained an effective tool for work management.

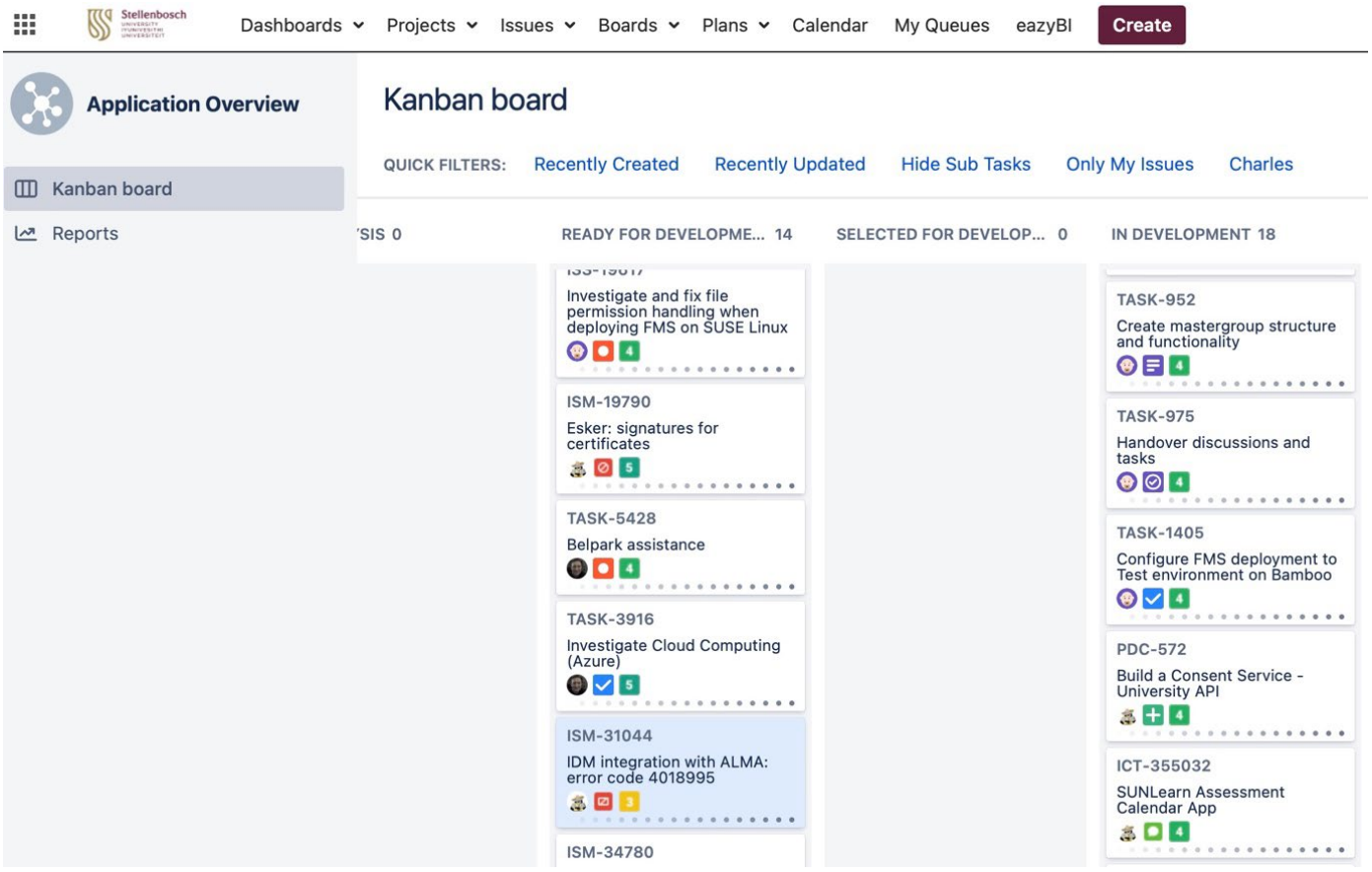


Using Kanban in Project Management: To Do, In Progress, Done

Benefits of using a Kanban board

Once the ITS team started building a Kanban board and filling it up with cards it started to experience one of the key benefits of Kanban: seeing cards that bunch up reveals a bottleneck in a workflow. With Kanban the ITS team is now actively, and in real-time, measuring cycle times of work done. This eases the process for the ITS team leader to measure how effectively and productively the team members are meeting their Key Performance Areas (KPA). The Kanban methodology relies upon transparency of work and real-time communication of capacity, and the Kanban board therefore is seen as the sole source of truth for a team's work. A further benefit of Kanban is to actively manage the WIP limits of team members. This can be visually measured to help avoid too much multitasking which kills the efficiency of team members. Before using Kanban, the ITS team leader would assign new tasks to team members which team members were still stuck on previous assignments. Leaders were not always actively realizing where team members needed assistance to finalize a task. Kanban has helped the ITS team leader to become a better manager at an operational level. It has helped the team to become more self-organized and self-managed at fulfilling ICT support tasks. Kanban helps each team member understand why and how tasks are assigned to them.

