SPECIAL BIG DATA ISSUE

JANUARY 2022

Trends and Issues in Library Technology

International Federation of Library Associations

IT Section IFLA IT

TRENDS & ISSUES IN LIBRARY TECHNOLOGY

IFLA Information Technology Section

Special Issue on Big Data

Feature Articles

Invited Articles

Departments

Editor's Notes

Special Big Data Issue

Dear Colleagues,

I would like to introduce myself as the new editor of our IFLA IT Section "Trends and Issues in Library Technology". I feel honored to be beginning as editor for this "Special Big Data edition" as it has already been both a unique and exciting experience to be working with our global authors and communications team.

I'm also very proud of this issue with articles on the latest trends in Big Data, Research Data Repositories, Metadata Education, Infrastructure Possibilities and Data Recommender Systems from around the globe. Sara Wingate from University College, London, UK presents a unique case study of data curation as collaboration between academia and the British Library. Tan Wen Sze, Lim Chee Kiam and Chow Ying Yi of the National Library Board of Singapore innovatively show how Singapore's National Library uses machine learning, data, and a recommender system to recommend books for patrons. Ingrid Mason and Steven McEachern from the Australian National University and Australian Data Archive overview diversifying collaboration, network complexity and big data. Samantha Fritz of the Archives Unleashed Project, Canada speaks about challenges and solutions for Web Archives and Big Data Collections in Libraries and I present a best practices case study from Texas State University Libraries, US for Research Data Repositories and global ecosystem possibilities.

To begin, there is also an excellent historical contextualization and synthesis of our articles from larger Big Data perspectives by Wouter Klapwijk, South Africa and our new chair, Edmund Balnaves overviews our ambitious IFLA IT Section upcoming plans. As an invited piece, I would also like to mention Celine LeClaire's excellent digital roadmap article, from the National Library of France.

Thank you also to our IT Section Communication Director, Francois-Xavier Boffy, IFLA IT Section authors and wider team. They have all done an amazing job contributing to an excellent kick off issue

I hope you enjoy reading and profit from the articles presented. We have an exciting year planned. Our next issue will focus on another very current topic, AI and library possibilities arising from several recent AI global symposia (Paris, France Fantastic Futures, South Africa, AI, and Robotics). A new AI SIG for our IFLA IT section is also planned.

I look forward to your comments as well as your contributions. Our submissions' door is currently open, so please feel free to send a note, or submit a proposal or query, for subsequent newsletter issues.

I also encourage all of you to join and visit our new IFLA IT Section Social Media Facebook Group available at: https://www.facebook.com/groups/iflaitsection

To all our readers, I would like to extend my best wishes for a safe, healthy, productive, and peaceful 2022. I look forward to meeting you, both online and at our upcoming annual conference in Dublin.

Sincerely,

Ray

Ray Uzwyshyn, Ph.D. MLIS ruzwyshyn@txstate.edu
Editor, Trends and Issues in Library Technology

IT Section Chair's Corner

IT Section Update and Action Plan

2021 has seen Artificial Intelligence blossom into various Library discussions. The IT section partnered with University of Pretoria for their symposium on AI and Robotics and there is clear interest in the library community in all aspects of AI. This includes the ethical landscape around algorithms, practical examples of use of AI, the impact of AI on the workplace and emerging AI research. This has confirmed our key action plan proposal to sponsor the formation of a Special Interest Group on AI within IFLA. We are also planning webinars and a Dublin WLIC conference session or satellite on the AI theme. We are also looking for engagement with other groups such as ai4lam in this fast-developing area.

All of this also complements our continued engagement with the Big Data Special Interest Group. There have also been interesting discussions in the IT section regarding points of overlap and difference in the areas of Big Data and AI. The overwhelming feeling has been that that there is very strong interesting in both areas. New innovations are emerging in Research Data Services and Big data. These continue to demonstrate the value of this interesting SIG. We are also looking to expand our engagement with Research Data services, in conjunction with other sections in IFLA.

Cory Lampert is continuing our engagement with Linked data with the PCC pilot project and 2021 LD4 conference.

The IT section has also committed to continuing our successful webinar series through 2021/22. Some previous topics included:

- The 4IC and its relevant to libraries
- Research data services

Future topics of interest include:

- Cloud services & IoT
- Transition to Bibframe
- "Is it still worth to have a library website"?
- Library Research
- Data management services update
- Decentralization and interchangeable users built with Blockchain technology
- The future of the library.
- Standards that certify interoperability for technologies that are part of the Open

Knowledge ecosystem of tools that underpin library infrastructure

The communications team with our Information Officer Francois-Xavier Boffy continues to break new ground in Twitter and social media. And TILT of course continues as a fantastic forum for providing projects and news in the Library IT space.

We look forward with anticipation to the conference in Dublin 2021, emerging as we are from the Covid hiatus. Covid itself has tested all aspects of managing library and information services and maintaining contact and assistance to clients. The Covid crisis has also bought new insights into workplace models, online support, and information delivery.

All up, we will have an exciting year ahead in the IFLA Information Technology Section.

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

Contextualizing Big Data in Libraries

Wouter Klapwijk, wklap@sun.ac.za
Director IT Services, Stellenbosch University, South Africa

The Big Data revolution has provided a more powerful information foundation than any previous digital advancement. Industries, research institutions, academia and businesses can now measure and manage massive amounts of information with remarkable precision. This is also dependent on the extent of the investment made in Big Data technologies. This evolutionary step allows managers to target, provide finely tuned solutions and to use data in areas historically reserved for the "gut and intuition" decision-making process. The past few years have seen a significant rise in tools to deal with Big Data and its numerous data types. Libraries as well as enterprises are still only just beginning to understand how to best deal with their new assets and their varying attributes.

Libraries in the Big Data evolution

For the sake of simplicity, we could broadly organize Big Data types in libraries into three data groups, depending where on the digital libraries development timeline that we find ourselves:

The first data type group is mostly transactional metadata accumulated during the 1965 – 2000 period and which is stored in library management systems and institutional repositories. Web crawling projects as well preserve data in web archives and digitization projects that safeguard data in digitization platforms. Although mostly of a structured type and not Big Data per se, over time this data has accumulated substantially.

The second data type group is that data which libraries became familiar with during the period that is considered the formal birth and classification of Big Data, 2000 – 2015. During this timeframe web-based systems and social media networks gave rise to large volumes of unstructured and semi-structured data that were generated at a pace never experienced before. For libraries today these data types represent that data which we mine through data and text mining projects. Academic and research libraries participate in research projects by brokering access to data sources with copyright restrictions and helping to describe and classify large datasets. This is a continuation of mass digitization projects, indexing and retrieving data from non-traditional data sources such as government databases, and a massive step-up in the crawling and preserving of web-based data in web archives.

Although libraries were participating in generating and curating large amounts of these data types in this period, they were they not yet investing in Big Data technologies and human skillsets for further discovery and insight. Big Data properties of variety, veracity and volume can be attributed to data sources accumulated during this period. Libraries were starting to ask the question how the value in these data sources could be optimized for their patrons.

The third group is that data generated over the last five years by mobile and sensor-based systems which could be referred to as the Internet of Things (IoT). This area remains largely untapped for libraries. Building sensors, camera technologies, and the geographical plotting of our user base are starting to become a growing concern for libraries to service our patrons better. We can expect the introduction of intelligent and virtual agents and robotic applications in libraries to add to the growing domain of sensor and user-based data.

