Purdue University

Purdue e-Pubs

Proceedings of the IATUL Conferences

2019 IATUL Proceedings

Swimming In The Deep End: Curtin Library's Deep Dive Into Data...

Drew Fordham
Curtin University, Australia, d.fordham@curtin.edu.au

David Lewis
Curtin University, Australia, d.lewis@curtin.edu.au

Amanda Bellenger
Curtin University, Australia, a.bellenger@curtin.edu.au

SWIMMING IN THE DEEP END: CURTIN LIBRARY'S DEEP DIVE INTO DATA…

Drew Fordham

Curtin University, Australia

d.fordham@curtin.edu.au

David Lewis

Curtin University, Australia

d.lewis@curtin.edu.au

Amanda Bellenger

Curtin University, Australia

a.bellenger@curtin.edu.au

Abstract

In 2017 Curtin University Library staff commenced an initiative to explore library data through analysis and visualisation tools. A Working Group was formed of self-selected staff from across the library who had a common interest in opportunities to use data and analytics tools. A key focus was to seek innovative ways to use data and have a measureable impact on library operations and future planning.

An environmental scan of the data landscape in academic libraries suggests five core areas of activity – the use of library data to gain insights, the need to upskill library staff in data methods, the demand for data literacy programs for library clients, the value of building partnerships and collaborations within the institution to support data programs, and the increasing use of library facilities to provide data-rich learning spaces. The group's work fell across all of these areas.

This paper will describe the development of the group and its practical outcomes including the innovative use of data to inform a multi-year building project at Curtin University's Robertson Library in Western Australia. The application of data analysis and interactive visualisation methods has enabled the Library to model different ways of identifying print materials that can be deselected. The Working Group provided a leadership role in fostering data-related professional development and learning opportunities for library staff and clients. One such example was the hosting of regular 'hacky hour' sessions which inspired others to embrace new tools such as R, RShiny, Python, Jupyter Notebooks and GitLab to solve real-life challenges in their day-to-day work.

Keywords: library data, data analytics, data visualisation, refurbishment, collections, resource usage, data literacy, data skills

Other Library Experiences with Data

Murray [2019] summarises data analysis quite succinctly as a process which begins with a question, and through the use of data, along with a healthy dose of curiosity and tenacity, attempts to find answers which can be difficult to find. Murray also notes consumers of effective data visualisations need less time to process the information and gain quicker insights. Experiences with embedding data analysis methods into the library field can be categorised into five areas: the use of data analysis and visualisation to provide insights into library activities; the need to upskill library staff in data tools and methods; the increasing demand for data literacy programs from library clients; the importance of partnerships and collaboration in meeting needs for data services; and the use of library facilities to provide data-rich learning spaces for visualisation, simulation, and immersion.

A number of academic libraries have reported on their application of data analysis and visualisation processes to internal library data. Client use of library print and electronic collections is a commonly cited data source [Perry et al., 2018] and learning analytics projects at the University of Huddersfield and the University of Wollongong aimed to demonstrate a relationship between library use and student attribution and attainment [Perry et al., 2018; Cox & Jantti, 2014; Pepper & Jantti, 2015; Stone & Collins, 2013]. Academic libraries are displaying their internal data in interactive, real-time 'dashboards' to highlight library activities to other areas of their institution, for example Georgetown University Library's data dashboard using Google Sheets and Google Docs and Northeastern University Libraries' use of Tableau to show basic assessment statistics [Guhde, 2018; Braun, 2018].

Beyond demonstrating statistics, libraries are using data to inform business decisions. Singapore Management University Libraries use heat maps to display physical occupancy through library buildings which is made available to clients on their website – this data helps clients know which areas are occupied in real-time, as well as informing library space planning [Yahava, Sangaram, Ali & Tay, 2017]. Libraries have shared examples where data visualisation has informed planning for inter-library loan services [Guhde, 2018] and mapping collections usage against floor space to determine which items to relocate off site [Braun, 2018].

In their report on ‘Shifting to data savvy’, Burton, Lyon, Erdman & Tijerina [2018] describe the need for academic libraries to support data-focussed research, to support clients in cleaning, manipulating, and wrangling of datasets, using a range of tools such as OpenRefine, Python, Jupyter, R, Tableau and D3. University of Arizona Libraries initiated their data science support program in response to high demand for data literacy support, which had to be balanced with limited resources [Oliver, Kollen, Hickson & Rios, 2019].

As part of the drive towards data-focussed services, libraries need to consider upskilling existing staff as well as seeking expertise outside the library sciences [Oliver, Kollen, Hickson & Rios, 2019]. The ‘Shifting to data savvy’ report characterises a skills gap, where librarians perceive that they lack technical skills to support data-focussed research, and a management gap, where managers lack an understanding of the benefits of data skills. The library community has worked towards addressing this skills gap by establishing programs such as Library Carpentry and 23 Research Data Things [Burton, Lyon, Erdmann & Tijerina, 2018].

Along with the skills gap, is the crucial need for libraries to build partnerships and collaborations with other areas of their institutions, such as IT, the Research Office, Faculty, and the wider community [Burton, Lyon, Erdmann & Tijerina, 2018]. Libraries can leverage partnerships by inviting partners to deliver sessions within library spaces [Oliver, Kollen, Hickson & Rios, 2019]. Library staff can apply their skills to data projects, for example library staff from Shell Australia participated in a project to manage geophysical data, providing advice on metadata, controlled vocabulary, naming structure and other best practices, while collaborating with a range of specialists such as geophysicists, IT, data analysts, and database architects [Johnson, 2017].

Finally, libraries are responding to the demand for a range of next generation learning spaces for clients to visualise their data projects. Libraries are well placed to do this, as they provide cost-neutral facilities, a welcoming environment, and spaces to facilitate cross-disciplinary partnerships [Burton, Lyon, Erdman & Tijerina, 2018].