Although Kanban might seem like something simple that you have been applying in your work environment for some time already, explicitly thinking about the Kanban principles when applying these, changes the way you approach and use it. Putting a name to a workflow practice can be the first step in tackling it. This is what Kanban does. Once you name a workflow you can start to identify with working more easily with its methods and to its strengths.



A Kanban board for software integration projects at SU

Summary

Transitioning from the ITS team's previous GUI-based helpdesk management software, Heat, to a web-based and agile-supported helpdesk system such as Jira, allowed the ITS team to become familiar with agile methodologies, methods, and tools. As a method, Kanban remains a darling among service-oriented teams like IT or human resources. If you are a queue-oriented team like IT, then Kanban provides a solid foundation for your agile practices.

Little about fulfilling ICT operations relates directly to the Agile manifesto's four values. Over time though, the ITS team did manage to adopt an agile mindset to work with greater autonomy, trust, and flexibility. Without adopting an agile mindset, the greatest risk for any team is that it continues to work as before, and that no change will happen in its workflows and operations. The way a team lives those values today – whether it uses a specific methodology by the book or blend it with elements of another, such as Kanban – is entirely up to the team and the organizational context in which it functions.

References

Beck et al. (2001). Manifesto for agile software development. Retrieved from http://agilemanifesto.org/

Cohen, Laura (2006). A Librarian's 2.0 Manifesto. Retrieved from https://web.archive.org/web/20070209003635/http://liblogs.albany.edu/library20/2006/11/a_librarians_20_manifesto.html

Cervone, H. F. (2011). Understanding agile project management methods using Scrum. *OCLC Systems & Services: International Digital Library Perspectives*, *27*(1), 18–22. https://doi.org/10.1108/10650751111106528

Hendry, J. D. (2000). Creating the Agile Library: A Management Guide for Librarians. *Library Review, Vol. 49 Iss: 1*, pp. 40 – 48

Kanban Tool. Retrieved from https://kanbantool.com/kanban-guide/kanban-history

Lawrence, P., and Lorsch, J., "Differentiation and Integration in Complex Organizations" *Administrative Science Quarterly 12*, (1967), 1-30.

Morgan, G. (1998). *Images of Organization*. Berrett-Koehler Publishers. https://books.google.co.za/books?id= a_W8jXt4_eMC

Schwaber, K. (2009), detailed information on Scrum methods available at: http://www.scrum.org

New Horizons for Artificial Intelligence in Libraries Conference Review

IFLA IT Artificial Intelligence (AI) satellite meeting
Galway, Ireland

Lynn Kleinveldt, kleinveldtl@cput.ac.za
Lecturer, Faculty of Business and Management Sciences
Cape Peninsula University of Technology, South Africa

The Hardiman Research Building for research in the humanities and social sciences, National University of Ireland

Introduction

In 2022, The IFLA IT Section sponsored an important new special interest group in Artificial Intelligence (AI). In July, the group hosted, an inaugural satellite conference satellite New Horizons in Artificial Intelligence in Libraries taking place at the National University of Ireland in Galway. This first formal event hosted by the IFLA AI SIG, was a historic gathering of participants drawing wide attention in speakers and audience and follow-up sessions at IFLA's main Congress in Dublin.

Introductions: The National University of Ireland at Galway

IFLA IT AI Satellite conference delegates arrived in Galway and were graciously offered a tour of the historic Irish university. Delegates were informed about the rich cultural background of the National University, (founded, 1845) still thriving today. The university opened its doors, starting with enrolment figures of 68 and currently celebrating 175 years with 18,000 students and ranking 259 in QS world university metrics. During the tour, IFLA IT AI delegates had the opportunity to witness the newly constructed engineering building and marvel at the exterior of the Ryan Institute, named after the late Dr Tony Ryan, founder of Ryanair. The Ryan Institute is the university's largest research institute and significantly contributes to the United Nations 17 Sustainable Development Goals.

Pre-Satellite Meeting

O'Connell's Pub provided an opportunity for all presenters to engage with the chair and IFLA IT section members before the start of satellite. It was also an opportunity for them to experience Galway, a town known for both its friendly academic and arts and humanities orientation. Over a pint of Guinness or Galway-produced Pils, the group of AI library delegates exchanged in discussions ranging from recommender systems to AI library reference services running to chatbots and AI paths for libraries.