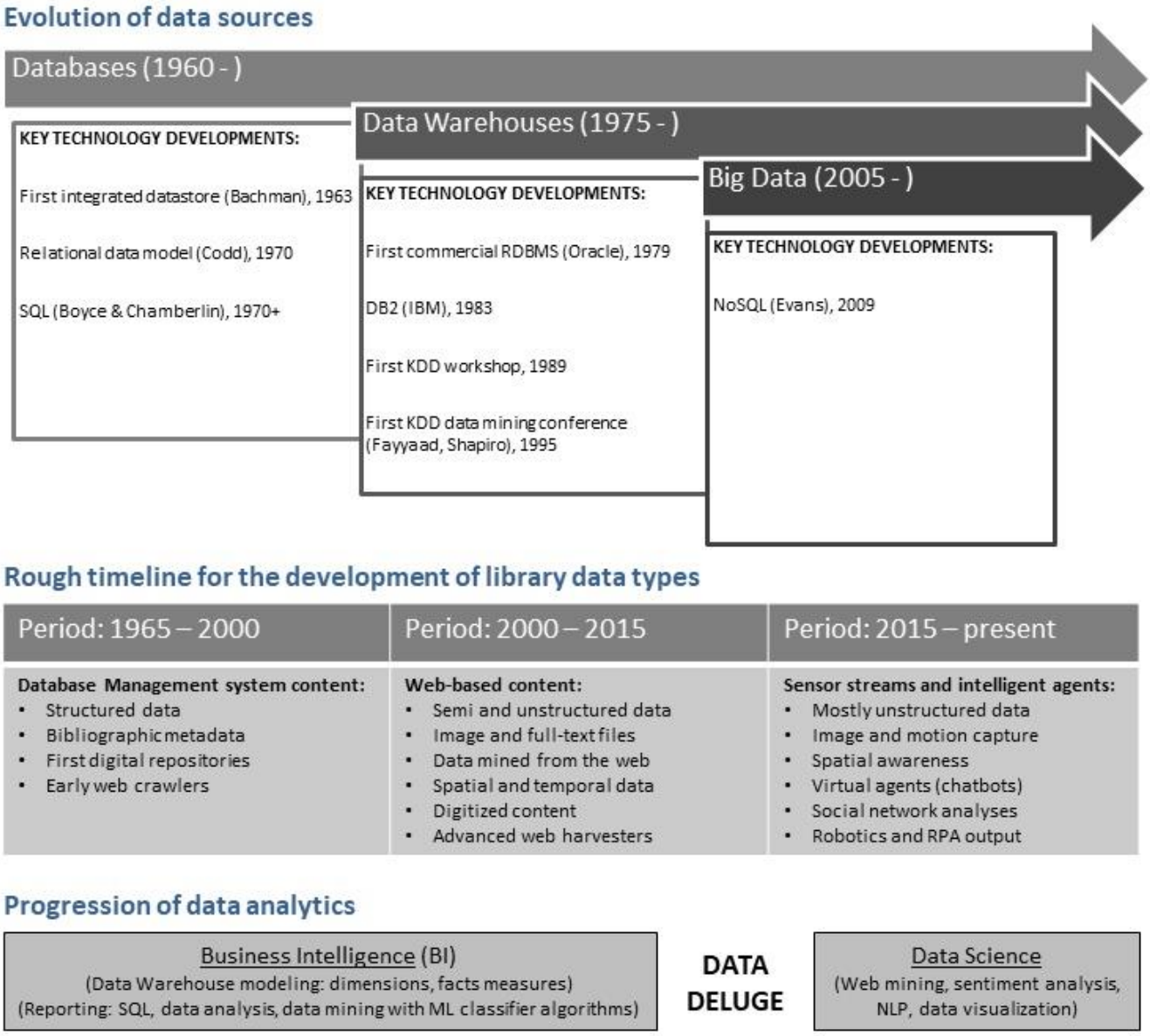


Figure 1: the evolution of data sources, data types, and data analytics in libraries

Trends in libraries working with Big Data

One could say that most libraries are behind the industry curve in adopting Big Data technologies to analyze and understand our new and continuing data sources in these three data groupings. Flexibility and agility are two states of mind useful in dealing with Big Data. Successfully exploiting the value of Big Data requires experimentation and exploration. As we have seen over the last 5 – 7 years this is no different to the experimentation and exploration needed to fruitfully engage with Artificial Intelligence (AI) applications. These are broad, diverse, and complex technologies with which all industries must come to terms. What is different is that libraries are adapting to the application of Machine Learning (ML) practices faster than to Big Data practices. Thinking that these two areas of concern in libraries are the same, or even dependent on one another, would be a mistake. ML as one area of AI can be applied on datasets that do not possess Big Data properties. Not all large amounts of data must be analyzed with ML methods.

Three concessions however should be made for libraries in the world of Big Data. From a library perspective the long-term benefit of data science is in finding the scientific questions we can ask, not in managing the velocity and volume of Big Data. Second, libraries have already adopted many methods to analyze Big Data which have commonalities with ML practices. Finally, it is generally accepted that the "bigness" in Big Data is relative if methods and applications of these new assets are applied appropriately, and to an organization or industry's competitive advantage.

Example Big Data applications in libraries

This raises the question of how progressive the library industry really is in responding to the challenges of Big Data. Our world is one where 90% of the approximate 70 zettabytes of data in existence in 2021 is unstructured, and generated outside of the realm of scholarly communication workflows. By way of reflection, we can take what successful big technology companies such as Google, Netflix, and Meta have in common, and group this issue's Big Data's Trends and Issues in Library Technology newsletter article contributions into those areas to illustrate what libraries are currently doing to address the Big Data challenges. These areas may be grouped as:

1. Digital in Nature

Big tech firms are born digital rather than having to go through a process of digital transformation. This is different from libraries which are some of the oldest institutions in the world. For example, in 300 BC the ancient Egyptians already tried to capture all existing "data" in the library of Alexandria. Being digital in nature implies that your systems are digital-ready for a Big Data world, such as deploying NoSQL and aggregate key-value stores as opposed to relational databases. The National Library Board of Singapore's article is exemplary in showing how libraries evolve their systems to be Big Data compatible by default, utilizing cloud technologies and machine learning techniques to create for example recommender services.

2. Data Intensive

Big tech firms use data as their natural resources to manufacture (digital) products and services. Samantha Fritz's Archives Unleashed Project article demonstrates by way of practical examples the impact tool building, community engagement, and collaborative partnerships have on expanding the visibility, access, and use of web archives. Data is captured in web archives in digital-born format since the inception of the Web in the mid-1990's and is curated according to international web standards.

3. Disruptive

Big tech firms start new disruptive trends, impacting on a global scale. Ingrid Mason's article shows how the CADRE platform project aims to diversify collaboration amongst libraries in a Big Data world. Through the process of harmonizing library and IT work, new collaborations are created on a global scale to further curatorial analytics in Big Data. This proposes a completely new, and perhaps unavoidably disruptive, mindset to the traditional way libraries collaborate on data work.

4. Information as Strategic Asset

Big tech firms utilize information strategically for their competitive advantage. Sara Wingate Gray's article similarly show how the British Library works with students in academia to augment and reuse digitized and digitally born data in data intensive student projects. In doing so they capture the imagination of students when using their cultural heritage collections, while at the same time helping to impart new skills and competencies.

5. Global

Big tech firms have a digital service offering that is available globally. Although libraries have a global presence, a lot more can be done to advance the uptake of digital technologies in libraries. Ray Uzwyshyn's article on Texas State University's Data Research Repository Ecosystem proposes concrete next steps to facilitate the roll-out of digital scholarship ecosystems on a global scale to institutions who either cannot implement it alone, or simply have not yet gone through these time intensive and laborious processes.

These use cases demonstrate excellence in working together to further the use, augmentation, and re-use of library data collections within different Big Data settings. Big Data is only going to continue to grow. New technologies will be developed to better collect, store, and analyze the data as the world of data-driven transformation moves forward at ever greater speed.

How will libraries keep up? Two clear opportunities present themselves. The first is to evolve our existing library technologies into the Big Data technology space This involves using document stores and graph databases as the default data management solution rather than as mere add-ons for unstructured, semantically rich, data. Secondly and to conclude, if there ever was a time to re-envision library collaboration and leverage the global network of libraries to their fullest extent, that time would be now.

Sara Wingate Gray, sara.wingate.gray@ucl.ac.uk
Arts and Sciences, University College London, UK

Figure 1. Screenshot of British Library's "Microsoft Books/BL 19th Century collection" homepage from URL: https://www.bl.uk/collection-guides

Introduction:

Information Through the Ages (BASC0033) is an undergraduate module, offered as an interdisciplinary elective through the Arts and Sciences department at University College London (UCL), exploring the concept of information. The module engages students in a critical, interdisciplinary examination of the role institutions and collections play in validating and verifying information and information sources. The elective scrutinises the interplay among audiences, politics, aesthetics, material forms and the socio-economic, technological, and socio-cultural elements in which information is situated.

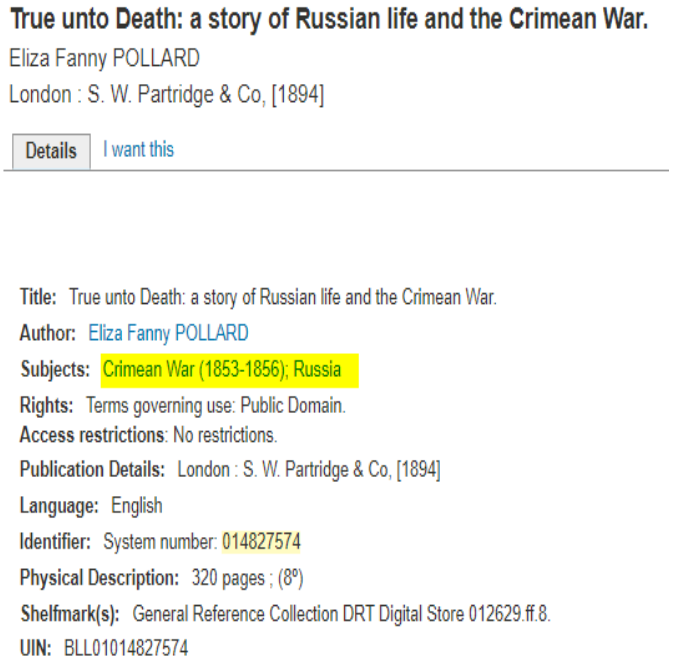
Alongside more traditional forms of assessment such as the scholarly essay, the module also provides students with the opportunity to work as data curators, through a partnership with the British Library (BL). Students work in groups using the BL's public domain datasets to produce projects which explore topics and themes within the data.

By transforming the British Library's public domain datasets, the assignment focuses on the BL's 19th Century/MS Books Collection's 60,000 records of metadata. These relate to approximately 25 million pages of out of copyright 18th and 19th century texts held in the BL's collection.

Students work in groups to discover, research, clean, refine and curate their own subsets of this original BL metadata. They structure their projects under topics or themes they discover in the data, such as travel literature, British colonies, queens of Britain, and dramaturgical literature to name a few.

Student co-collaboration in real-world environments

The module follows a research-based educative framework, providing students with the opportunity to engage in a real-world collaboration. This is both public-facing and provides the potential for real world impact factors. Student datasets are then fed back into the British Library's catalogue, providing richer, more accurate datasets, of benefit to the British Library's national research remit and future researchers alike.



Figures 2. Screenshots of catalogue records in British Library with student metadata added in the form of FAST headings. From URLs: http://explore.bl.uk/BLVU1:LSCOP-ALL:BLL01014827574

Students are afforded the opportunity to see their work disseminated to the public. This become of real-world use while the students have gained technical and archival research skills. These skills include software training and experiencing collaborative work practices such as report writing and project management.