Insights into Library Activities - Curtin’s Experiences with Data

Background

Whilst physically located at Curtin University’s Bentley campus, the Library delivers services to staff and students regardless of location, with a strong emphasis on delivery of online resources and educational materials. The Robertson Library building at Bentley houses more than 500,000 physical items across four floors as well as holdings in offsite storage. As of the end of 2018, the Library staff group comprised approximately 75 FTE (full time equivalent) staff members, equating to a headcount of 87 continuing/fixed term staff.

Similar to other academic libraries, Curtin Library faces many upcoming challenges, including a multi-year building refurbishment. Design stages are complete and construction is scheduled to begin in 2019. One of the key deliverables for this refurbishment is increasing seating capacity by 50% across the fixed floor area to accommodate projected future enrolments. This objective has put great pressure on the library’s physical collection, which will need to reduce to one third of its current footprint to accommodate a wide range of user-centred learning spaces.

Coinciding with the refurbishment planning, the Library established a working group to focus on how the Library might implement data analysis and visualisation methods in operational and strategic activities. The working group were keen to experiment with new tools, in order to determine how they might create efficiencies in tasks as well as inform business decisions based on quantitative data.

Tools and Software

Initial explorations by Library staff made use of the open source programming language R and the development environment RStudio. Later staff used the interactive web based visualisation tool RShiny to analyse and visualise library related data, providing an evidence-based approach to inform decisions about library activities. R/RShiny far exceeded the capabilities of the existing tools used for analysis and visualisation by the Library. Other visualisations were implemented using the subscription based Tableau platform, as the experience of other libraries suggested this was a valuable tool [Braun, 2018]. The group found Tableau to be expensive and thus had less potential than R/RShiny, which could deliver the same outcomes at no cost.

Having created RShiny visualisations, the next challenge was how to make them available to their intended audience. Users who have RStudio installed can easily execute RShiny apps. The challenge was to make the apps available in web browsers for an audience unskilled in R, including Managers who only need to view the apps for reporting purposes. There are free RShiny app hosting services available on the web, however these typically wouldn't limit use to only library staff, deemed necessary given the potentially sensitive nature of some internal data. As an alternative, the group selected internal library servers and Gitlab, a web-based tool which allows non-IT staff to upload their RShiny apps and has a continuous deployment feature to ensure automatic server updates without the need for intervention by IT staff.

Other tools investigated included the browser based Jupyter Notebooks, which contain a mix of rich text elements and executable R and Python code. These notebooks are ideal for data literacy lessons which mix an instructional narrative with coding activities. They can also be used for reporting purposes as they are easily updated simply by executing the code. The group also investigated and used the equivalent tool in the R universe called R Notebooks. The programming language Python has a library, Pandas, which mirrors a lot of the functionality in R and was also used in data literacy programs delivered by the group. Open Refine was used for data cleaning efforts.

Library Data Projects

Many areas relating to the usage of the physical library were explored using data analysis and visualisation tools by the working group. Detailed below are three examples and how the data was used to influence planning for the refurbishment project.

1. <u>Client Interaction Statistics</u>

Australian University Libraries submit an annual set of statistics to CAUL (the Council of Australian University Librarians) which includes data on collections, physical item use, expenditure, and other operational areas [CAUL, 2019]. The annual collection process can be time-consuming as reported data must meet certain definitions and criteria. Client interaction statistics, part of annual CAUL reporting, was a good starting point for developing interactive data visualisations. Previously performed by manual tallying of statistics on paper and then utilising a myriad of spreadsheets, this task was automated by the group using Google Forms and R/RShiny, automating the annual collation for CAUL and delivering interactive web pages which could be accessed by library staff to provide up to the minute visualisations of client interaction statistics. An example is shown in Figure 1, detailing headcounts by floor at various times. The group also undertook a major data cleaning exercise using Open Refine by extracting the data from the previously used spreadsheets to expose over 10 years of previously 'hidden' historical data, revealing long term trends in client interactions.