NUI, Galway University Library Conference Overview

The conference began with a welcome by the IT section's chairperson, Dr. Edmund Balnaves, introduction by Mr John Cox, University Librarian. This was followed by a keynote by Dr. Paul Buitelaar, Vice Director of the NUI Data Science Institute, NUI.

Dr. Buitelaar focused on key aspects regarding artificial intelligence, setting the scene for the satellite conference including a discussion of AI, knowledge graphs, chatbots, ethics and data visualization. AI and ethics turned out to be a key theme of the conference ranging over topics from AI fake news to AI and bias, genders exclusion and AI data language models and translation. Buitelaar gave a great example of "she is a nurse" which, when translated from Bulgarian by an AI, becomes "he is a doctor" in many AI language models. Two days of presentations followed. This summary does not overview but glances at a few of these to give a flavour of this great event. Full abstracts and slides are available here.

Dr. Ray Uzwyshyn: Building Library AI Infrastructures

Dr. Uzwyshyn's presentation focused on enabling library AI capacities and how research data repositories may serve as a foundation for building library AI infrastructures. An academic library may play a strong role in helping a university's researchers set solid foundations for AI and their research with a data research repository. Digital scholarly ecosystems around the repository, library staff and this library digital ecosystem can also serve as a great foundation for university AI researchers. Open-source software components such as ORCID, Dataverse and Open Journal Systems were discussed as support

mechanisms for enabling researchers with visibility of their AI projects, sharing of data and dissemination of research. An example was also given to show open science possibilities from Stanford University and a cancer detection AI algorithm utilizing online digital images needed for the AI neural net training. The potential for global open science was illustrated through the continued trajectory of this data lifecycle Recommendations included recently formed global library AI related groups such as The Carpentries, AI4LAM and recent global library AI conferences such as Fantastic Futures.

Dr. Mojca Rupar Korosec: Artificial Intelligence is Already in Libraries, Let's Master It

Dr Rupar Korosec employed a philosophical approach focusing on AI ethics and libraries. Key events were highlighted with regards to AI in libraries ranging from the European Union Intellectual Property Office (EUIPO) releasing its findings on their study on the impact of AI on the infringement and enforcement of copyright and designs (March 2022) to Data Ethics in the EU (Dataethics.eu) and GDPR guidelines. Dr Korosec also discussed how in 2019 the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) released a report on the ethics of artificial intelligence. In 2021 UNESCO also adopted a recommendation on the Ethics of Artificial Intelligence at the 41st general conference session. Rupar Korosec mentioned the OECD.AI, a Policy Observatory and unique source of real-time information, analysis and dialogue designed to shape and share AI policies across the globe.

Dr. Lynn Kleinveldt: AI to Assist IL Integration to Support Students Acquiring 21st Century Skills for Work

Dr Kleinveldt highlighted the challenges students face with research and how AI incorporated into reference management tools and Microsoft Office could address this challenge. She also touched on the topic of algorithmic literacy and the importance of adding this new AI literacy as a key literacy in concert with the more well-known historic library literacies of information literacy and digital literacy.

Lynn Kleinveldt Presenting, IFLA AI

Panel discussion comprising of all presenters at the satellite

Jean-Philippe Moreux: AI and Image Analysis at the BnF: Experience Feedback

Jean-Philippe highlighted aspects on innovative and creative image based projects conducted by the Bibliothèque Nationale de France (BNF) and opened to the general public online. It was pointed out that within a library setting developers now work at the forefront of AI projects with librarians and archivists on many prototype projects testing the potential of new AI developments. Moreaux's discussion included BNF's successful AI projects GallicaPix – an experiment with deep learning on heritage images to retrieve and create hybrid images from archival originals, GallicaSnoop – a framework dedicated to large-scale image retrieval and DataLabs, an initiative for better AI opportunities for transdisciplinary projects.

Scenic view from the Corrib Princess River Cruise

Galway River Cruise

An evening reception, dinner and river cruise on the Corrib Princess sponsored by Ex Libris was also held after the first day of presentations, panels and workshops. It was the perfect way for the group to unwind, informally gather and enjoy local Irish folk songs, fine wining, dining and great conversation while passing castles and swimmers, along the wild Atlantic way.