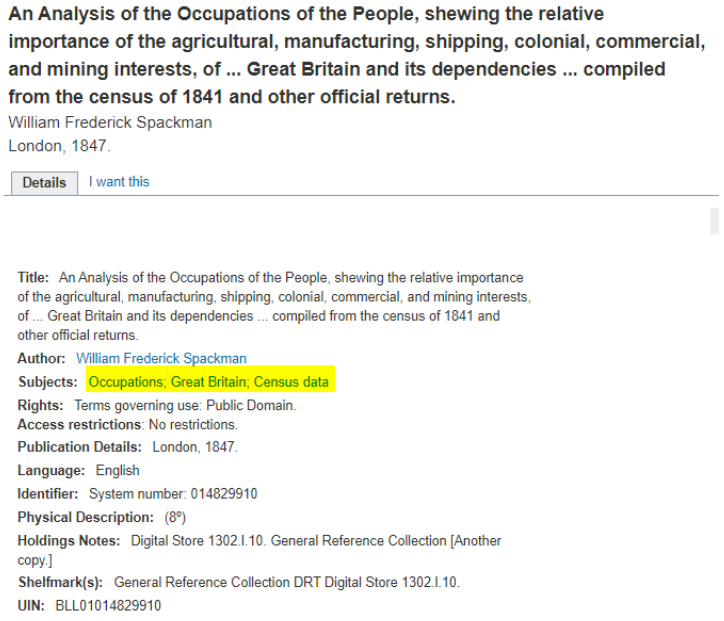
Providing students with external industry partners to collaborate with, can contribute an important fillip to their motivation and the learning experience overall. Importantly, they see their assessed work move beyond the confines of the academy to have an impact in the wider world. The work also provides a form of assessment which matches up to the reality of the world of tasks and work outside of the academy and provides evidence of their capabilities. This is an important aspect for employability and their future careers. Datasets produced by students are also held within the British Library's Research Repository, and as such suggest a long-lasting documentation legacy, provided in open access format.

Transforming and cleaning datasets

The module has currently produced student datasets which have led to the upgrade of 4,050 of BL records, in the form of FAST headings for topical, geographical and form/genre terms, based on the themes of the students' datasets, enabling richer catalogue records to be presented and searched.

Student's topic-based metadata sets importantly chosen by the student groups themselves provide interesting subject additions.

For example, two drama-related datasets have now enabled the addition of form headings for the performing arts such as playscripts and operas in this data subset. Topical terms for specific wars and battles were able to be added from one student group's focus on producing a data subset related to armed conflicts in the 19th century. Of course, important caveats, always apply to working with metadata, and as Victoria Morris (one of the British Library guest lecturers) noted in correspondence, "It's probably worth saying that not all of the records will now have subject/genre access that we can regard as complete or comprehensive. For example, record 014827574 (Figure 2 above) has had subject access for Russia and the Crimean War added, but no genre term to indicate that this is a work of fiction. However, the record is likely to be much more discoverable than it was previously; this is definitely progress in the right direction".



Figures 3. Screenshot of catalogue records in British Library with student metadata added in the form of FAST headings. From URLs: & http://explore.bl.uk/BLVU1:LSCOP-ALL:BLL01014829910

Other examples of records from the BL online catalogue updated using student datasets in this way include:

http://explore.bl.uk/BLVU1:LSCOP-ALL:BLL01014879786
http://explore.bl.uk/BLVU1:LSCOP-ALL:BLL01014831386
http://explore.bl.uk/BLVU1:LSCOP-ALL:BLL01014829910

Data curation training for students is focused on the use of OpenRefine, "a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data" (Open Refine, 2020). Students are also provided with cataloguing and metadata training documentation.

This includes a *Metadata Fields Guide for the MS Books Collection*, and a (brief) *Guide to Metadata Standards* by BL staff, alongside an introductory lecture on the BL dataset and lectures discussing cataloguing and metadata frameworks and standards.

Providing extra documentation and enabling further in-class discussions in these areas are important aspects of balancing both the opportunities and challenges of this type of partnership. This has relevance for projects where metadata quality and integration are essential factors. Drawing on the expertise of BL metadata staff in future iterations of the module (2019-2020; 2021-22), has been a successful enhancement here.

Figure 4. Screenshot from Open Refine, https://openrefine.org/

Student engagement and feedback

Student responses to their group project assignments, solicited in anonymous module evaluation forms at the end of the ten-week course in its first iteration (2018) pointed to this area, enabling the module to be augmented in this way. Other student feedback underscored the module's objectives regarding developing students' data curation and data analysis critical and practical abilities, while facilitating students to bring qualitative and quantitative methods, approaches, and understandings, to their data curation group projects. This reflects the integrative aspect which this interdisciplinary module seeks to foster:

"This was my first-time doing data curation so it was an interesting experience to figure out what skills I needed to work on to contribute positively to my group. Through the various trainings we had, I was able to come up with a good understanding of what data cleaning is and the processes associated with it, so all I really had to do was come up with a criterion to determine what was "good" and "bad" data."

Respondent A, BASC0033 Student Module Evaluation Form

"I felt that the talks given by the British Library was very interesting. Collaborating with them for our data curation was very helpful in learning the aspects of data science. However, I learned that it takes time and patience to go through data which is why I don't think I would undertake it again."

Respondent B, BASC0033 Student Module Evaluation Form

"I enjoyed the collaboration with the British Library as it gave me more insight on how curating data sets can be useful in a practical sense. I found the lectures on information privacy, collections, and knowledge to be interesting and think that what we explored in these lectures will be useful going forward. The talks from the staff were insightful. They were knowledgeable and able to answer questions that we had regarding curating the dataset. I enjoyed the project but found the group work challenging due to group differences. Overall, I have enjoyed taking this module."

Respondent C, BASC0033 Student Module Evaluation Form

Data from the first two years of the module, provided by the module evaluation forms, showed that 87% (year 1) and 89% (year 2) of respondents agreed that they "enjoyed the range of assessments for this module because they incorporated both traditional and non-traditional academic assignment elements (GIF-making; blogpost writing; essay-writing; data curation)". 83% (year 1) and 71% (year 2) of respondents agreed with the statement that "[t]his module helped me to make connections across different academic disciplines and disciplinary fields". It should be noted however, that module size numbers are not extensive (less than one hundred students to date) so the dataset should be treated with caution considering such a small sample size.

Conclusion

Providing impactful learning experiences in the context of industry partnership requires significant planning, pedagogical design, and development. Partners should also be willing to go the distance. Aiding and abetting the growth of public domain library datasets and enhancing resource discovery for library catalogues can also be observed as tangible outcomes reflecting partnership and academic commitments to the common weal: co-collaboration in its truest sense.

Project partners: ***Alan Danskin*** *(Collection Metadata Standards Manager, British Library),* ***Mahendra Mahey*** *(British Library Labs);* ***Victoria Morris*** *(Metadata Analyst, Metadata Standards team, British Library);* ***Sara Wingate Gray*** *(Lecturer, Arts and Sciences, University College London);* ***Stella Wisdom*** *(Digital Curator, British Library).*

Research Data Repositories and Global Scholarly Ecosystem Possibilities

Ray Uzwyshyn, ruzwyshyn@txstate.edu
Director, Collections and Digital Services, Texas State University Libraries, US

Figure 1. Texas Data Repository, https://data.tdl.org/

Introduction

Online networked data research repositories allow sharing and archiving of research data for experiments and research studies. This opens data to modern interoperability and metadata standards for search and retrieval. Located in open scholarly ecosystems, data research repositories are currently being leveraged to accelerate global research, promote international collaboration, and innovate on levels previously thought impossible.

Research data repositories open possibilities for collaboration with others. They link data to further content from online publications and aggregation tools making associated documents easily accessible. This paper pragmatically overviews such a data-centered ecosystem at Texas State University Libraries, a large state university research library in the United States. The research then goes on to speculate on possibilities for global research data repository networks utilizing the models presented.

Texas State Data Research Repository

A Data Research Repository is now a necessity For STEM disciplines (Science, Technology, Engineering and Math), the Social Sciences, Open Science, or any discipline which utilizes data-driven research methodologies. Data research repositories are suitable for any library, research, or academic institution with large scientific and social science research and datasets.

This new class of open-source software becomes the infrastructure of choice in a digital scholarly ecosystem and academic library infrastructure. Library human resources and information technology models are worked out and promising future possibilities are manifest. Texas State University Libraries uses a customized consortial version of Harvard University's open-source data repository software - Dataverse. Dataverse is a multi-tiered open-source platform for publishing, archiving, and sharing research data.

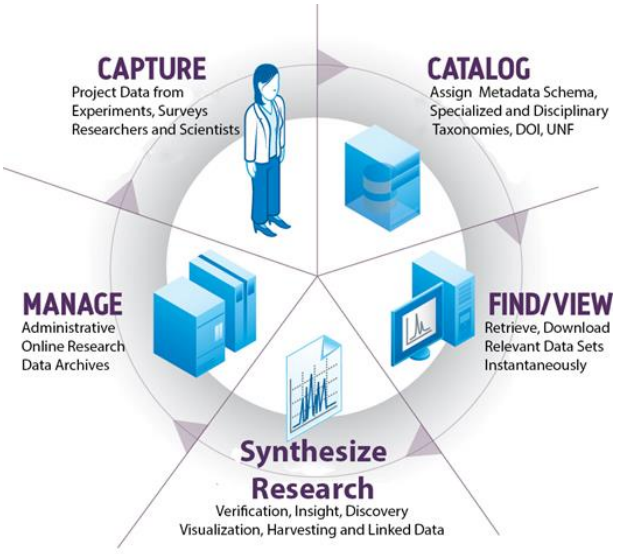


Figure 2. The Data Research Cycle

Dataverse and the Texas Data Repository allow researchers to publish, track, discover and reuse other researchers' data and participate in the larger data research cycle. The data repository enables researchers to operate independently, assigning metadata. Specialized disciplinary taxonomies are also possible. A data repository specialist assists in archiving data archives and creating unique metadata schemas. The Dataverse platform allows researchers to search internally within an institution and retrieve institutional data. Simultaneously, it also enables search access across other institutions.

