emerald PUBLISHING

Library Hi Tech

Analysis of EZproxy server logs to visualise research activity in Curtin's online library

Journal:	*Library Hi Tech*
Manuscript ID	Draft
Manuscript Type:	Original Article
Keywords:	Data visualisation, EZproxy server logs, User-centered design, Collection Management, Academic libraries, information resources management

Analysis of EZproxy server logs to visualise research activity in Curtin's online library

Introduction

Curtin Library has a substantial dataset of logged, authenticated use of its online library collection, comprising databases, eJournals and eBooks dating from 2013. The EZproxy software writes about 30 million lines a month, and this rich dataset was accessible for this project.

Making sense of and drawing meanings from the raw data, which is mainly in the textual format of URL codes, is nearly impossible. The Curtin library team reported an unsuccessful attempt ten years ago, reporting that it was impossible to comprehend so much data using the human eye, as it all looked similar, with no distinct observable trends in users' information-seeking behaviours.

Scholarly findings about visualisations based on EZproxy data have been traditionally static, as in the work by Bhaskar et al. (2014), Coombs (2005), Chan (2014), Grace and Bremner (2004), Lewellen et al. (2016) and Sharman (2017). Dynamic visualisations are uncommon in this field, but Archambault et al. (2015, p.1) argue that a framework for meaningful data visualisation has merit:

> There are advantages to presenting data visually rather than as a set of flat statistics. Proper data visualization facilitates the recognition of patterns and relationships to communicate a message in a more compelling and interesting way. It allows the complexity of the data to be understood more easily.

We agree with Archambault et al. (2015) that the creation of a set of visualisations that is dynamic and immersive would offer friendlier interactions with the dataset, allowing one to look at it in a way that normally is not possible. The creation of a 3D virtual space where one could move around and view the data freely was envisioned to provide the library team with an immersive encounter with the dataset and an opportunity to explore their users' information seeking behaviour.

An opportunity to work with the EZproxy dataset was made possible with technical and financial support from the Curtin HIVE (Hub for Immersive Visualisation and eResearch) Internship Program, which allows a Curtin student to undertake a ten-week, full-time investigation of the application of visualisation technologies to a discipline area. Interns have regular access to the HIVE, are supported by its expert staff and supervised by a library and information science discipline leader and the Curtin Library team; the latter were the clients for this project.

In this paper we describe a project that aimed to visualise the EZproxy dataset to draw inferences about Curtin library users' information seeking behaviour and collection use in the virtual/online environment. EZproxy offered a rich source of data for analysis, containing a detailed log of HTTP requests processed through the library's authentication servers. Another user identity dataset was merged with EZproxy to identify each users' profile: whether staff or student; their geolocation data; and if their requests were processed in or out of Curtin's IP subnet. Analysis of these logs led to visualisations of hidden trends in users' collection of information, providing insights into where and when users accessed the virtual library, what information seeking activities they performed, whether they browsed or downloaded full-text journal articles, and which online databases they accessed frequently.

Visual awareness and understanding of these trends will help the library to make important budget decisions about their collection subscriptions and assist with strategic planning about their service delivery options (Grace and Bremner, 2004).

Background

Curtin University Library

Curtin Library supports the University's learning, teaching and research requirements and supports more than 58,000 students and 3,700 staff worldwide. Its online collection comprises at least 250,000 eBooks, 15,000 streamed videos, 150,000 journal subscriptions, 600 electronic databases and 48,000 institutional repository records. The library's acquisitions budget is over $10 million.

Access and authentication are administered by EZproxy, software licensed by OCLC. EZproxy reduces the number of authorisations for users and ensures remote access to content is secure and complies with licensing arrangements.

What is an EZproxy server?

EZproxy is an internet proxy server that is primarily used to first authenticate then allow computers outside a library network to access content provided by the library without requiring additional logins. Figure 1 illustrates how this works: when a user accesses the Curtin catalogue and attempts to access content on an Online Database Vendor's website, for example ProQuest, they are authenticated by the library's authentication system, which then creates a session with the EZproxy server and redirects the user's browser there, allowing access to all subscribed library content.

[Insert Figure 1: How the EZprozy server works?]

All hypertext transfer protocol (HTTP) requests that are sent through EZproxy are logged by its server. For example, if a user is directed to a website through Google Scholar, once they are authenticated by Curtin's Authentication server, all requests for this website will be logged within the dataset, creating a valuable set of information about who accesses resources via the EZproxy server that can be utilised to visualise users' information search and retrieval behaviours.

A user may access the library's online databases from a personal browser by going to the library catalogue or A – Z list of subscription databases, or through Google Scholar. When a user is successfully authenticated, a cookie is sent to the browser by EZproxy. The browser then presents this cookie for each access to EZproxy, allowing EZproxy to check the user's access rights each time. EZproxy handles access by URL rewriting: taking the URL requested by a user and modifying it so that the web server holding the content accepts the request. This removes the need for the user to use a separate login, as they are already authenticated by EZproxy. An example is when a user tries to access the online database 'proquest.com': the user's browser will be informed of the URL that contains the cookie, 'proquest.dbgq.lis.curtin.edu.au' —everything after ProQuest is the cookie. ('URL Rewriting', 2018)

How other Academic Institutions use the EZproxy Server and Dataset

A literature review shows how other academic libraries handle their EZproxy data logs. Academic libraries from the United Kingdom (Sharman, 2017; Grace and Bremner, 2004) and the United States (Chan, 2014; Coombs, 2005; Lewellen et al., 2016) report their experiences working with EZproxy, how their findings assisted in improving their service, and understanding of the usage patterns of their electronic resources. The Open University library reported using their EZproxy server data to evaluate their electronic resource subscriptions comprehensively and monitor resource usage trends by staff and students; the findings enabled them to measure the performance of their library services, and led to the development of performance indicators to measure their service quality and

patron impact (Grace and Bremner, 2004, p. 164). Sharman (2017) reports how the University of Huddersfield combined their EZproxy dataset with book loans from the library management system and statistics from library visits; they found correlations between the low use of library resources and final grades among Chinese students compared to their UK peers.