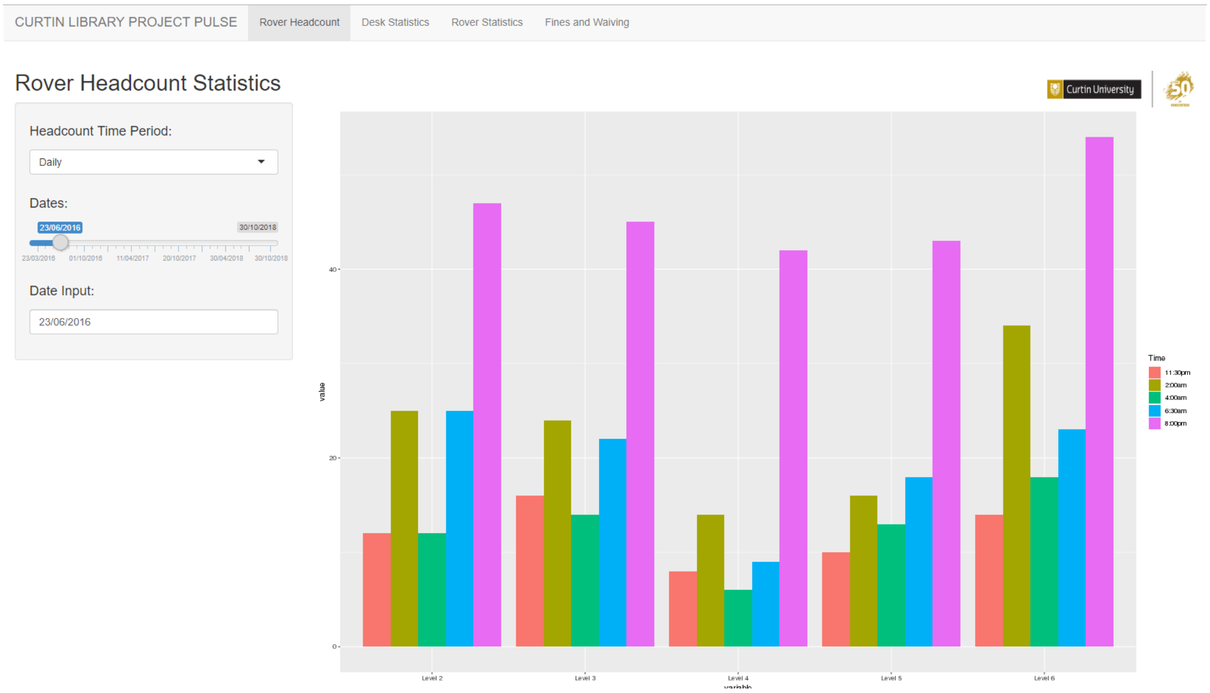

Figure 1. Curtin Library, Bentley Campus, Client Engagement statistics dashboard (RShiny).

2. Managing Physical Collections

One of the key criteria in deselecting physical collection items is to evaluate if the holding is the 'last copy' held across all library collections. An early application of R was developed to select holdings satisfying our existing criteria for deselection, including the last copy status. This filtering was previously done manually by staff, who identified candidates for deselection based on a hardcopy listing of all physical holdings whilst checking the shelves. The manual process led to inconsistencies and many staff errors, and was extremely inefficient. The use of code reduced these errors and inefficiencies significantly, and accelerated the timeframe for the deselection. Without the code, the project seemed insurmountable as potentially all 500,000+ items needed to be manually reviewed. With the code the deselection was targeted to around 20,000 items, and the deselection was achieved utilising existing library student assistants hours with no additional staffing support required.

R coding was also used to optimise the library student shelving effort, where better data and knowledge of the movement of physical items over the year allowed improved shelving rosters. Figure 2 is a visualisation which emerged out of this work, using a tree map to show for each Dewey number the relative size of the collection and popularity (loans). The darker shade of green represents more loans.

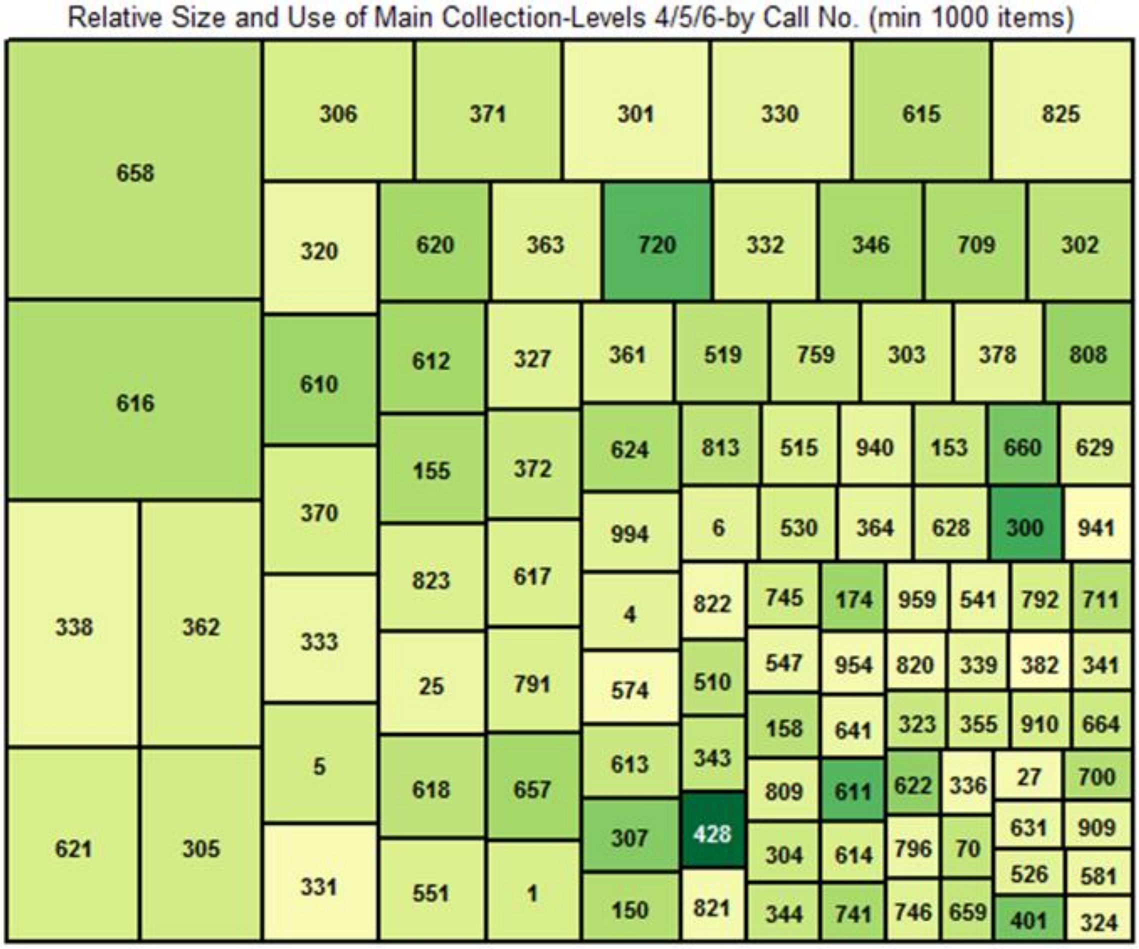

Figure 2. Relative size and use of Curtin Library's Physical Collection, Bentley Campus by Call or Dewey No. (min 1000 items, R).