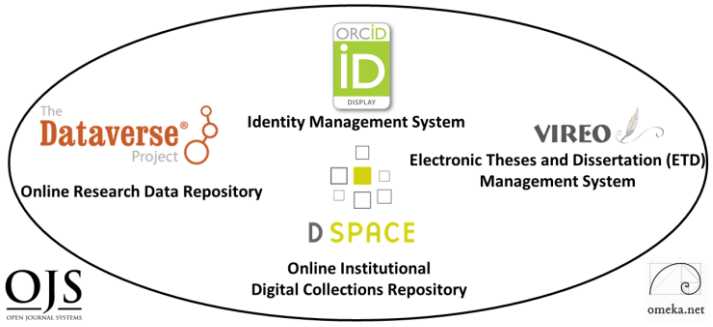


Figure 3. Texas State University Libraries Scholarly Research Ecosystem

Within the Dataverse Structural Model, the researcher may search and retrieve data from other researchers who may be working locally. Alternatively, they may search more inclusively for similar research data at other institutions in the Texas Digital Library Dataverse Consortium.

The Texas State University Libraries Online Data-centric Research Ecosystem consists of six basic software components, two primary components and four tertiary. The primary components allow content to be shared and archived. The tertiary components allow communication among internal and external networks.

Central to this scholarly ecosystem are primary components of an Online Research Data Repository (Texas State University DATAVERSE) and an Online Institutional Collection Repository (Texas State University DSPACE). These primary components serve to house content - research data and research papers.

Four other software communication components comprise tertiary components: an Identity Management System (ORCID), Electronic Theses and Dissertation Management System (VIREO), Academic Journal System (Open Journal Systems 3) and User Interface Software (OMEKA). The tertiary components connect the ecosystem components with each other to various areas internal to the research institution (i.e., Electronic Theses and Dissertation Management System) and externally to larger global networks (ORCID Identity Management Software). Typical characteristics necessary for this type of larger digital research ecosystem are open-source software, active developer communities, stable software releases, easy configurability (i.e., open API's), customizability, and connectivity.

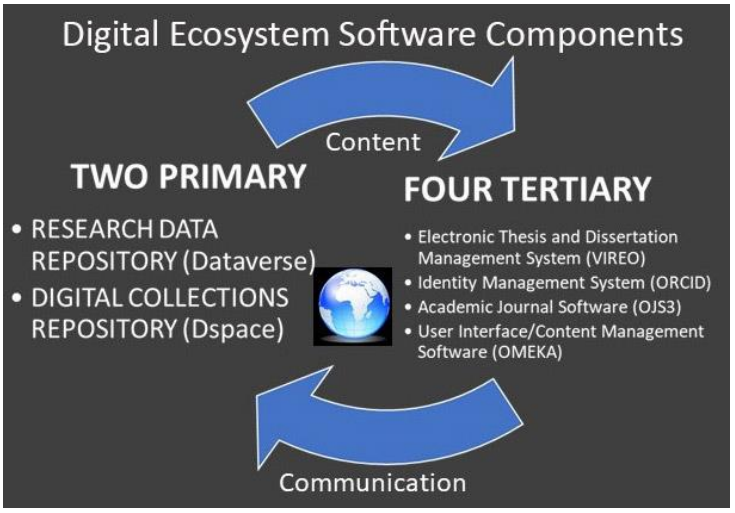


Figure 4. Digital Research Ecosystem:
Primary and Tertiary Software Components

All ecosystem components may be interlinked with the Data Repository. Research papers residing in the Digital Collections Repository (Dspace) may contain direct links to repository datasets. Datasets within the data repository may reference research papers published through Open Access Academic Journal Software (OJS3). These may also directly reference associated datasets through citation and reference links that the Open Journal Software allows. These allowances enable researchers reviewing a scientific paper to click directly to the data to scrutinize research integrity and possible download.

Further internal linkages are possible through the Electronic Theses and Dissertation Management System (Vireo) and externally, the Researcher Identity Management System (ORCID). ORCID may be used as a network hub for further linkages to funding agencies, other research identifiers and institutions.

Figure 5. ORCID https://orcid.org/

Data Research Repository Ecosystems: Global Possibilities

Currently, the Texas State Data Research Repository and ecosystem is part of a statewide 22 university library consortium. Participating members who possess a local instance of the Dataverse repository may search statewide for similar datasets. A researcher working on datacentric subject areas can connect with associated network ecosystems

Figure 6. Texas Digital Library University Members https://www.tdl.org/

The possibilities with the Dataverse model for larger geographically disbursed consortia on state, national or international levels, become interesting as data-centered networked research ecosystems scale. Surrounding tertiary software can be connected on consortial, state, national or international levels.

There are currently approximately 266-300 High Research Activity institutions (universities and national research centers) in the United States and Canada. The Times Higher Education Supplement estimates there are 1250 higher level research universities worldwide (2.7 - 4.2% of all universities worldwide). QS World University Rankings estimate that 40% of these higher-level research institutions are in Europe, 26.5% in the Asia/Pacific Corridor, 18% in North America, 9% in Latin America and 6.5% in Africa and the Middle East.

Future Possibilities for Data and Digital Scholarly Ecosystems: Global Data Sharing

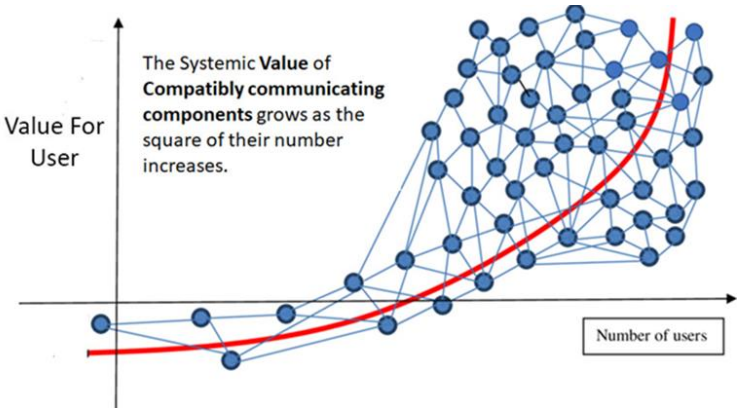


Figure 7. Metcalfe's Law for Networked Components

With the abundance of mature network-enabled open-source online research software possibilities outlined, there is little reason that the other top 2-3% of research libraries globally should not be empowered with these types of digital ecosystems at the least. The larger majority of the top 2% of these research libraries globally (approximately 1000) currently do not yet possess these newer software infrastructures. It would be easy enough with the proper organizational funding to open possibilities for the entire 1250 research level institutions globally.

As a global grand challenge, why not enable these research institutions and their academic libraries with these types of online networked research ecosystems? This would open abilities for dialogue, sharing research data and collaborating with established networks of contact. One need not look far for an immediate precedence for this type of global need beyond the continuing pandemic and needs for more organized future global data-sharing networks.

Future Global Possibilities One Server Per Research University: 2022-2027

With interest and funding it would be beneficial and easily possible to give at least 1000 Research University Libraries globally one configured online research ecosystem server with the six pieces of open-source research software described. Three times a year, weeklong training could be given over five continents and an online help network set up through IFLA as a worthy new millennia initiative. Later, assessment would be done to gauge the network effects of sharing data and research globally.

Physical servers could be given to research institutions with software installed, or fractional server space placed in the cloud over five continents, with mirror sites globally. With current server power, this would not be more than $1000.00 USD/server. One or two large funding agencies, or suitable international organizations, could shepherd the project with IFLA to develop version 1.0 of this Global Digital Scholarly Ecosystem and Research Data-Sharing Network. After one, three- and five-year periods assessment could be done. Subsequent versions could also be produced improving and streamlining the model and utilizing the latest agile information technology project management techniques. The positive effects for research, collaboration, data collection, and science would be enormous. The benefits for the progress of knowledge through research libraries in the 21st century would also be large if even a fraction of the global potential was achieved.

Figure 8. One Server Per Institution
Global Digital Research Ecosystem Initiative

Conclusions

This paper has given a high-level outline of a successful, pragmatically possible, digital ecosystem on local levels (Texas State University Libraries Model) and extended this model towards future global research possibilities. These systems are not overly challenging to set up at local institutions. Example links, papers, presentations, and open-source software for download from the working digital ecosystem links are given in the references below.

Possibilities with current open-source software for data research repositories and digital scholarly research ecosystems are wide open. These systems are pragmatically realizable on local university library levels. They should begin to be prototyped with larger national, international, and global networks. The time has come to begin connecting these types of digital research ecosystems. Present day research networks, libraries and universities are well-suited towards these new millennia digital stewardship infrastructures and roles.

Further References

Papers

Uzwyshyn, R. 2020 Developing an Open Source Digital Scholarship Ecosystem International Conference on Education and Information Technology ICEIT2020. St. Anne's Oxford, United Kingdom. February 2020.
- - -. Open Digital Research Ecosystems: How to Build Them and Why . *Computers in Libraries*, (40) 8. November 2020.
---.Online Research Data Repositories: The What, When, Why and How. Computers in Libraries. 36:3, April 2016. pp. 18-21.