In the United States, Coombs (2005) reports how the SUNY Cortland University used the EZproxy dataset to gain insights into online collection use. Chan (2014) discusses how student course enrolment data was merged with the EZproxy dataset at California State University to develop personalised e-library services. Lewellen et al. (2016) describe how the University of Massachusetts Amherst used the dataset to investigate the use of e-books compared to print books.

These studies demonstrate that it is possible to extract useful information about trends and usage patterns from EZproxy datasets, and that these can be used to improve library services in universities. However, working with the complex raw EZproxy dataset is neither easy nor user-friendly, as revealed by the East Kentucky University's experience. This university analysed its EZproxy dataset to determine use patterns because this information was not provided by the online database vendors. Hence, the university firstly used an OpenURL link resolver, which provides a set of reports that enable robust analysis of user behaviour, as shown in Figure 2.

[Insert Figure 2. An example of OpenURL results (adapted from Smith and Arneson, 2017, figure 3)]

This snapshot of a report generated from the EZproxy dataset is detailed, but the textual data is visually unappealing—in fact, it looks just how our project's EZproxy data displayed when exported into MS Excel.

The second method that East Kentucky University used to extract data from their EZproxy data logs was by using the command-line grep tool, as shown in Figure 3. The grep tool is a Unix based pattern matcher that searches through plain text data sets in the target file and outputs all lines that match a regular expression of these patterns. Both the construction of these patterns and working with the grep utility tool are daunting tasks for novices.

[Insert Figure 3. Example of grep on a small sample file
('Pipe, Grep and Sort Command in Linux/Unix with Examples', 2018)]

There are three disadvantages to using this method. First, the output of this tool is in the same format that it goes in, as it does not change the lines; instead, it writes the files that match the pattern to a new file. Second, the single grep command-line will not work for all queries as it cannot cope with the varying file formats used by different database vendors. Third, it does not enable the development of data visualisations, as its functionalities are limited to searching the file for certain queries.

Given our research aim to find out about Curtin's online collections usage trends, an understanding via the literature review on how others achieved this was of interest. Hence, Morton-Owens and Hanson's (2012) research about how they analysed EZproxy logs through the creation of dashboard charts based on EZproxy data was useful. They report running a few basic calculations to highlight significant changes within their data, which produced two charts indicating trends in resource use and variations in use of specific resources. It was our understanding that often additional analytical software is required to make sense of this complex EZproxy data. Thus, it was useful to read that Austin College used Google Analytics on their EZproxy dataset to identify which were the main resources being used, when (time/day) the most use took place each week, and what devices were being used to access their library (Ajamie et al., 2014).

EZproxy server logs are a rich source offering immense opportunities for harnessing their data to tell stories of how library users search for information and how the library's information

sources are used at any time, or 24/7/365. Presenting this information visually will assist both in planning and management of the library collection and in improving its service delivery (Grace and Bremner, 2004; Kay, 2014). Knowing how students use library resources can aid in focusing on teaching those information literacy skills most needed; and to identify groups of students needing specific remedial skill teaching (Kay, 2014). It can also enable the delivery of e-library services, such as the personalisation of library websites/portals for students and staff in specific discipline areas to provide rapid access to their most frequently used online resources (Chan, 2013; Grace and Bremner, 2004, p. 162;).

This review identifies a gap in the work about using data visualisation tools and techniques to interrogate the rich but complex EZproxy dataset. At the moment, all data is presented as static flat graphs and tables. No-one so far has developed an interactive search interface that is friendly to library decision-makers, who are not trained to work with or understand complex computing datasets. There is also a lack of built-in user-friendly search interfaces that offer library decision-makers the opportunity to explore visually their clients' techniques of information-seeking. In the absence of these functionalities, it is not yet possible for library professionals to have an immersive experience that enables them better to understand their clients' usage patterns. Hence, it is worthwhile to offer librarians both a familiar and an ease to use search interface that is similar to their online library catalogue interfaces to search for usage information about their online resources. Further, these data visualisation tools will provide a visually immersive experience whereby they can intuitively explore the dataset to identify usage patterns and statistics. This project offers a new and friendlier way for librarians to analyse datasets that have traditionally been difficult for them to understand.

Definition of information seeking behaviour

Given that the EZproxy server logs staff and student interactions with various online databases and electronic resources provided by the library. It is necessary to define the broad behaviours the findings from the EZproxy datasets reveal. The information seeking behaviour literature is vast and only key definitions that are relevant for this project are described in our article. We adopt Joseph et al.'s (2013) definition: that seeking information 'relates to the process of identifying an information need, then sourcing and accessing the necessary information avenues to address that need'.

We consider information retrieval is a crucial activity in the information-seeking process, especially in the context of this research: that users wish to fulfil their informational needs from the widest possible pool of electronic resources. Meadow et al.'s (2007, p. 2) definition that information retrieval 'involves finding some desired information in a store of information or a database' is applicable here. Information 'search behaviour' also needs to be defined. Joseph et al. (2013) state that information search behaviour comprises 'search processes' and 'search activities'. Search processes involve several sequential but iterative stages of judgement, option selection and decision-making (Ellis 2005; Henefer and Fulton 2005; Kuhlthau 1988; Leckie et al. 1996; Marchionini and White 2007; Meho and Tibbo 2003). Joseph et al. (2013, p. 6) state:

> Search activities refer to the actions users enact during the iterative process of moving the information search from start to closure. These actions include browsing, navigating or extracting information. Both information search processes and search activities comprise the search behaviour of users of these systems.

These simple and practical definitions are useful guides to our discussions on the research aims of our project described next.

Research aims

The current research does not have research questions; however, our project's research aims will enable the development of research tools that answer questions raised in later research.