Trove is a searchable discovery layer administered by the National Library of Australia and aggregates holdings across all Australian libraries. The group produced another visualisation using the publicly available Trove API to show Australian holdings by subject area, including common holdings between institutions. R/RShiny was used, in particular the powerful igraph and visNetwork packages. An example visualisation is shown in Figure 3, using the subject 'Gardens, Japanese'. The node size represents the size of a library's holdings, and the paths or edges between the nodes represent the common holdings between libraries. Hovering over the node will give full library details, whereas hovering over the paths or edges will reveal the number of holdings in common. The user can select the minimum number of holdings a library needs for it to be displayed, allowing display of simple through very complex shared holdings networks. Essentially the visualisation shows each library's holdings for that subject in context with other libraries.

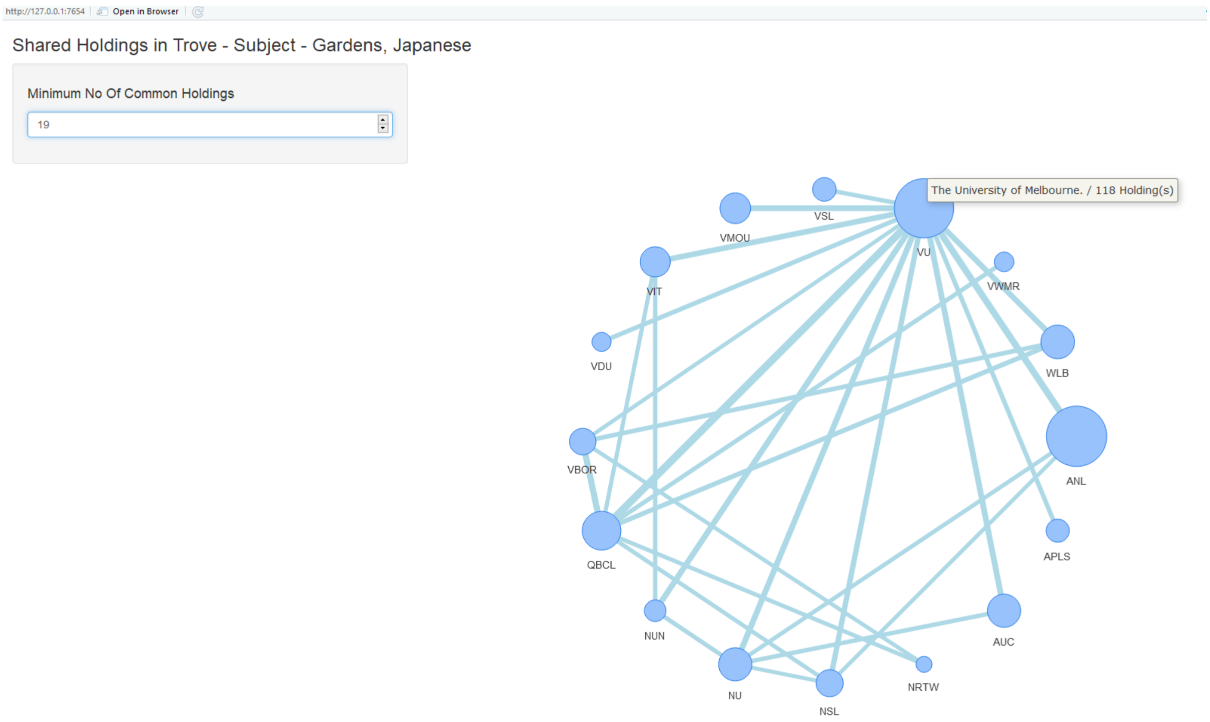

Figure 3. Shared Library Holdings in Trove – Subject : Gardens, Japanese, Minimum common holdings : 19 (RShiny).

One challenge the group faced was the growing use of data stored in human readable hierarchical type formats such as XML or JSON, which for spreadsheet users is very different to work with. The group encountered several of these data sets, including those from the TROVE and ORCID API's. The data extracted from the Trove API was challenging; multi layered and deeply nested. The group created R based methods and code for working with this data.

Given this grounding in the effective use of R coding and visualisation with the management of the physical collection, staff then focussed on the need to address the refurbishment project objective to reduce the physical collection footprint and to consider relocating it from the current open access shelving to a large closed access compactus model. Given its height and complexity of operation, initial thoughts were to limit retrievals and reshelving to staff only. To inform this business decision, exploratory work took place to consider the staff effort required to manage this new configuration, effort which would be required for the lifecycle of the closed access solution. Data modelling was undertaken using R/RShiny on the most recent semester's data on physical item retrievals and reshelving, including loans, returns, hold requests, rejected holds and acquisitions. The model simulated a mixture of a large closed access compactus along with some small open or 'high demand' access shelving, graphically detailing over the semester the number of physical items moved between the open and closed components and also the size of the open access component which would be needed. The RShiny visualisation allowed variation of model parameters including the hold request rejection rate and a 'loan history profile' of what constituted a 'high demand' holding. Interestingly the data and modelling revealed that what was seen as a relatively low use collection with a relatively small 'high demand' component would still require a considerable staff effort to effect retrievals and reshelving in the proposed closed access version, versus client effort in the existing open access version, and based on this a more balanced mix of open and closed access compactus and more generous open access shelving was pursued. Figure 4. shows the modelling for one particular set of circumstances.

Figure 4. Collection modelling for Robertson Library refurbishment planning (RShiny).

The refurbishment still requires significant physical collection deselection. To aid in this effort Australia wide physical holdings data were downloaded in bulk using the Trove API. Visualisations were created in RShiny (refer Figure 5.) and Tableau to explore how many other libraries in Australia had the same holdings as Curtin, moderated by the library being in WA or interstate and whether a University library or otherwise. Of the approximately 300,000 items retrieved, other libraries held around 90% of items.