Presentations

Uzwyshyn, R. Digital Research Ecosystems for Open Science (Presentation). AI for Data Discovery, Reuse & Open Science Symposium. Carnegie Mellon University. October 20, 2020.
---.Developing a Digital Scholarship Research Ecosystem. (PDF) Association of Southeastern Research Libraries Members' Meeting Presentation, April 29, 2020.

Open-Source Digital Ecosystem Software (Downloads & Info.)

Data Research Repository: Dataverse https://dataverse.org/
Digital Collections Repository: Dspace https://duraspace.org/dspace/
Content Management System: Omeka https://omeka.org/
Academic Journal Software: OJS3 https://pkp.sfu.ca/ojs/
Identity Management Software: ORCID https://orcid.org/
ETD Management Software: Vireo https://www.tdl.org/etds/

Texas State University Libraries Data Repository & Digital Ecosystem

Texas State Data Research Repository
Texas State University Libraries Digital Scholarship Ecosystem.
Texas State Digital Collections Repository
Texas State Online Research Identity Management System (ORCID)
Open Journal Systems @ Texas State

You May Like – How the National Library Board of Singapore uses Machine Learning to Recommend Books

Tan Wen Sze, Assistant Director, tan_wen_sze@nlb.gov.sg

Lim Chee Kiam, Principal Solutions Architect

Chow Ying Yi, Senior Project Manager, chow_ying_yi@nlb.gov.sg
National Library Board, Singapore

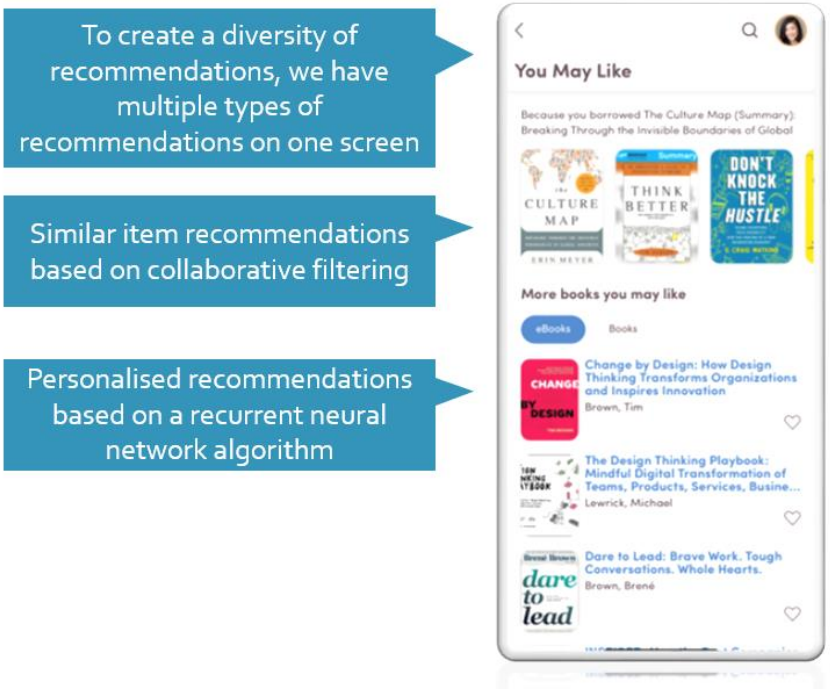


Figure 1: Personalized recommendations on the NLB Mobile app

The National Library Board of Singapore (NLB) has been developing recommendation engines for our collection of books and articles and has deployed them on our digital touchpoints (e.g., website) since 2009. More recently, in 2019, we began testing and implementing a new recommendation engine for our book and ebook collections using Amazon Personalize, a cloud-based machine learning personalized recommendation service.

Since 2020, we deployed our first version of this service on our website and the NLB Mobile app. We are now working to release new and enhanced recommendations based on the same underlying service.

By using a fully managed cloud-based service both infrastructure resources and the entire machine learning pipeline are taken care of by the service provider, Amazon Web Services. There are no minimum fees and no upfront commitments. The service is also continuously enhanced and better algorithms and features are made available without additional investments except for the work required to leverage them

However, there remain many tasks for us to deliver a complete recommendation service. These tasks include: choosing the right features in our datasets, setting up different types of recommendations for good performance as well as diversity, and designing how our patrons can interact with the service.

Choosing the right data

The National Library Board of Singapore already has a process to bring in transactional data into our data warehouse. Having this infrastructure in place reduces our effort to generate datasets with new features. For example, we created a new feature that classifies borrowing patterns – whether a patron borrowed mostly books for children, books for adults, or a mix. This was important because many parents borrowed books mainly for their children, and this would affect the recommendation outcome for other adults.

Using the data warehouse, we also created a new dataset based on a unified work-level identification of books and ebooks. This allows us to have a unified view of the patron, and reduced recommending duplicate titles (e.g., the same title in different formats appearing in the same set of recommendations).

With work-level identification in place, we further explored subject-based recommendations. By harmonizing the metadata for books and ebooks under a common work-level identification, we created new subject-based recommendations to provide the patron with even more personalized choices for various subject areas.

These two new recommendations are works-in-progress as of 2021 and will be deployed into our digital touchpoints in 2022. The work will not stop here as there are many more data features that we can use to increase and enhance the types of recommendations.

Using our own web services to enhance the recommendations

Instead of using the recommendations from Amazon Personalize directly we chose to develop our own web service layer, and have our front-end applications retrieve the recommendations from our web services. This lets us customize our recommendations further, layering additional logic to provide better experiences.
For example, new and inactive patrons would not have recommendations available. Instead of not having any recommendations for these patrons, we have our digital touchpoints additionally pass the patron's age when they try to retrieve recommendations. This allows us to retrieve generalized age-appropriate personalized recommendations for that patron's age group.
Another possibility is to use web services to mix and match of different recommendations from Amazon Personalize. We can liven up our current list of recommendations with titles from a more "exploratory" set of recommendations. This lets us introduce some diversity while maintaining a high degree of familiarity. Familiarity is important for the user to trust the system and keep coming back.

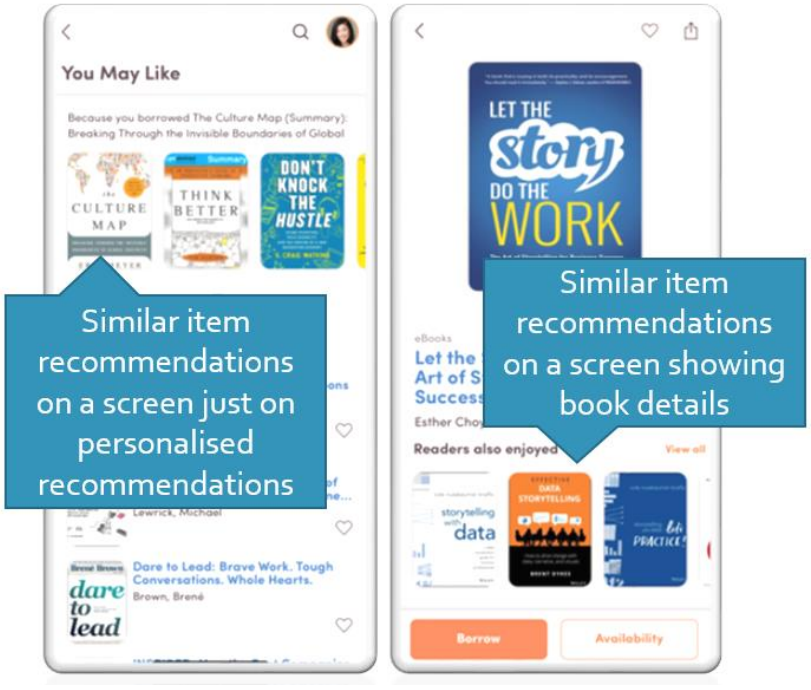


Figure 2. Deploying recommendations in multiple screens

Designing the interaction on the digital touchpoints

Providing too many choices can be overwhelming for patrons. While we have a webpage / app screen that focuses on recommendations, we are selective about the total number of recommendations we put in and where we choose to surface them. These considerations will become more important when our new recommendations are ready in 2022.

We need to design the user journey and how our patrons interact with our services. There are different touchpoints in a user journey where our patrons may be interested in a new book, e.g., checking his or her loan records, searching for titles, and looking at title details. These touchpoints are where we can place in similar item recommendations to enhance the user experience.

Conclusion

Personalized recommendations are a good way of introducing new books to our patrons. With fully managed machine learning recommendations services available, personalized recommendation services are now accessible to libraries.

At the same time, with more ebooks available, we can now provide a seamless user journey from piquing the patron's interest to getting and reading the recommended ebooks. These are empowering times for libraries to engage their patrons more deeply.

Adapting to Challenges of Big Data Collections in Libraries: An Archives Unleashed & Web Archives Use Case

Samantha Fritz, MLIS, svefritz@uwaterloo.ca
Project Manager, Archives Unleashed Project, Canada

Libraries have adapted to the field of big data and analytics, yet the realities of complex and challenging data affect our memory institutions and the relationship with end-users like scholars and researchers.

Libraries use big data in various ways, from developing collections to tracking usage trends. They also act as depositories/repositories for big data - from digitized archival collections to geospatial data to accessing various government census or statistical datasets.

Increasingly, web archives have been adopted as part of collection holdings within academic and national libraries alongside other heritage-based institutions.

By exploring the big data nature of web archives and the Archives Unleashed project as a case study, we can identify the impact that tool building, community engagement, and collaborative partnerships have on expanding visibility, access, and use of web archives.