Our research aims are to develop interactive and immersive data visualisation prototypes with user-friendly interfaces to enable the Curtin library team to explore how the online library is being used by its patrons. These prototypes will allow them to explore answers to the following questions:

1. Where is the online library use taking place geographically?
2. What are the peak and off-peak time periods of use?
3. How do use patterns change over the course of the academic year?
4. What types of resources are being used?
5. What is the entry point for users?

Research methodology

Three sequential steps were performed to develop the desired visualisations for this project. First the EZprozy dataset was curated into a format that enabled easy extraction of useful data from the logs and the ability to remove all lines in the log file that did not represent actual search data. This was done by extracting the data into a TSV (Tab Separated Value) format, then running a script written to parse over all the files that had been collected and remove all lines with a URL that contained any file formats, indicating data unrelated to actual search data. This included such things as the required webpage data (HTML or CSS) and images/logos (GIFs, JPEGs, PNGs).

Second, the data was placed into a MySQL database which could then be queried by Unity, the game development engine that was chosen to create the 3D visualisations. Unity scripts were written using the C# programming language to query and extract the data required to develop the visualisations.

Third, we wrote scripts to allow Unity to read in the data we wished to visualise. The scripts were written in Unity to process this data and create the visualisations, plus the environments they take place in. The decision to choose the Unity3D graphics engine was made early in the project because of its ability to create virtual 3D environments by simply dragging and dropping models and textures/images into the editor and then dragging and dropping these into the visualisation. Unity's ability to create and run C# scripts made it possible to read the data from either text files or the MySQL database, and to instantiate objects based on this data. Unity's simple user interface made it user-friendly and easy for beginners to use and learn.

Curation of the EZproxy Log File Dataset

Our project's dataset includes five years of EZproxy log data; Curtin University Library has been logging data since 2013. The data is stored within text files, one file for each day, to 2016. From 2017, owing to the rise in usage and in the resultant volume of log files generated, separate files for morning (AM) and post-morning (PM) use each day were created.

The original data was stored in the Combined HTTP Log format ('Combined Log Format—Just Solve the File Format Problem', 2018), but while the data provided for this project resembled this format, along certain fields were anonymised, while geolocation data along with a staff/student field were added. The fields within the log file were:

- IP address: normally this is the IP address of a user, but to comply with research ethics and privacy requirements this field was hashed and salted using AES-256 encryption to ensure that users remained anonymous ('AES 256 Hardware Encryption—Safe and Secure Encryption', 2018).
- Patron ID: usually the ID of the user; but again, to comply with research ethics and ensure personal privacy, this field was hashed as described above.
- Date-time: this was the date and time of a request sent through the EZproxy server; it was logged in the 'day/month/year hour:min:sec' format, along with the time zone (UTC+0800 Perth time).
- HTTP request method and URL: 'RFC 7231 - Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content' (2018) describes individual users' search request:
    - There can be multiple requests for each web page, based on items that need to be sent.
    - A given URL may not necessarily represent whether what was sent was a full text or a different type of item.
    - It is hard to determine what requests are for: as an example, if an eBook is ordered whole eBook is not downloaded at once; nor does an eBook have a specified format— some come as PDFs, others as images of each page.
- Status code from whether a HTTP request was successful.
    - Codes starting with 2 are successful requests: the data that the user requested arrives.
    - Codes starting with 3 are requests that were successful, but some prior redirecting was required.
    - Codes starting with 4 indicate that an error happened on the client side of the request; e.g. the data was not found or permission was denied.
    - Codes starting with 5 indicate that an error happened on the server side; e.g. internal server error or service unavailable.
- Size of HTTP response (in bytes)
    - These give the file size of the data sent to the user. One can infer that smaller responses are probably search requests, while larger ones are usually a resource like a PDF.
- Referrer URL
    - This field outlines the URL from which the user's request was sent. This data can be used to determine the web page the user was looking at before initiating the current request.
- HTTP User Agent
    - This refers to the user's web browser (e.g. Google Chrome or Internet Explorer) sending the request, and identifies the browser and device being used to make the request. However, as it is difficult to tell just from looking at the user agent what browser was used, this remains an educated guess.
- Geolocation
    - This field is generated from the IP address before it is anonymised. It includes a flag indicating if a request came from within Curtin's network or if the requests are external. In some cases this data may not be accurate if the location service has not been able to provide 100% accurate data.
- Patron Type
    - This field classifies library patrons as either 'Staff' or 'Student'. For privacy and anonymity reasons, patrons' identities cannot be revealed.

Issues with Data

A few key issues with the dataset had to be resolved before we could begin work on any visualisations of the dataset. All but one specific dataset persisted to present issues in retrieving meaningful data from the log files. The first of these was that EZproxy logs all HTTP requests, and many of these are for parts of a website that would not be useful to visualise as they do not represent a user using the library but rather parts of a webpage being sent to the users' browser, such as images on the webpage or CSS stylesheets. To remove these requests from the data, we created a file parser, which reads each of the files and checks if the HTTP Request URL contained any file format that represented this unwanted data, such as .css, .html, and .gif. This process took approximately 60 hours to run over a weekend, given the large amount of data, roughly 1.2 Terabytes of raw text data.

The second issue was the format of the data and how to get the Unity engine to interact with it in its current form. A decision was made to store this data on a MySQL database, allowing us to write C# scripts that Unity could run to query the database. The process of uploading the data to the database was quite slow, and storage of the database also caused problems.

The third issue, that of difficulty in extracting meaningful data, was the format of the URLs. Each database provider that Curtin Library subscribes to uses a different URL format, and some of these are mainly hashed. Despite this we were able to extract meaningful data that enabled the creation of visualisations.

Results – Visualisation Prototype Developments

Our project developed two main visualisations to showcase research activity in Curtin's online library space, described next. Both these prototypes provide a friendly, interactive and immersive user experience to provide meaningful approaches to visualising and making sense of otherwise complex EZproxy datasets.