Panel Discussion and AI Ethics Workshops

Presenters also engaged in a lively panel discussion, and later, a question and answer period with the audience giving participants the opportunity to discuss further AI topics and discuss key issues facing libraries globally in the current AI landscape today. Dr. Andrew Cox, of the University of Sheffield, subsequently conducted an AI Library Ethics workshop with case studies of current AI ethics scenarios. Participants divided into groups to consider prevalent thorny issues of current global AI ethical issues.

National University of Ireland Library and Archives Tour

An NUI library tour also allowed participants to engage with local library staff and visit the NUI library archives and special collections as well as library Makerspace and Information/Data Literacy Hub, supporting the library's students and academic user community.

Dr. Andrew Cox: Keynote Address: AI is the new TA in the classroom

Dr. Cox spoke broadly on AI and ethics ranging from current student learning to what it means for libraries to come to terms with what AI means for libraries and considering ethical issues in tandem with the technological. Chatbot reference queries were discussed and, also the current lack of widespread implementation in libraries not yet ready to hand over reference questions and the need for further consideration of ethical issues. Cox also touched on AI's analysing, predicting, or nudging of user behaviour with individuals and groups and how this could be both positive and detrimental to individuals and democratic societies.

Karolina Andersdotter: How 100 Library Professionals Increased their AI literacy

Dr Andersdotter spoke about AI literacy and being able to "critically evaluate AI technologies, communicate and collaborate with AI, and use AI as a tool online". She discussed an Online Course (OER) developed by the University of Helsinki to "demystify AI" that library professionals in Helsinki had enrolled in and created a study circle for as a "pedagogical method to further discuss and absorb what they have learned". Padlet, a noticeboard online tool, was used to share discussions and recommendations and the group study circle made learning AI together less cumbersome and more fun. Various other examples of free online courses on AI were given (i.e. Elements of AI.com) and the study circle method explored for learning the sometimes daunting new world of AI programming, algorithms, methods and complexities.

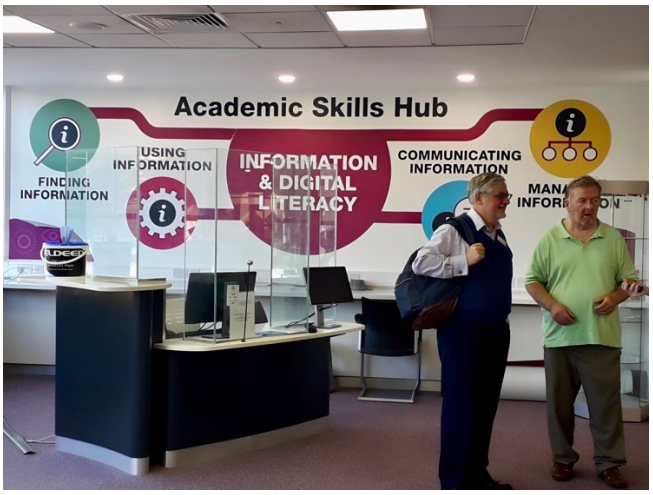
Florian Engel: Automatic Indexing Using AI Methods – A Project Insight at the German National Library

Florian Engel, from the German National Library, presented a fascinating presentation on AI and subject cataloguing at the German National Library (DNB) using the Annif AI open source tool for the automatic subject classification of its Integrated Authority File. Florian raised both large ethical questions but also central questions regarding subject cataloguing and how this is currently accomplished from the top. He showed how natural language processing could be used to create new methodologies for creating taxonomies and the possibilities and their intriguing challenges.

Dr Edmund Balnaves: Building Open Source Chatbots Workshop

In Dr. Balnaves' Chatbot workshop, conference participants were introduced to how to build chatbots using an open-source platform and application called Botpress. Participants were challenged to take up this new software which presents a spectrum of fascinating choices and easier methodology than programming through raw code. The workshop. more than anything, allowed participants to witness first-hand the extent to which chatbot technology and implementation have advanced. Chatbots are able to be customized to an amazing degree. Dr. Balnaves importantly pointed out

that to build a successful chatbot, one needs to start with designing the chatbot's dialogue flow and to structure the chatbot's bot dialogue hierarchically and purposefully and build a workflow which corresponds to larger purposes and conceptions of the bot's structure. It was a fascinating introduction to a new AI area and a great conclusion to a great satellite.



Concluding Remarks and Acknowledgements

The satellite conference on AI, held in Galway, was a great success for the group and IFLA. It was a satellite conference with the highest 2022 in-person attendance and subsequent months showed continued large interest in the presentations, now online and available here for further review and download of presentations.

To note, presenters have also been approached to develop their AI and library presentations into a book to be published by De Gruyter in late 2023, early 2024. Overall, this was a valuable, perhaps historic opportunity, to engage with experts, unpack new challenges and showcase best practices.