Due to the variations in the datasets contributed by Australian libraries, the data obtained from the Trove API required significant cleaning and validation. One large stumbling block here was the lack of consistent and unique holding identifiers across all library holdings, and the ability to separate electronic from physical holdings.

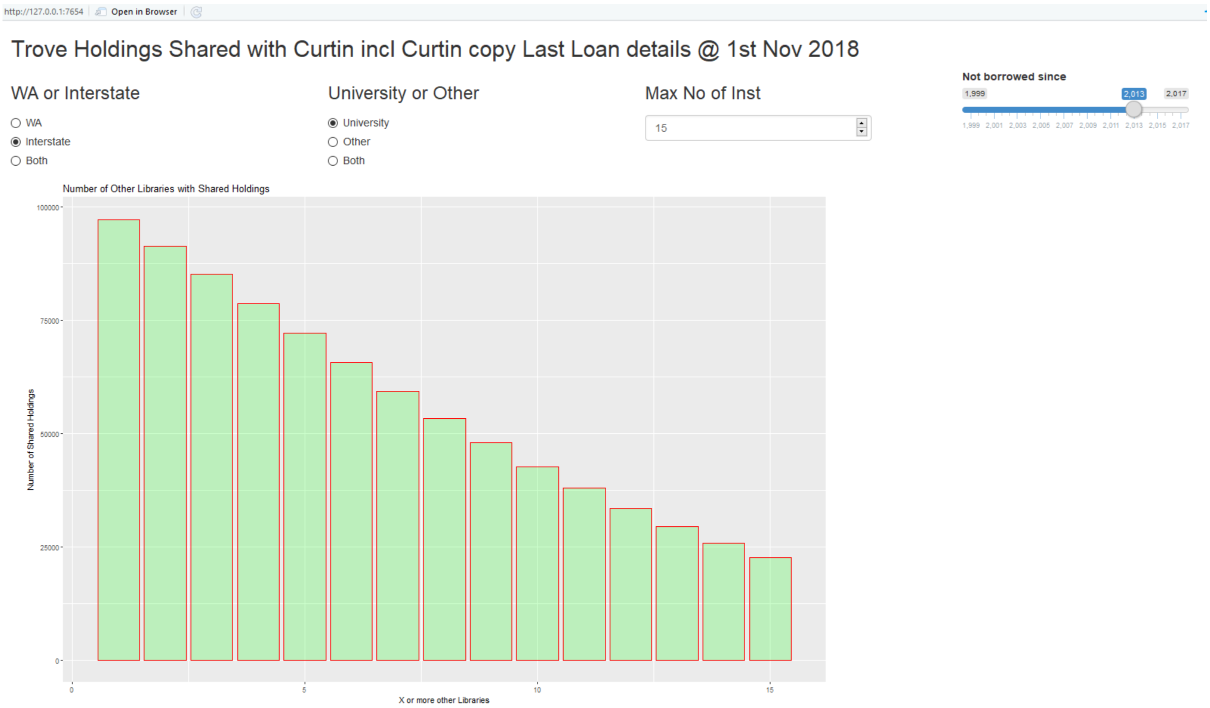

Figure 5. Robertson Library, Curtin's Bentley Campus, holdings in common with other Australian Libraries – Totals (RShiny).

3. ICT Resource Usage

The third area of focus in planning for the refurbishment was to consider the number of fixed computer workstations installed, at the time around 440 in total across the library. A visualisation was created in RShiny (refer Figure 6.) using data from Labstats, the logging software employed by Curtin's IT services, which revealed at which locations in the physical building our library clients preferred to use fixed workstations. The RShiny app allows the user to select any time of day to the nearest 15 minutes interval to see what workstations were in use and the session length. Incomplete tagging of library computer workstations required significant cleaning to allow the usage visualisations to be accurate.