Global Visualisation of Curtin Research Activity

The first visualisation prototype is referred as the Global Visualisation of Curtin Research Activity (henceforth Global Visualisation). It uses a geographical map of the world as a platform to show from where each research request comes from, the time the request is made, and how large the file size of the request is (Figure 4).

[Insert Figure 4. Global Visualisation of Curtin Research Activity]

Global Visualisation allows a user to search the dataset by specifying a date via the search menu to trigger the visualisation. Global Visualisation follows a reverse waterfall model, where each request is represented by icons to indicate 'user types' and their 'request status type'. Staff and student user types are represented by a circle and square respectively. Square icons are also used to represent the status of users' requests, colour coded to indicate the HTTP status of the request. The traffic light style colour code was selected to indicate if a request was successful, redirected, not found, or an error. Finally, the location of the object on the world map represents the geographical location from which the request was sent.

A simple user-friendly interactive single search interface was developed to query the EZproxy data set by day, month, or year with the click of a button. Menu options using a pick list to select the day, month and year were developed to ensure the correct formats for these search terms were entered each time (Figure 5).

[Insert Figure 5. Single search interface]

A key to the prototype was added to tell users what the icons and colours represented (Figure 6). This key was easy to implement using Unity's user interface feature and is overlaid on whatever the camera sees; an effortless way to show users this information.

[Insert Figure 6. Key/Legend explaining what shapes and colours]

The visual simulation for the Global Visualisation was developed by programming the Unity software to read TSV (Tab separated value) files from an external file. For each line in the file that it read, it was programmed to parse the data and then instantiate the associated prefabricated object into the scene, which begins to rise in a reverse waterfall effect as more and more objects begin to float into the sky (Figure 4).

It is possible to obtain specific details about each research request by hovering the mouse over the icon to generate a pop-up menu displaying specific information about the request (Figure 7).

[Insert Figure 7. User interface displaying specific details about an individual search request]

Database Usage Visualisation

The second visualisation showcases the use of the various databases available through the Curtin library by both staff and students daily over a month in April 2017. This visualisation is the Database Usage Visualisation and (henceforth 'Database Visualisation'). It is achieved using three-dimensional bar graphs designed to be interactive (Figure 8) and based on a month's data from April 1 to 30, 2017. It is possible to run a custom script on the EZproxy server to allow more data to be added later.

[Insert Figure 8. Prototype of the Database Usage Visualisation]

This visualisation follows a 3D bar graph design with three axes, each representing a different category of data. Along the X axis is a listing of databases accessed, along with the URL for each database. The Z axis, 'Date', represents the time period when these databases were used, with each row representing a day's data usage of each database. The Y axis is represented by pillars whose height indicates the number of HTTP requests made and thus the level of use of a particular database on a certain day.

The labels on the scene floor in Figure 8 show the reader which online database each column correlates with. The colour of the pillars does not yet not represent anything, but later may be used to represent the height/usage of a certain pillar, or other useful information.

Another visualisation of this prototype (Figure 9) showcases the different use of databases by staff and students. The left visualisation represents use by staff; the right, by students. There is a stark contrast between the two. This prototype provides a quick visual comparison of database use by staff and students, and indicates whether the groups are using the same databases.

[Insert Figure 9. Database Usage Visualisation]

There are currently two primary ways of looking at this visualisation. The first is a free fly camera system view (Figure 10), which allows the user to move freely around the scene and investigate the data from any preferred angle, enabling the user to gain different insights based on how they choose to review and focus on the dataset.

[Insert Figure 10. Free Fly Camera view]

The second is a ground-based camera system view, where the camera does not move along the y-axis (Figure 11). It offers a more immersive experience, as a viewer can move between the pillars of

the graph as if walking through the data. This immersive experience is enhanced when experienced using the HIVE's cylinder screen, which provides a 180-degree field of vision. The use of the cylinder is helpful in the Database Visualisation prototype as it allows a wider field of view, giving the feeling of being immersed in the data. Viewing the Database Visualisation on the cylinder at the correct viewing angle creates a 'pegboard'-like effect, with the data appearing to extend from a wall (Figure 10).

[Insert Figure 11. Ground View Creating an Immersive Experience]

Users can manipulate these graphs to interact with the information they provide. The Database Visualisation provides users an interactive experience by allowing them to move around the graphs and rotate them to different angles, gaining different perspectives of the dataset. The Database Visualisation serves to help visualise how much the library is being used by staff and students, and total usage volumes within a specified timeframe. A current shortcoming of the prototype is its inability to search and retrieve a dataset by specific criteria; for instance, to present the use of databases by date, time, or database types. Consequently, in the current prototype it is not possible to dive deeper into the dataset to observe information behaviour patterns. This would be a valuable functionality to add in future enhancements of the Database Visualisation; it will be discussed later.

The Database Visualisation was built by writing a script to read through a selected subset of data, which then collated each of the base URL domains each time it saw this URL, then incremented the counter associated with that URL. Upon reaching the end of the file, the script created two new files containing the total amount of data for each URL domain, for both staff and students for that given day.

Digital Story of Curtin's Online Library Usage

An online platform to showcase the digital story of Curtin's online library usage was created to provide viewers with a background to the project and enable users to view both the Global Visualisation and Database Visualisation prototypes described above. The HIVE's cylinder screen platform was selected to tell this digital story as it provides a 180-degree stereoscopic 3D display of the visualisation. However, the production can also be viewed on desktop computers.

The viewer begins the digital story via the screen shown in Figure 12.

[Insert Figure 12. Welcome Screen for the Digital Story]

When they press the 'Begin' button viewers are presented with the second screen (Figure 13) and given a brief description of the project, including background information on Curtin Library and its collection.

[Insert Figure 13. Background information about the Project]

Next, viewers are invited to enter a door in front of them and move into the main room (Figure 14). They can then engage with and move between each of the two visualisations described earlier.