BIG DATA

Popularized in 2005 (1), the term Big Data, reflects the ways technological developments of the 20th and 21st centuries have profoundly changed the way we produce, interact, and preserve information. It doesn't take much to realize our daily habits and interactions keep us plugged in some way or another. As a result, the ways in which we use technology both professionally and personally have contributed to the exponential growth of our digital data footprint.

While a large corpus exists around the study of Big Data, we can begin with the 3Vs Framework outlined by Doug Laney in 2001 to identify Volume, Velocity, and Variety as key characteristics of Big Data (2).
Although there is no single definition, we can understand Big Data as "data whose scale, diversity and complexity require new architectures, techniques, algorithms, and analytics to manage it and extract value and hidden meaning." (3)

WEB ARCHIVING + LIBRARIES

With this general definition, we can consider web archives to exemplify the concept of big data while recognizing the considerable challenges this type of data presents to the cultural heritage sector.

Web archiving is the process of preserving vulnerable born-digital information in an ISO standard file format known as a W/ARC. Since 1996, various organizations and institutions worldwide have participated in archiving web data, from national libraries to academic institutions to international organizations.

The prolific preservation of web data has altered our research landscape for scholars in many disciplines. As historian Roy Rosenzweig describes, web data shifts our scale from resource scarcity to abundance. The use of digital-born data is critical to our understanding of the recent past. As digital historian Ian Milligan suggests, the absence of web archives from our historical record would result in significant knowledge gaps when exploring topics post-1990s. (4)

For over two decades, memory institutions and organizations worldwide have engaged in web archiving to ensure the preservation of born-digital content that is vital to our understanding of post-1990s research topics. Web archiving programs are increasingly adopted as part of institutional activities and agendas. In general, there is a recognition from librarians, archivists, scholars, and others that web archives are critical resources and are vulnerable to stewarding our cultural heritage. (5)

As intellectual property expert Ben White expresses, "libraries have long been in the business of preserving documentary heritage," and we continue to see the fundamental processes of selection, organization, description, and access apply to big data collections such as web archives. (6)

CHALLENGES WEB ARCHIVES PRESENT

Yet, despite the volume of data captured over two and a half decades, web archives have not become a dominant resource for researchers. The adoption has been relatively slow. We can look at a few different reasons for this:

● Lag in tool development. The tool landscape for working with web archival data at scale is sparse. While collection and preservation efforts have excelled and standardized practices have emerged, the development of analytics tools and infrastructure has lagged.

● Available tools require technical competencies. Working with any type of data at scale is messy and difficult. Web archives are no different. This is especially apparent when conducting large-scale analysis, which commonly requires an advanced degree of technical knowledge. Unfortunately, computational expertise and familiarity with the command line tend to be out of reach for most scholars, who already contend with time, support, and resource limitations.

●Lack of visibility. There is a silo effect when observing the collecting practices of curating web archiving collections. Curation happens at the institutional level, which does not lend well to the search and discovery of collections that span across a broad range of subject matter and institutional/organizational bodies across the globe.

This is all to say that the scale of web archives present high barriers to access, use, and ultimately scholarly exploration.

OPPORTUNITIES: WORKING WITH WEB ARCHIVES

To address these challenges, we can look at three methods the Archives Unleashed Project has engaged in, to lower barriers of working with web archival data at scale.

1.Make scalable, user-friendly tools.

The Archives Unleashed Project (2017-2020) developed two open-source, transparent, and user-friendly tools, namely the Toolkit and Cloud, to conduct scalable analysis of W/ARC files and extract scholarly derivatives for research. This directive was informed through community consultation and an environmental scan, which identified a lack of available tools for researchers, primarily within the digital humanities field. The Toolkit and Cloud offer different ways of interacting with W/ARC files. The Toolkit allows users to work with raw files via command line. The Cloud provides a simple user interface to generate a variety of derivatives quickly and easily. Both tools have adopted practices and methodology widely accepted among digital humanities scholars.

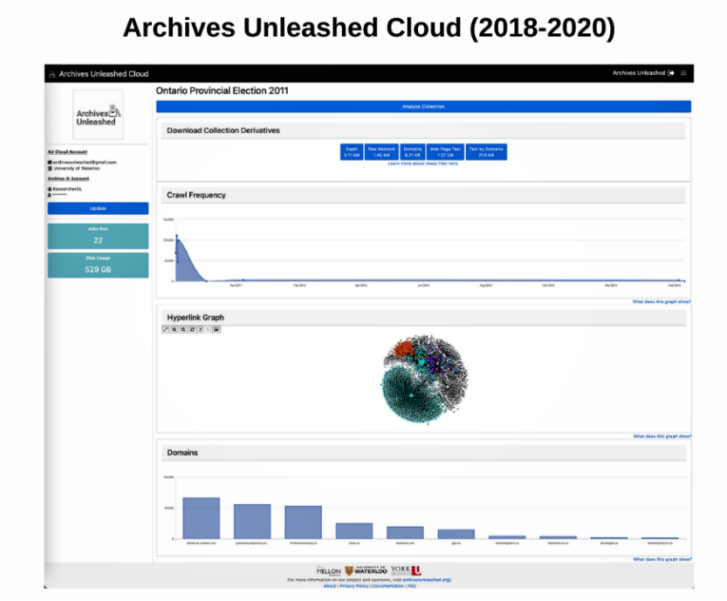


Figure 1: The Archives Unleashed Cloud Interface

2.Create learning resources to inspire confidence & use

Several resources were created to address the more technical components of the Archives Unleashed Toolkit and the complexities of working with data at scale.

In addition, these materials were designed to inspire confidence in working with the WARC file format, encourage scholarly exploration of web archives, and support a wide range of skills and experience.

First, the Toolkit User Documentation (https://aut.docs.archivesunleashed.org) presents a cookbook approach, wherein users can copy and adapt pre-built scripts (or recipes) to address common analytic tasks. In providing scripts, the documentation offers building blocks. Researchers can begin to explore the types of questions that can be asked while also presenting examples of the types of information that can be extracted.

Second, several written and video tutorials were designed to provide step-by-step instructions to investigate scholarly derivatives in conjunction with external tools (e.g., Gephi, Voyant, AntConc) and methods (textual, sentiment, network analysis). In both cases, these learning resources guide researchers along a journey of exploration and highlight the possibilities and use cases of working with web archive collections.

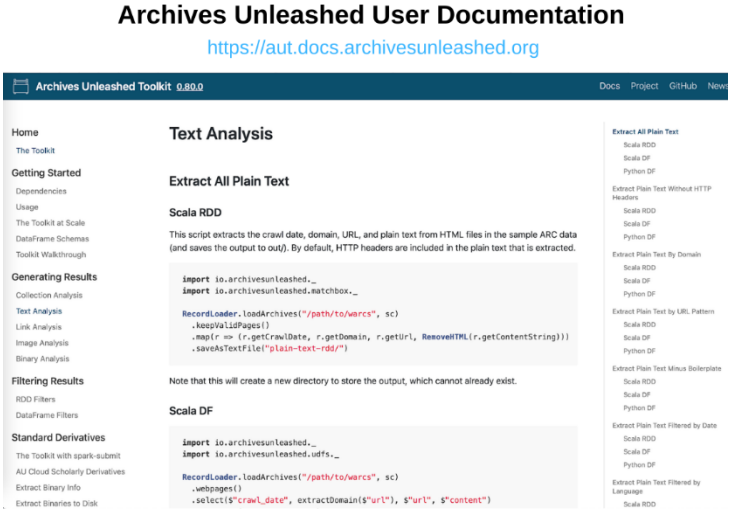


Figure 2: Archives Unleashed User Documentation

3.Build and Engage with Community

Building and engaging with community has been an essential pillar for Archives Unleashed because projects (and in our case, tools) can't live in silos.

The main community-building activity undertaken by the project was to run a series of datathons, an adaptation of the hackathon model. These events engaged interdisciplinary groups of people in a hands-on learning environment to work directly with web archive data and new analytical tools to produce creative and ingenious projects that explore W/ARCs.

Datathons also helped build a user community around our tools and foster a sense of belonging. One of the most notable impacts of our events has been that datathon alumni have become ambassadors among peers, colleagues, and broader communities, such as the digital humanities, GLAM (Galleries, Libraries, Archives, and Museum) institutions, and the web archiving community. (7)

COLLABORATION EXPANDS VISIBILITY, ACCESS AND USE

Archives Unleashed has proactively developed collaborative relationships with various stakeholder groups to expand access and use of web archival data and increase the visibility of web archive collections. Through our experience with researchers, we have recognized that although collections may exist, if researchers aren't aware of where or how to use the data, these carefully and thoughtfully curated collections will have limited adoption in research workflows.

The project team has connected with research communities by collaborating with scholars in a variety of disciplines (e.g., digital humanities, social sciences, and journalism). This has proved an invaluable opportunity to identify and highlight the applications of web archival research.

We have also collaborated with several University libraries in North America to expand the accessibility and visibility of their web archive collections. These collections were processed through the Cloud to generate scholarly derivatives and made openly available through repositories like Zenodo. (8) This approach has been valuable in highlighting institutional collections, lowering barriers researchers face in gaining direct access to web archive datasets, and demonstrating examples for the possibility of web archives for scientific inquiry.

The project has recently engaged in a formalized collaboration with the Internet Archive's Archive-It, which will run until 2023 to integrate and blend complimentary services. The result will be an end-to-end process for collecting and studying web archives. (9)

ADOPTING PERSONAS TO MEET BIG DATA CHALLENGES

The Archives Unleashed Project highlights methods for tackling challenges of investigating web archive collections, spanning from tool design to documentation and learning resources to community building.