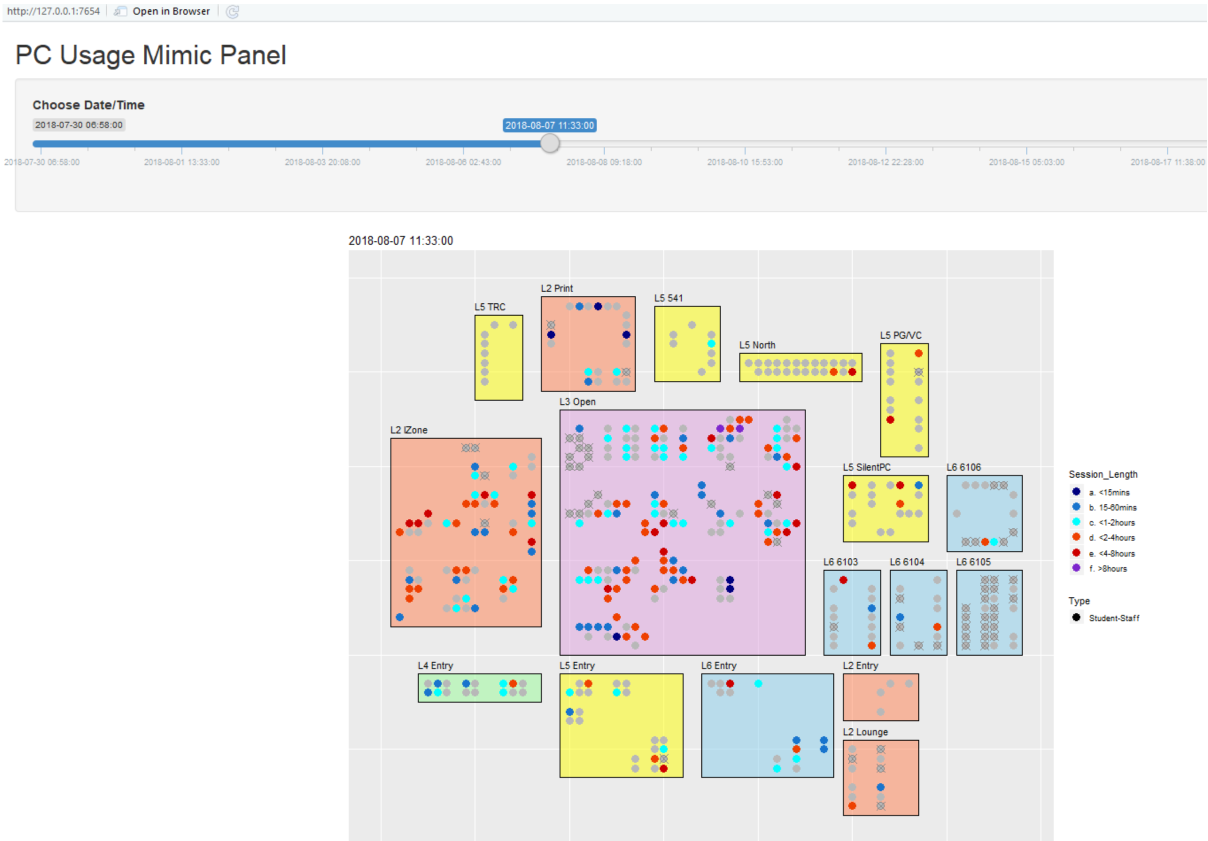

Figure 6. Robertson Library, Curtin's Bentley Campus, PC Usage Mimic Panel over period Jul-Aug 2018 (RShiny).

The library is a focal point for students with an assorted collection of personal mobile devices. The library's wi-fi service is freely available to students with a Curtin identity and password and is one of the densest and most active wi-fi points on campus. The library, however, did not have a clear picture of wi-fi usage in the building. At our request, Curtin's IT services provided raw wi-fi AP usage data for the first few weeks of Semester One, 2017. Library staff imported the data into Tableau Desktop and produced an animation that captured the essence of usage patterns over time. This visualisation (Figure 7), with a clear demonstration of heavy use over peak periods, reinforced the importance of wi-fi access points for library users going into the refurbishment plans.

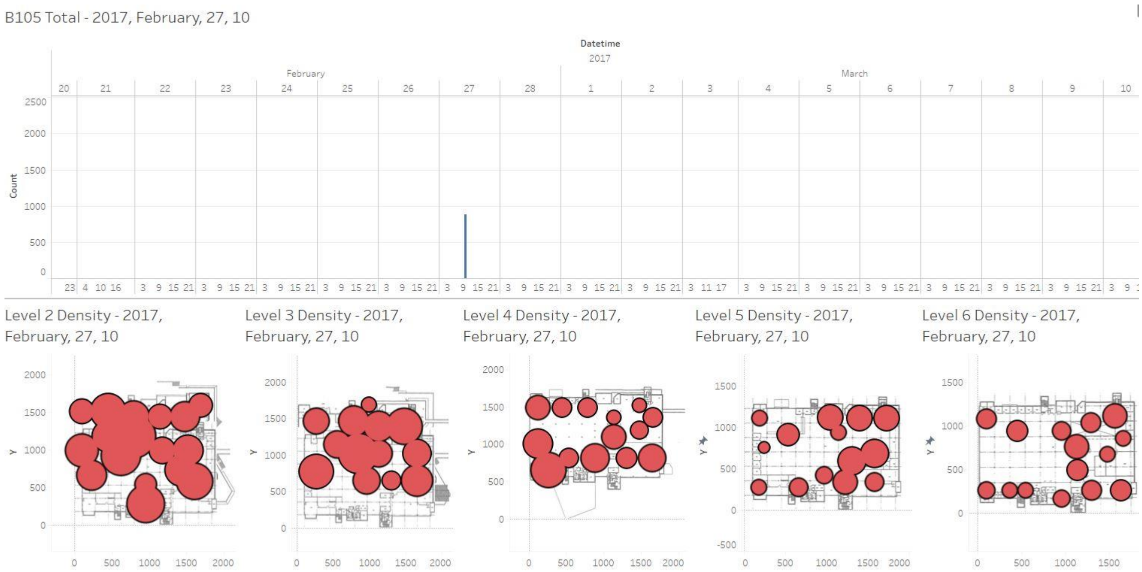

Figure 7: Robertson Library, Curtin's Bentley Campus, wi-fi usage per floor level using red bubbles to visualise density of wi-fi usage over time (Tableau).

Data literacy for Library Staff and Library Clients

In 2019, the group reflected on their own data journey, and began development of lesson modules aimed at increasing student's data literacy. The intent was not to duplicate the considerable range of resources currently available. Two approaches were used.

The first approach was to create a series of Jupyter notebooks, each introducing a basic principle or theory useful when working with data, supported by coding in R or Python. Not aimed at beginners, these lessons are intended to fill perceived gaps in online data literacy training. Fordham [2019] provides some examples of the lessons, including that shown in Figure 8. They are freely available on Github and easily run using Google Colab, and also include lessons working with the challenging hierarchical data types such as XML or JSON.