[Insert Figure 14. Main room enabling the user to move between visualisations]

Discussion

Given the limited research on dynamic 3D visualisation of EZproxy datasets, it was not possible to compare our project with other work. The literature review reports the use of data visualisation software like Tableau for visualising EZproxy data and 2D graphing software to create simple visualisations from EZproxy datasets. In comparison, the two Global and Database 3D visualisation prototypes developed for this project are more immersive and dynamic than the 2D presentations reported by Bhaskar et al. (2014). The ability to move around a scene provides a user-friendly interface to navigate and digest the complex EZproxy dataset information, as a large set of bar graphs cannot.

Design Considerations for the Prototypes

When designing the visualisations for this project, we considered how the data would be viewed and how different interactive environments could be developed using the Unity software. Unity was chosen for its ability to create immersive environments quickly and easily, and to present the content on multiple platforms using the same scripts and assets; it also has options to develop and present content for all computing and mobile platforms and via web browsers.

When we commenced the project, we discussed what we wanted to achieve. We decided that the visualisations should be in 3D, and that the viewer should be able to move freely around the data to gain the insights that they wished to. In short, the specifications were to design immersive and user-friendly interfaces to explore and make sense of the rich information dataset the EZproxy server logs files contained. Which of the many HIVE screens the project would be developed for was determined later, with the decision to showcase the visualisations on the HIVE cylinder screen interface. The cylinder screen's 180-degree, 3D field of view enables more content to be presented on screen, and in a more immersive manner, than a normal desktop computer screen creating a more immersive feel and effect, as shown earlier in Figures 9, 10 and in 15.

[Insert Figure 15. An example of the Global Visualisation on the HIVE's Cylinder screen]

The dome display, presented in Figures 16 and 17 was a secondary choice for the Database Visualisation prototype, as it fully encompasses the viewer's field of vision to create a feeling of being fully immersed (Figure 17) with the data.

The use of the HIVE screens assists in the presentation of visualisations, more than is possible with a regular desktop screen display. Our experience echoes Lugmayr et al.'s (2016) observations that it is possible to display large-scale data on these screens to create immersive experiences unlike anything a regular display can offer.

The environments in which the data is showcased are open spaces, allowing the user to move freely around and investigate the data. Within the Database Visualisation the user can fly around the environment to change the angle from which they look at the data. The ground-based movement is for a more immersive examination of the data, allowing the user to move at a slower pace and to feel as if they are immersed within the data; the effect is increased by the cylinder screens' panoramic field of vision.

[Insert Figure 16. An example of the Database Usage Visualisation prototype displayed on the HIVE's Dome screen at the HIVE Intern Showcase presentation]

[Insert Figure 17: An example of an immersive experience, moving though the Database Visualisation display on the HIVE'S dome screen at the HIVE Intern Showcase]

Future Work

Given this was a short 10-week proof of concept research project, there are many opportunities for future research and further development of the Global Visualisation and Database Visualisation prototypes, and to build new prototypes. Some of these opportunities are described below.

Further technical development

Continue development using the Unity 3D platform

As the foundational work for this project was developed in Unity, it is recommended that future work continues using this platform. Preparations for using the MySQL database rather than the raw text files have already been completed to allow Unity to query the database.

Unity query from database

The current visualisations rely on the data being in a text file. Preliminary work has been done to export the data to a MySQL database, so this is an area that could be further developed. Once all the data has been stored and hosted on a database that can hold large data quantities, a few changes will need to be made so that rather than reading files, Unity can query the database and then use that data to develop the visualisations. These queries would enable insights into specific usage patterns for PDFs or eBooks. Time-based queries would also be possible, allowing viewers to select a more dynamic time, rather than starting at midnight and only one day at a time.

Existing Visualisation Models: further development

Global Visualisation

The global visualisation can be expanded, first to improve the way that icons are instantiated in Unity. Currently one icon is instantiated every frame, which causes issues when moving through the timeframe. If a way is found to handle this differently, it could allow distance between clusters of icons to become more meaningful in regard to time gaps in the data. Second, the search functionality could be improved by providing more detailed search criteria instead of the simplified version implemented here.

Database Usage Visualisation

Two improvements are identified for this visualisation in future. The first is to build a search interface that enables queries about the use of specific online databases currently subscribed. This would be a desired feature for a library team wishing to find out how specific databases are being used and to make strategic decisions about future subscription to them. The second is to build a sorting functionality in the search interface so that visualisation could be based on parameters such as use, alphabetical, or other ways of sorting the data. This feature would allow users to view what is most important to them, quickly and easily.

Develop New Visualisation Prototypes

Inclusion of Faculty information

Adding academic faculty information to the EZproxy dataset, such as student enrolment data by academic staff, by faculty, or by discipline area, would enable insights into information behaviour and resource use by these faculties. When we began this project we considered adding which faculty the user who sent a request came from, to see how different faculties used the online library, and if there were any major differences; however, as this data was not requested at the very start, we did

not have it to work with; nor could we get it within the timeframe of the project. Adding this faculty information would offer a deeper understanding of how different faculties use the library.

Precedent research in other universities (Chan, 2014; Coombs, 2005; Grace and Bremner, 2004) using EZproxy datasets indicate that different faculties use the library differently, with each having preferences for databases and eBooks, and different levels of use. Having such information would enable the provision of targeted e-library services like those reported by Chan (2014), and Grace and Bremner (2004, p. 164): for instance, the development of personalised library websites/portals for different disciplines.

Chan (2014, p. 453) reports how California State University San Marcos Library merged its student course enrolment data with the EZproxy dataset 'to develop a system that automated the process of connecting users with the library resources most relevant to their research needs', reporting that such services offer 'simpler pathways for accessing online resources and enrich the overall user experience'. This model of library service delivery aligns with Curtin University's strategic vision to 'deliver a seamless, responsive and innovative digital environment' for learning and student experience (Curtin University, 2017, p. 5).