Let's return to a broader discussion around the challenges that big data as collections present to librarians, technologists, and users. Perhaps we can find solutions to overcoming the challenges of data visibility, use, and access, simply by reflecting on and adapting personas that seem to naturally develop within the LIS profession.

First, the persona of community stewards. Libraries foster, support, and engage many overlapping communities. As such, they are well-positioned to create a conversational space to identify user needs.

By fostering community and a sense of belonging, libraries also promote the transformation of users into champions and advocates of services and holdings, positively affecting the visibility of data collections.

This is closely connected to the role LIS professionals play as collaborators and teachers. By understanding the community served, collaborative relationships can be built to encourage and instill confidence in the use of data. This is already present in the ways libraries support data literacy and skills development through initiatives like learning guides, tutorials, and workshops.

Finally, by embracing the identity of a resourceful problem solver, LIS professionals can meet data access challenges. This means assessing the landscape of available tools and methods, identifying reliable resources that can be shared with patrons, and, when necessary, developing in-house workflows or systems to help fill gaps.

Drawing from these personas and roles, libraries and LIS professionals have an incredible opportunity to craft solutions that will ultimately foster deep and meaningful connections between data collectors and data users. By adapting to the challenges of big data collections, we can make big data collections more visible, accessible, and usable.

Endnotes

[1]A Short History of Big Data: Big Data Framework. Enterprise Big Data Framework©. (2019, March 26). https://www.bigdataframework.org/short-history-of-big-data/ .

[2] Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research

[3] Noam Slonim et al., "Knowledge-Analytics Synergy in Clinical Decision Support," *Studies in Health Technology and Informatics* 180 (2012): 703–07. From https://crl.acrl.org/index.php/crl/article/view/24918/32769

[4] Milligan, I. (2019). *History in the age of abundance? How the web is transforming historical research.* McGill-Queen's University Press.

[5] White, B. "Guaranteeing Access to Knowledge: The Role of Libraries". WIPO Magazine, April 2012. From https://www.wipo.int/wipo_magazine/en/2012/04/article_0004.html

[6] Canadian Association of Research Libraries (2014). "Archiving the Web." Working paper submitted to the CARL Committee on Research Dissemination. From https://www.carl-abrc.ca/doc/Archiving_the_web.pdf

[7] Fritz, S., Milligan, I., Ruest, N., and Lin, J. (2021). "Fostering Community Engagement through Datathon Events: The Archives Unleashed Experience". *Digital Humanities Quarterly*, 15:1.

[8] For a full list of collaborative datasets see, https://archivesunleashed.org/publications/#datasets

[9] Petrik, J. (2020). Archives Unleashed Project scales up with Archive-It for better collection and analysis of digital history. University of Waterloo, Arts News. Press Release, July 2020. From https://uwaterloo.ca/arts/news/archives-unleashed-project-scales-archive-it-better

Diversifying Collaboration: Network Complexity and Big Data

Ingrid Mason, ingrid.mason@anu.edu.au
Project Manager, Australian Data Archive, The Australian National University

A/Prof Steven McEachern, steven.mcEachern@anu.edu.au
Director, Australian Data Archive, The Australian National University

Introduction

Library practice is founded upon the ability to have an impact, at the personal, community and societal levels. Library professionals understand that changing collections and practices can be the means to assist in bettering lives and having a positive social impact. For the library world to have impact and change lives, it also requires that libraries change their collection management practices and with whom they collaborate.

"…great works of culture were almost invariably created to redeem, console and save the souls of their audiences. They were made with the idea of changing lives" (The School of Life, 2018, p.11)

The library world benefits from strong and well-established collaboration approaches and structures to provide access to digital collections in and across national borders. Existing library collaboration structures however are bounded by in-group definitions, requirements, expertise, leaders, and the small technical scales of collection management practices of the past. New forms of collaboration can directly be linked to striking examples of leadership outside of the norm. This is evident in the reflections of library practitioners navigating the global pandemic in the "New Model Library".

"Having staff collaborate across their typical boundaries proved effective at meeting community needs regardless of staffing changes and fluctuations in demands for services." (Connaway et al., 2021, p.12)

New drivers for collaboration

Given the network complexity that libraries now operate in, new drivers for collaboration and models of partnership around big data such as data collaboratives, collectives (GovLab, n.d.) and community led initiatives (Thorpe, 2021 p.347) need to be established. Key questions for the library world are:

- How can the global library community scale up its efforts to unlock social value in big data?
- Who can libraries collaborate with to innovate and scale effectively and ethically?
- What situation are libraries in socially and culturally, and how do these need to change?
- How do library underpinning infrastructures support a move to build community trust in their capacity to manage big data?

Big data management is resource intensive and presents significant complexities in terms of collection, access, and governance for libraries. Collaboration in the library world needs to change to continue to have social impact and to be resourceful. For the library world to radically collaborate, work and relate differently, this means centering different priorities, expertise, and voices around big data management.

"…it isn't just us we learn about through culture. It is also the minds of strangers, especially those we would not ordinarily have ever learnt much about." (Ibid, p. 78)

Some practical assumptions also need to be tested:

- Large datasets will be shared to improve training and testing machine learning models e.g., digitized newspapers.
- Data too big to move will require compute to be co-located and high-capacity (low latency) networking e.g., web archives.
- Existing artificial intelligence technologies and techniques need critical evaluation because of bias in big data e.g., prediction and optimization.

Big data curatorial practices

The social change and ethical complexities of big data governance is critical as a part of the development of curatorial practice. At present the technical scale and complexities of big data management dominates practitioner discourse in technology, analytics, data and computer science – but not big data curatorial concerns. Library collaboration around big data management is needed so that social responsibilities (i.e., social equity and plurality of view in intellectual access practices) can be embedded into curatorial drivers.

But what does this mean in real terms? The library world needs a new socio-technical vision – to show leadership in social and technological change – and to weigh in on discourse on big data curatorial matters and lay the foundations of "curatorial analytics" (Kenderdine, Mason & Hibberd, 2021). For rather too long the social and technological aspects of library practice have been given "front and back of house" or "library and IT" organizational treatment and this needs to change.

Network complexity

Loosely articulated network complexity in combination with big data involves libraries working within cloud architectures and operating in specific contexts with old and new relationships to consider.

New risk management models and trust building methods need to be established for big data management. Leadership in social and technical brokerage will be a critical feature of working in new collaborative arrangements around big data management.
Cloud architectures as underpinning infrastructures are terrain loaded with commercial and sovereign interests. The variety of cloud arrangements needed for big data management i.e., local/private services, consortia/private services, hybrid private/public services and local or remote vendors, pose a socio-technical challenge. How best to store big data, where, and how to back it up?

So much of library technical work associated with data storage and movement is invisible, and yet with big or sensitive data the necessity for rigorous management and practice needs strong policy and practice guardrails. There is a difference in the practical and ethical concerns with providing public access to big digitized historical newspaper collections in comparison with big web archive collections regarding personal privacy and social consequence. Where, how, and what big data is kept, moved, and computationally processed becomes significant. This collection management and collections infrastructure work needs to be undertaken out in the open.

What if we want to bring big data collections together across cultural boundaries and national jurisdictions, or relocate them? The library world needs to develop ethical and practical underpinnings in curatorial methodologies for big data management that are well documented and where data science practice is integrated into library practice. Library work in big data management needs to stand out in sharp contrast to notorious big data management practices of technology companies and computer science, where material is coarsely amassed and indiscriminately lumped together for computational convenience.

More importantly libraries with big data collections that represent the history of colonization and Indigenous peoples as part of their collection policy and curatorial practice, need to recognize the social structural position played in the world of cultural knowledge regarding personal, cultural, and social categorization and marginalization. This is the moment for libraries to own up to the epistemological and tokenistic practices of the past (Thorpe, 2021) and shift away from didacticism and a centralized position of authority control. A new humanistic and user-centric position is needed and to undertake community-led initiatives around big data (Hicks, Nicholson & Seale, 2021). It is important for libraries to understand their situation i.e., where they are connected into networks actually and historically, nationally and internationally. There will be different social motivations and levers to radically change the way collaboration occurs and different community interests that play out in library big data management.

The CADRE platform project

The CADRE platform project is operationalizing the Five Safes framework (Ritchie, 2017) to improve researcher access to sensitive data. The Five Safes framework is a conceptual model – a means to examine the safety of data, people, projects, settings, and outputs and all the intersections and alignments of each of those “safes”. The new curatorial approaches to data management, informatics and analytics are emerging through a collaboration across diverse organizations and sectors: universities, government agencies and national research infrastructure providers. The CADRE Five Safes framework and information model (McEachern et al, 2021) emerging from the project draws heavily on social science research, data and computer science practice, expertise. Quantitative and qualitative research methods as well as data and custodial responsibilities and knowledge of secure technologies for sensitive data develop new curatorial approaches to enabling collection access. The project partners and affiliates share an interest in addressing ethical concerns, optimizing workflows, and services, and improving access to sensitive data through innovation and digital transformation.

The technical work in CADRE project bears all the features of digital library development i.e., governance, common information standards, interoperability, authorization and authentication technologies, new curatorial terminology, and risk assessment and decision-support systems. The context for this collection infrastructure work is Australian. There is legislative change around access to data held by the federal government as the Data Availability and Transparency bill (ONDC, n.d.) is moving through parliament.

Figure 1. Australian Data Archive

The awareness of privacy challenges and interest in accessing data about people – whether big or small and often sensitive or in need of scrutiny for bias – is growing and it is global.

The network of shared interests in ethical practice and use of new technologies lies at the center of this collaboration in the CADRE project, and it is a step change for all the project partners to work this way.