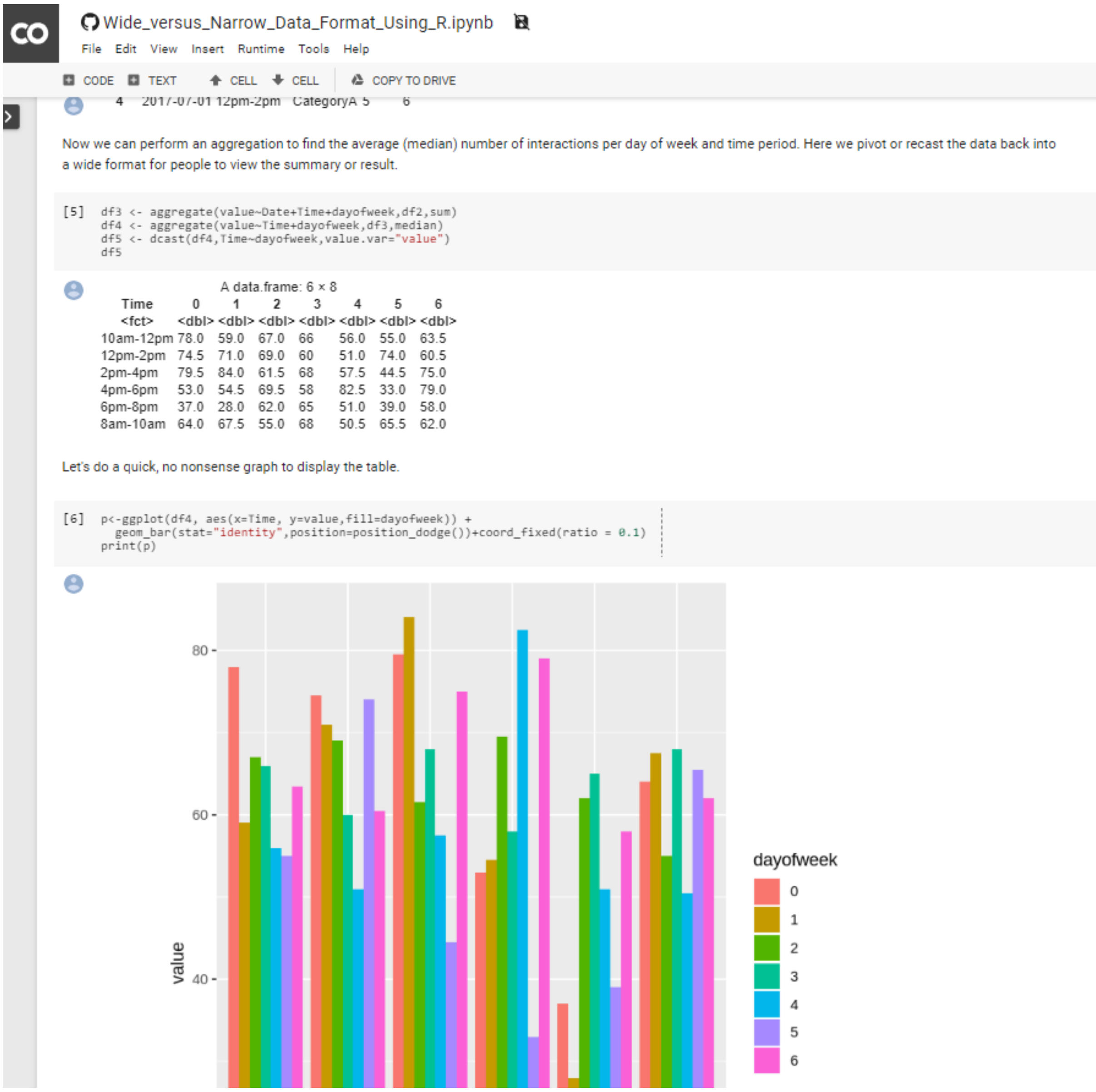

Figure 8. Example of Jupyter notebook data literacy lesson.

The second approach involves the bold vision, only at early stages so far, to take students who are beginners at using R, and prepare them with everything that they need to run cloud based big data storage, analysis and visualisations.

The group also identified the need for library staff to upskill in data literacy, and the accompanying need for all library staff to provide data literacy skills programs to library clients. The Library organised regular ‘hacky hour’ sessions open to all library staff and students from the Library School at Curtin to discuss data issues, and data literacy sessions were delivered using the Jupyter notebook lessons described earlier. Regular meetups between staff led to the development of informal relationships to solve work-related challenges. A particularly inspiring example was an engineering student working in the library. This student would always reach for a spreadsheet in the first instance, but under advice resisted that urge and developed skills in R/RShiny, and successfully constructed interactive web based reporting summaries and visualisations for Managers to use, shown in Figure 1. This developed skillset was immediately useful, assisting the student to obtain their subsequent graduate engineering position.

Partnerships and Collaboration

The issue of relationship-building between the Library and other areas of the University, including central IT services, was crucial. Relationships were built with staff who not only had an interest in data, but could provide the Library with access to actual data sets which were simply waiting to be analysed.

Strategically, library staff regularly liaised with IT staff to keep track of forthcoming projects to build data infrastructure, for example the use of hardware and software to track facilities use. Operationally, regular processes were established where data was fed from central systems to the library, allowing analysis and visualisation on wi-fi data and PC workstation usage.

Group members participated in informal community of practices, exchanging ideas, sharing visualisations and providing insights with academics, researchers and other professionals from other areas of the Curtin University community. From a session with a potential PhD student seeking ideas, a sharing of R skills with the Research Office, and working with University Properties regarding wi-fi usage patterns within the library building, the group expanded the library's professional network.

Another rich area of collaboration was the Library's involvement in a Curtin HIVE (Hub for Immersive Visualisation and eResearch) Intern Project. The Intern Projects are open to Curtin students who develop visualisations to screen in the HIVE physical facility, under supervision of an academic staff member over an intensive ten week period [Curtin University, 2019a]. The Library made available an anonymised and cleaned up version of its EZProxy data log files. EZProxy is the authentication system used to log in library clients to online databases and the log files were a source of information on date, time, URL, IP address and other transactional information. Green [2019] provides further detail into this project, the range of data visualisations developed by the student, and benefits of the outputs.

Use of Library Facilities

An outcome of the refurbishment will be a focus on new smart technologies which will transform the library into a 'living laboratory'. Digital screens within the current Library provide a selection of real-time graphics and figures relating to the building's energy usage [Curtin University, 2019b]. The vision of the 'living laboratory' expands this concept significantly with more interactivity and a wider range of data visualisations, all accessible via the Curtin website.