Profiling the average Curtin Library User

Profiling a visual image of Curtin University's 'average' user will assist with planning initiatives in library service. Given that EZproxy logs contain all the search history of all who use the library, the ability to track a single user's resource usage patterns over a period of time would be an interesting visualisation to develop. It would provide a glimpse into research activity over a longer period; and if performed with users from different discipline areas and faculties, could lead to the formulation of archetypes of researchers in different faculties, discipline areas and campuses. For example, as Curtin operates in several locations internationally, it would be a useful exercise to compare the use of someone at the Bentley campus user versus someone at the Singapore campus. User confidentiality could be retained by hashing and salting ID and IP addresses, allowing researchers to pick at random a few of these hashed values, and use all requests sent by them to create a timeline visualisation. It could be fruitful to compare and contrast the 'average' user for different faculties and campuses to visualise their differences and similarities.

Understanding Users' Information Search Habits

From the EZproxy server logs, it is possible to observe aspects of users' 'search processes'—to examine the paths users take to reach databases. Are they being referred by Google Scholar to Curtin Library's online databases, or are they being referred by the library's 'FindIt' resolver link from the catalogue? It is also possible to gain insights into aspects of their information 'search activities': that is, which databases are they browsing, and what information are they extracting from the online resources? Are they downloading e-books (Lewellen et al., 2016), abstracts, or full-text journal articles (Coombs, 2005; Grace and Bremner, (2004)?

Functionalities for understanding users' information search habits could be included as an improvement. One key factor of visualising research activity is to see what users search for, as well as what they actually download, when they browse the online library. By analysing the URLs in a thorough and detailed manner, it may be possible to ascertain whether a user has downloaded an item, looked at an item, or just searched the database—although this could be difficult as each content provider has differently structured URLs, which would not be possible to automate at this time. Visualising information browsing habits could provide fruitful insights into how the library is being used at a deeper level, seeing how individuals *use* the collection rather than what they use *in* the collection.

Conclusion

This short 10-week HIVE Intern research project highlights opportunities for developing interactive, user-friendly and immersive ways to visualise and make sense of the rich EZproxy dataset. It also indicates various avenues for expansion.

Both the Global Visualisation and Database Usage Visualisation prototypes provide visual evidence of the high volume of usage of Curtin Library's digital resources—eBooks and databases, and of the accessibility and usage of the library's digital contents at any time and from anywhere. This offers a visual demonstration of the often hidden demand for the library's online services and the frequency with which it is used by staff and students. It offers evidence of how the library supports the university's strategic goal of becoming a global campus by 2020, delivering courses internationally (Curtin University, 2016).

It empowers the library with evidence-based data visualisations that communicate in a compelling and interesting way (Archambault et al. 2015, p. 6) how its services and subscriptions support Curtin University's mission statement to 'transform lives and communities through education and research' (Curtin University 2016, p. 4).

Continuing with this project will provide more detailed insight into how the online library is used for research, including an understanding of the research habits of staff and students. Given that the online library is an important part of the university's infrastructure, it is vital to understand how it is being used so that future strategies to provide focused and efficient services are based on sound and thoroughly understood evidence. By expanding on the research opportunities discussed earlier, a more detailed and complete set of visualisations can be used to showcase how the library is used for research activity, ultimately leading to improvements in library service (Archambault et al. 2015, p. 6).

This HIVE intern project successfully achieved what it set out to do. First it curated EZproxy log files into formats required to feed into Unity software and develop visualisation prototypes. Second, it developed search interfaces that provided options to dissect the dataset into searchable chunks to understand usage patterns in the digital collections. Third, the visualisations revealed distinct demographic usage patterns by staff and students, including the geographical locations from which, and the times at which, the library's digital collection was accessed. Fourth, it provided visual displays of the usage patterns of different digital collections by staff and student user groups. This project provides a unique way of looking into research activity at a single University, and offers diverse options for expanding the project and creating new ways to analyse EZproxy log datasets.

References

Ajamie, L. *et al.* (2014), "EZproxy for electronic resource librarians: conference report", *Journal of Electronic Resources Librarianship*, Vol. 26 No. 3, pp. 203-205.

Archambault, S. G. *et al.* (2015), "Data visualization as a communication tool", *Vine,* Vol. 32 No. 2, pp. 1-9.

Bhaskar, R. *et al.* (2014), "LibraryVis: towards visually understanding library resource usage", workshop presented at IEEE Vis, 2014, 9-14 November, Paris, France, available at: http://homepages.ecs.vuw.ac.nz/~craig/publications/businessvis2014-bhaskar.pdf (accessed 21 February 2018).

Chan, I. (2014), "Leveraging student course enrollment data to infuse personalization in a library website", *Library Hi Tech*, Vol. 32 No. 3, pp. 450-466.

"Combined Log Format - Just Solve the File Format Problem" (2018), available at: http://fileformats.archiveteam.org/wiki/Combined_Log_Format (accessed 21 February 2018).

Coombs, K. A. (2005), "Lessons learned from analyzing library database usage data", *Library Hi Tech*, Vol. 23 No. 4, pp. 598-609.

Curtin University (2016), "Strategic plan 2017 to 2020", available at: https://strategicplan.curtin.edu.au/wp-content/uploads/sites/12/2016/12/2017-curtin-strategic-plan.pdf (accessed 12 February 2018).

Ellis, D. (2005), "Ellis's model of information-seeking behaviour", in Fisher, K.E., Erdelez, S., and McKechnie, L.E.F. *et al.* (Eds.), *Theories of information behavior*, Information Today, Medford, NJ, pp. 138-142.

Grace, C. and Bremner, A. (2004), 'Getting the value from evaluation: where to get the data and what you can do with it', *Vine*, Vol. 34 No. 4, pp. 161-165.

Henefer, J. and Fulton, C. (2005), "Krikelas's model of information-seeking", Fisher, K.E., Erdelez, S., and McKechnie, L.E.F. *et al.* (Eds.), *Theories of information behavior*, Information Today, Medford, NJ, pp. 225-229.