IFLA IT Section meetup, Toners Pub (est. 1734)

Fitting for Irish culture, the tradition at IFLA conferences is for the IT section committee to meet and network. This occurred throughout the satellite, continued at the main WLIC conference sessions and concluded with a final group meeting at Dublin's oldest pub, Toner's, established 1732. This experience, in its entirety, was a great opportunity to reflect together on the satellite, and to plan future engagements with other sections during the main conference.

It was also a great honour to represent my university, the Cape Peninsula University of Technology (CPUT) and the IFLA IT Section in Ireland. I was able to develop collaborative relationships and share best practices in taking teaching, learning and research support forward to the next levels. I look forward to collaborating with international colleagues and making a further contribution to teaching, learning and research during the 2022-2023 IT section's activities.

IFLA IT Section

The Information Technology (IT) Section promotes and advances the application of information and computing technologies to library and information services in all societies, through activities related to best practices and standards, education and training, research, and the marketplace. The scope covers IT for creation, organization, storage, maintenance, access, retrieval, and transfer of information and documents for all types of libraries and information centers; IT for the operation of libraries and information centers; and related management and policy issues. Of primary importance are applications of IT for supporting access to and delivery of information. In recent years, the uses of use of technology in libraries have expanded to cover improved machine learning and AI techniques, digital humanities, and data analytics.

The section meets annually at the IFLA Congress; in between congresses, members collaborate with other Sections on programs and workshops. There are election ballots every two years as members complete their 4-year term. The IT Section is one of the largest in IFLA with over 300 members from nearly 80 countries, all types of libraries, and a range of disciplines. We welcome all members (http://www.ifla.org/membership).

The IT Section's website at http://www.ifla.org/it has news and resources regarding activities of the Section, session minutes, publications, and membership details.

The IFLA-IT email list provides a forum for members to exchange ideas and experience in the use of information and communication technologies in libraries. The list address is ifla-it@iflalists.org, and subscription is at https://mail.iflalists.org/wws/info/ifla-it.

The Trends & Issues in Library Technology (TILT) newsletter is published twice a year in June/July and January.

Primary Contacts

Chair
Edmund Balnaves, Prosentient Systems, Australia; ejb@prosentient.com.au

Section Coordinator
Elena Sánchez Nogales, Biblioteca Nacional de España, Spain; elena.sanchez@bne.es

Secretary
Cory Lampert, University of Nevada Las Vegas, United States; cory.lampert@unlv.edu

Information Coordinator
François-Xavier Boffy, Université de Lyon, France; francois-xavier.boffy@univ-lyon1.fr

Standing Committee Members

- Alenka Kavčič Čolić, National and University Library, Slovenia; alenka.kavcic@nuk.uni-lj.si
- Almudena Caballos Villar, Biblioteca de la Universidad Computense de Madrid, Spain ; acaballo@ucm.es
- Anna Tereza Barbosa da Silva, Oslo Police District, Norway; anna.tereza.barbosa.da.silva@politiet.no
- Katherine Zwaard, Library of Congress, United States;kzwa@loc.gov
- Leda Bultrini, ARPA Lazio, Italy; leda.bultrini@gmail.com
- Lynn Kleinveldt, Cape Peninsula University of Technology, South Africa; lynn.kleinveldt@gmail.com
- Maria Kadesjö, National Library of Sweden; maria.kadesjo@kb.se
- María Loretto Puga, Biblioteca del Congreso Nacionalde Chile, Chile; mpuga@bcn.cl
- Ngozi Blessing Ukachi, University of Lagos, Nigeria; nukachi@unilag.edu.ng
- Peter Leinen, Deutsche Nationalbibliothek, Germany;p.leinen@dnb.de
- Qiang Xie, National Library of China; XIEQ@NLC.CN
- Ray Uzwyshyn, Mississippi State University Libraries, United States; rju13@msstate.edu
- Sogoba Souleymane, University of Ségou, Mali; Sogoba.souleymane@gmail.com
- Sylvain Bélanger, Library and Archives Canada; sylvain.belanger@canada.ca
- Wouter Klapwijk, Stellenbosch University, South Africa; wklap@sun.ac.za
- Yeon-Soo Lee, National Library of Korea; myth4ys@korea.kr

Mentor

May Chang, University of Cincinnati, United States; may.chang@uc.edu

Big Data SIG Convenor

Patrick Cher, National Library Board, Singapore; patrick_cher@nlb.gov.sg

Artificial Intelligence SIG Convenor

Andrew Cox, University of Sheffield, United Kingdom; a.m.cox@sheffield.ac.uk