Library and information sciences draw heavily upon social science in theory and in practice. Substitute "big data" for "sensitive data" or overlay the two in this scenario with the CADRE project and there are some useful lessons for libraries. The Five Safes framework (see Figure 1.) has significant merit as a conceptual lens if the presumption is made that all big data in library collections is likely to be sensitive or there will be issues with bias.

Figure 2: The Five Safes framework as conceptual lens

The technical requirements associated with digital library development are recognizable in the CADRE project. The non-technical requirements however are highly specialized, complex and expensive to develop. The need for new user services, access and to share costs, has drawn collaborators together across organizational and sector lines to meet this challenge.

The social and financial costs were high with leaving siloed and duplicated infrastructure in place. By tackling digital transformation of access management services with special technical requirements (e.g., sensitive and/or big) the gains made by joining forces to establish shared services are higher. Working collectively is nothing new to the world of libraries with a history of metadata sharing and interlibrary loan services. The collective action was however predicated on easy and extensive access to library tools and services. Network complexity and big and sensitive data are here. Librarians must face the associated practical and ethical dilemmas of big data openly to thence be able to unlock social value.

Curatorial Analytics

"A computational archive is a refactoring not just of digital archives but also of curatorial practices of authority, interpretation and the philosophical and positional shifts required in digital transformation.
Curatorial analytics will build on cultural analytics (as material deconstruction of the archive through computation and generalization) [24] by maintaining a focus on semantic refinements, variation, contextualization and significance." (Kenderdine, Mason & Hibberd, 2021, p.5)

New curatorial approaches and data documentation processes i.e., informatics and analytics developed by the library world around big data, need to be considered as part of the computational turn. For the metadata nerds, new informatics and curatorial analytics need to be developed at the core of big data management as part of library collecting practice. How the library community tackles this socio-technical challenge with big data will be determined through repositioning, refactoring practice and new partnerships that work across diverse organizations and sectors. Few libraries can afford to go at this big data management exercise alone as collections reflect users and communities. The library world will benefit from the emergence of new leaders, radical collaborations around big data, and growth from new relationships with community and other professional groups.

References

CADRE Platform Project (n.d.) https://cadre5safes.org.au
Connaway, L. S., Faniel, I. M., Brannon, B., Cantrell, J., Cyr, C., Doyle, B., Gallagher, P., Lang, K., Lavoie, B., Mason, J. & van der Werf, T. (2021) New Model Library: Pandemic Effects and Library Directions. OCLC, October 2021.
GovLab (n.d.) https://datacollaboratives.org/
Hicks, A., Nicholson, K. P. & Seale, M. (2021) Towards a critical turn in library UX, College & Research Libraries, 83(1).
Kenderdine, S., Mason, I. & Hibberd, L. (2021) Computational archives for experimental museology, International Conference on Emerging Technologies and the Digital Transformation of Museums and Heritage Sites, 2021/6/2, Springer, Cham, 3-18.
McEachern, S. et al. (2021) In press, The CADRE Five Safes Framework.
Office of the National Data Commissioner (ONDC) (n.d.). Data Availability and Transparency Bill, https://www.datacommissioner.gov.au/data-legislation/data-availability-and-transparency-bill
Ritchie, F. (2017) The 'Five Safes': a framework for planning, designing and evaluating data access solutions, Data for Policy 2017 Conference DOI: 10.5281/zenodo.897821
The School of Life (2018). What is Culture For? London: The School of Life.
Thorpe, K. (2021) The dangers of libraries and archives for Indigenous Australian workers: Investigating the question of Indigenous cultural safety, IFLA Journal 47(3), 341-350.

Acknowledgement

This paper is based on a presentation given in the Big Data Special Interest Group session at the 2021 International Federation of Library Associations and Institutions (IFLA) World Library and Information Congress (WLIC).

Digital Transformation for All at the National Library of France (BNF): Mapping the BNF's Digital Realities

Céline Leclaire, celine.leclaire@bnf.fr
Strategic content production officer
National Library of France

Figure 1. National Library of France Digital Roadmap

Where am I? History of the BnF's Digital Roadmap and associated challenges

Created in 2008 and continuously developed since to follow the way digital technologies and projects constantly evolve at the BnF, this Digital Roadmap intends to help people (staff) gather information and find their way in a more and more complex digital ecosystem.
The 2020 Roadmap is the result of a ground-breaking collective approach that began in 2019 and involved about 150 people. It was designed from the start to offer a more familiar media to staff, and to enrich its content by collecting different views of a same product or project.
This collaborative work brought out how our professional life is characterized by the notion of networks: many links can be created between the different elements of the map, and its boundaries are open, suggesting that the library interacts with a larger ecosystem. This work also drew attention to the increasing importance of issues related to ethics and to the working environment at the library, for both staff and users – something we had in fact already envisaged before the 2020 lockdown.

These shifts led us to create an edition quite different from the previous ones (see the Digital Roadmap 2016): not only the medium chosen in 2020 is original, but it supports an entire system, that includes presentations, meetings, workshops, etc.

Making diverse and moving components understandable

To help users understand digital issues as a whole and in detail, we first put human beings at the core of the representation: digital technologies are closely linked to jobs, to acting people, in every field. You can see many characters on the map and find people to contact on the fact sheets. Abstract spaces come to be inhabited.
Besides, the 2020 cartography draws its inspiration from the library's collections of maps and plans: it reminds colleagues of familiar contents… and makes them dream!
That's the second point: using metaphor and drawings to invite users to explore unknown countries and give free rein to their curiosity. The map and fact sheets can be printed: we also wanted to offer something that could be held in hands.

Figure 2. Digital Roadmap (Detail)

Finally, the document has been given a clear structure, with highlights: ten red discs stand out on the map. They represent the key elements of the library's digital life: Digital legal deposit, Digital collections, Digitization, Data, Working environment, Digital hospitality, Text and data mining tools, Gallica, General catalogue, and Metadata management. The map is also divided into nine regions (Digital collections; Data; Infrastructures; Collections processing; Avenues and perspectives; Technologies; Exploring and sharing resources (I and II); Metadata management and catalogues). The set of cards is organized into six card categories. Among them, five strategic priorities can be distinguished (Managing the evolution of work; Studying, examining and anticipating uses; Sharing infrastructures, know-how and skills; Managing information technology applications; Overcoming the digital divide), as well as a five-fold vision for the future, closely linked to the founding missions of the BnF and to its obligations as a major public actor, particularly in the fields of data processing, information dissemination, privacy, and sustainable development.

(Un)expected assets
The BnF Digital Roadmap 2020: a resourceful support for internal training and dialogue

By combining different scales (small scale, larger one…) and adding a light touch in the presentation of the different ways heritage is digitally collected, preserved, catalogued, and disseminated, this Roadmap makes it easier for librarians (and staff in general) to speak to each other and to manage the evolution of their work. It has proven to be particularly useful when welcoming new staff and partners.

The workshops we organize mainly aim at giving participants time to discover the content of the document, and at gathering different people to bring together different views of digital technologies. During these workshops, professionals can talk together, ask questions, and teach each other what they know about digital issues and projects. It also gives us an opportunity to identify new themes to develop or explain in relation to this field (such as digital responsibility, sustainability, etc.).

Today, a dozen professionals can facilitate such workshops based upon the Roadmap.

An inspiring project

Other institutions (museum, university…) inquired about the way we managed this project, and within the BnF it appears as a step towards better connections and understanding between the different jobs, fostering collective day-to-day work in this complex ecosystem, and linking this day-to-day work to great strategic priorities.

A dynamic process

Some clouds are visible on the map because it has been designed to be enriched and updated. Elements can be added on the map, new fact sheets can be created…
As said before, development opportunities have already been identified: issues related to digital responsibility, sustainability, but also to public service, digital training, content dissemination…
To answer some requests, we have launched several itineraries such as "Must-see spots on the digital roadmap" and "The fundamentals of AI" (a travel companion to the BnF's Digital Roadmap on the footsteps of Artificial Intelligence).
Finally, to complete this overview, we shall present two recent prominent projects: the Data Lab and the BnF's AI Roadmap.
Over the past ten years, the huge increase of digital collections at the National Library of France (BnF) - Gallica, Internet archives, metadata, etc. – has opened new research perspectives. To meet new researchers' uses and needs, the BnF has set up a service: the BnF Data Lab, aimed to support research teams in constituting and processing digital corpuses. This space, designed for individual and collaborative work, is available in the BnF's premises since fall 2021 and complemented by a website, api.bnf.fr.
In 2021, the BnF also shaped a roadmap regarding AI issues and projects. AI technologies can be used in different fields at the Library: to create and enrich catalogue records, to manage collections, to search content and improve access to content, to put content in perspective, to make decisions... Five priorities have been defined for the 2022-2026 period: making IA issues and projects a part of the Institution's global strategy, improving the R&D's organization and implementation at the BnF, developing skills, adapting infrastructure and data management, and designing a multi-year program with other stakeholders What an exciting never-ending story!

Further Links
English version:
https://www.bnf.fr/en/digital-roadmap
Spanish version:
https://www.bnf.fr/es/plan-desarollo-digital
French version:
https://www.bnf.fr/fr/schema-numerique-bnf
Contact: schema.numerique@bnf.fr

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 applicationsof 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 andworkshops. 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 andresources regarding activities of the Section, session minutes, publications, and membership details.

The IFLA-IT email list provides a forum for members to exchange ideasand 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

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

Big Data SIG

Convenor
- Cory Lampert, University of Nevada Las Vegas, UnitedStates. cory.lampert@unlv.edu

Information Coordinator
- Patrick Cher, Singapore National Library. patrickcher@gmail.com