IETF (2018), "RFC 7231 - Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content", available at: https://tools.ietf.org/html/rfc7231 (accessed 21 February 2018).

Joseph, P., Debowski, S. and Goldschmidt, P. (2013), "Search behaviour in electronic document and records management systems: an exploratory investigation and model", *Information Research*, Vol. 18 No. 1, Paper 572.

Kay, D. (2014), "Discovering the pattern, discerning the potential: the role of the library in unraveling the cat's cradle of activity data", in *2014 Library Assessment Conference*, 4-6 August 2014, Seattle, Washington, Association of Research Libraries, Washington, DC, pp. 269-282.

Kuhlthau, C. C. (1988), "Perceptions of the information search process in libraries: a study of changes from high school through college", *Information Processing and Management*, Vol. 24 No. 4, pp. 419-427.

Leckie, G.J., Pettigrew, K.E. and Sylvain, C. (1996), "Modelling the information-seeking of professionals: a general model derived from research on engineers, health care professionals, and lawyers", *Library Quarterly*, Vol. 66 No. 2, pp. 161-193.

Lewellen, R., Bischof, S., and Plum, T. (2016), "EBL ebook use compared to the use of equivalent print books and other resources: A University of Massachusetts Amherst - MINES for Libraries case study", *Performance Measurement and Metrics*, Vol. 17 No. 2, pp. 150-164.

Lugmayr, A. *et al*. (2016), "Cultural visualisation of a cultural photographic collection in 3D environments - development of 'PAV 3D' (photographic archive visualisation)", in Wallner, G. *et al* (Eds), Entertainment Computing – ICEC 2016, Lecture Notes in Computer Science, Vol. 9926, pp. 272-277.

Marchionini, G. and White, R. (2007), "Find what you need, understand what you find", *International Journal of Human-Computer Interaction*, Vol. 23 No. 3, pp. 205-237.

Meadow, C.T. *et al.* (2007), *Text information retrieval systems* (3rd ed.), Academic Press, London.

Meho, L.I. and Tibbo, H.R. (2003), "Modeling the information-seeking behavior of social scientists: Ellis's study revisited", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 570-587.

Morton-Owens, E., and Hanson, K. L. (2012), "Trends at a glance: a management dashboard of library statistics", *Information Technology and Libraries*, Vol. 31 No. 3, pp. 36-51.

"Pipe, Grep and Sort Command in Linux/Unix with Examples" (2018), available at: https://www.guru99.com/linux-pipe-grep.html (accessed 21 February 2018).
OCLC (2018), 'URL Rewriting', available at: https://www.oclc.org/support/services/ezproxy/documentation/rewrite.en.html (accessed 12 February 2018).

Sharman, A. (2017), "Using ethnographic research techniques to find out the story behind international student library usage in the Library Impact Data Project", *Library Management*, Vol. 38 No. 1, pp. 2-10.

Smith, K., and Arneson, J. (2017), "Determining usage when vendors do not provide data", *Serials Review*, Vol. 43 No. 1, pp. 46-50.

Woods, A. J., Datta, S., Bourke, P., and Hollick, J., "The design, install and operation of a multi-user, multi-display visualisation facility", (In press).

Zybersafe (2018), "AES 256 Hardware Encryption - Safe and Secure Encryption", available at: https://zybersafe.com/aes256hardwareencryption/ (accessed 21 February 2018).

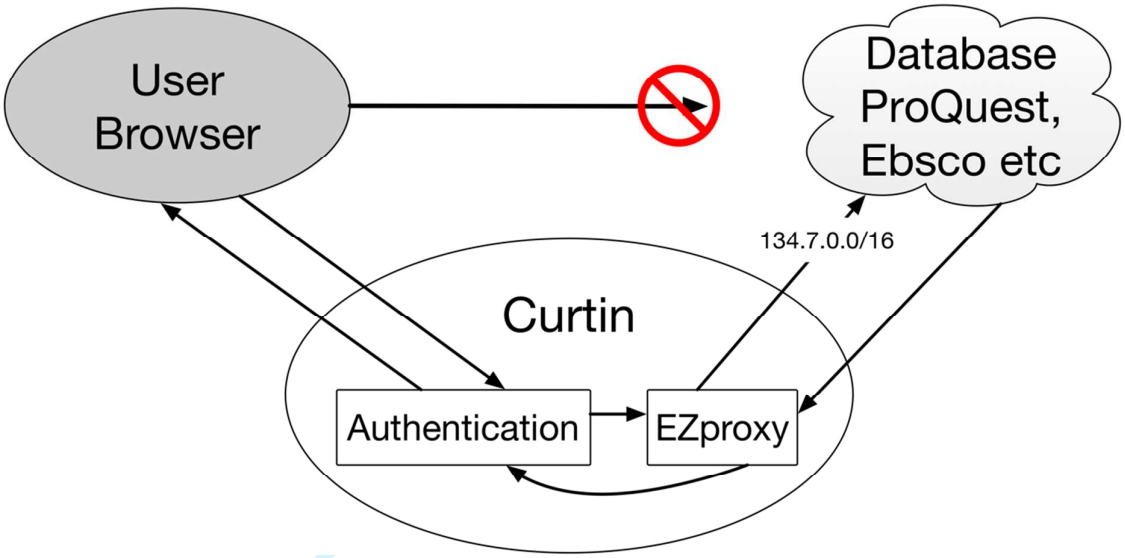

Figure 1: How the EZprozy server works?