Working group members provided feedback in refurbishment planning focus group sessions relating to the 'living laboratory' concept. This included proposing the inclusion of a wide range of sensors, which allow opportunities for creating interactive installations reflecting the building's current state; for example a wall that changes colour or shape based on sound levels. Sensors also allow for historical analysis and identification of long term trends. The current plans include a dedicated 'data lab' along with large screens for visualisations located at the main entrance.

Conclusion

What began as an interest group exploring possibilities led to a direct impact on planning and professional development opportunities. Data was gathered from traditional and new technologies, and also analysed and visualised in innovative and creative ways. The visualisation of the library datasets provided both intuitive and quantitative perspectives across a broad range of library activities, and inspired staff to form informal relationships to work with their data to solve real-life challenges in their day-to-day work.

From the outset, the Group could not have predicted the wide range of activities that were undertaken as part of the project. By focussing library staff time on developing their expertise on data analysis and visualisation methods, and making ourselves visible with data work to other areas of our institution, the Library was exposed to many opportunities and partnerships. These experiments with data led to a greater understanding of the value of library data, and the challenges in working with data.

Biographies

Drew Fordham is in the Client Engagement team at Curtin Library, and draws on previous experience as an automation engineer to use data to shine light onto library problems.

David Lewis is a Systems Analyst with Curtin Library, and is interested in the application of data analysis, visualization, and processing to real-world problems.

Amanda Bellenger is the Manager, Research and Copyright at Curtin Library, and is passionate about upskilling library staff to prepare them to meet future challenges.

List of Software Tools

EZProxy - https://www.oclc.org/en/ezproxy.html

Github - https://github.com/

GitLab - https://about.gitlab.com/

Google Colab - https://colab.research.google.com/

igraph – R/Python - https://igraph.org/

Jupyter Notebook - https://jupyter.org/

Labstats - https://labstats.com/

Open Refine - http://openrefine.org/

ORCID API - https://orcid.org/organizations/integrators/API

Python – Pandas - https://pandas.pydata.org/

R - https://www.r-project.org/

RShiny - https://shiny.rstudio.com/

RStudio - https://www.rstudio.com/

Tableau - https://www.tableau.com/

Trove API - https://help.nla.gov.au/trove/building-with-trove/api-version-2-technical-guide

visNetwork – R - https://datastorm-open.github.io/visNetwork/

References

Braun, S. (2018, September 12). *Using analytics to extract value from the Library's data: part one: analytics behind the scenes* [Video file], *https://www.niso.org/events/2018/09/part-one-using-analytics-extract-value-librarys-data*

Burton, M., Lyon, L., Erdmann, C. & Tijerina, B. (2018). Shifting to Data Savvy: the future of data science in libraries. Project report. University of Pittsburgh. http://d-scholarship.pitt.edu/33891/1/Shifting%20to%20Data%20Savvy.pdf

CAUL (2019). Statistics. Retrieved from https://www.caul.edu.au/programs-projects/statistics

Cox, B.L. & Jantti, M.H. (2014). Unlocking the potential of library generated data to assess value, impact, and influence. Paper presented at IATUL Conferences, Purdue University, United States, pp. 1-9, https://scholars.uow.edu.au/display/publication91982

Curtin University (2019a). Curtin HIVE. Retrieved from https://humanities.curtin.edu.au/research/centres-institutes-groups/hive/

Curtin University (2019b). Welcome to our Living Campus. Retrieved from http://livingcampus.curtin.edu.au/

Fordham, D. (2019) data-literacy-workshops. Retrieved from https://github.com/drewfrobot/data-literacy-workshops

Green, P. (2019). *Dynamic and immersive visualizations of library user research activity using data from EZProxy log files*. Paper presented at the THETA Australasia 2019 Conference: The Tipping Point, Wollongong, NSW, https://doi.org/10.6084/m9.figshare.8180753.v1

Guhde, E. (2018, September 19). *Using analytics to extract value from the Library's data: part two: actionable data analysis* [Video file], *https://www.niso.org/events/2018/09/part-two-using-analytics-extract-value-librarys-data*

Johnson, V. (2017). Leveraging technical library expertise for big data management. *Journal of the Australian Library and Information Association*. 66(3), 271-286, https://doi.org/10.1080/24750158.2017.1356982

Murray, E. (2019). What is the difference between data analysis and data visualization? Retrieved from https://www.forbes.com/sites/evamurray/2019/03/29/what-is-the-difference-between-data-analysis-and-data-visualization/#49c54afe6411

Oliver, J. C., Kollen, C., Hickson, B., & Rios, F. (2019). Data Science Support at the Academic Library. *Journal of Library Administration*, 59(3), 241-257, https://doi.org/10.1080/01930826.2019.1583015

Pepper, A. & Jantti, M. (2015). *The tipping point: how granular statistics can make a big difference in understanding and demonstrating value*. Paper presented at ALIA Information Online 2015: at the edge, Sydney, pp. 1-15, https://scholars.uow.edu.au/display/publication98516

Perry, M. R., Briney, K. A., Goben, A., Asher, A., Jones, K., Robertshaw, M., & Salo, D. (2018). SPEC Kit 360: Learning Analytics (September 2018). https://publications.arl.org/Learning-Analytics-SPEC-Kit-360/

Stone, G. & Collins, E. (2013). Library usage and demographic characteristics of undergraduate students in a UK university. Performance Measurement and Metrics. 14(1), 25-35, http://doi.org/10.1108/14678041311316112

Yahaya, N., Sangaram, D., Ali, S.M., & Tay, A. (2017). SMU Libraries: Sensory in library spaces. *Singapore Journal of Library and Information Management*. 45, 61-73, http://ink.library.smu.edu.sg/library_research/96