Target	Clickthroughs 2010	Clickthroughs 2011	Clickthroughs 2012
Total:	38101	79192	80218
DOCDEL_ILLIAD	5745	13833	17464
EBSCOHOST_ACADEMIC_SEARCH_PREMIER	3543	8930	6690
ELSEVIER_SD_SCIENCE_DIRECT_COMPLETE	1417	3319	3889
OVID_JOURNALS_AT_OVID	768	2991	3834
MISCELLANEOUS_FREE_EJOURNALS	1770	2947	2691
LOCAL_CATALOG_EXLIBRIS_VOYAGER	2548	3823	2423
EBSCOHOST_BUSINESS_SOURCE_PREMIER	1132	2302	2070
EBSCOHOST_CINAHL_WITH_FULL_TEXT	1184	1648	1522
DOAJ_DIRECTORY_OPEN_ACCESS_JOURNALS_FREE	788	1538	1483

Figure 2. An example of OpenURL results adapted from (Smith and Arneson, 2017, figure 3)

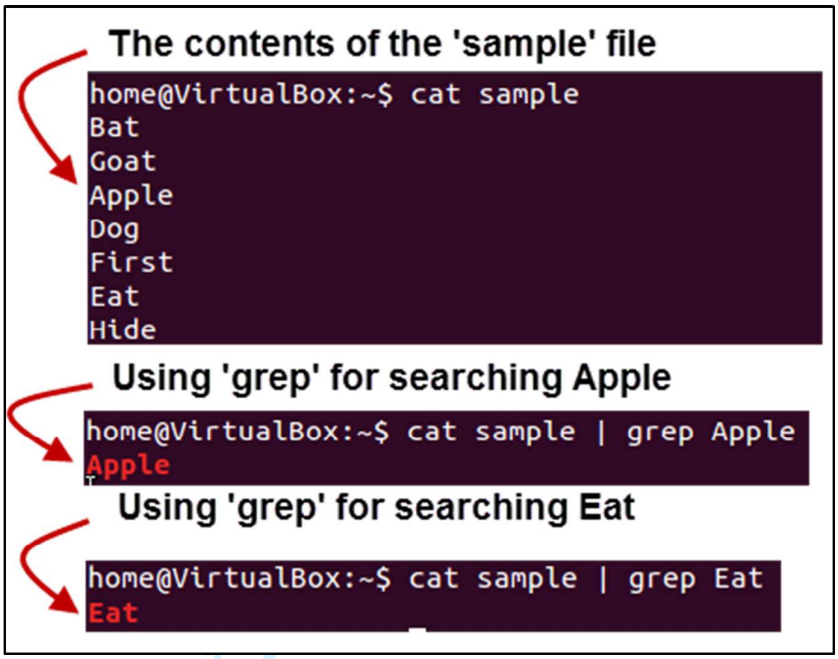

Figure 3. Example of grep on a small sample file
("Pipe, Grep and Sort Command in Linux/Unix with Examples", 2018)

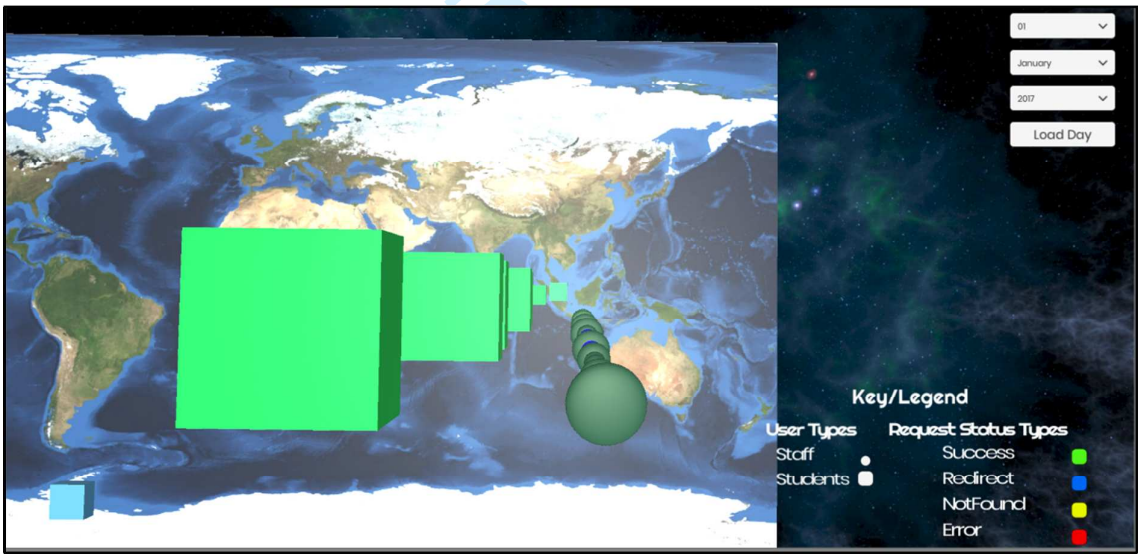

Figure 4. Global Visualisation of Curtin Research Activity

Figure 5. Single search interface

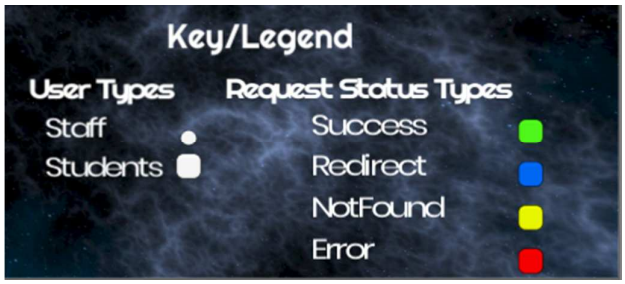

Figure 6. Key/Legend explaining what the shapes and colours mean

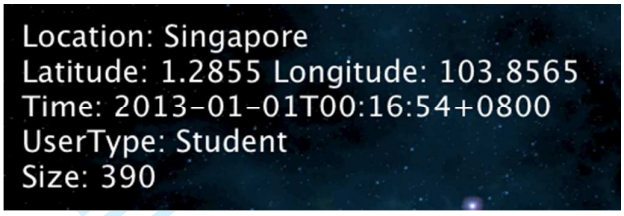

Figure 7. User interface displaying specific details about an individual search request

Figure 8. Prototype of the Database Usage Visualisation

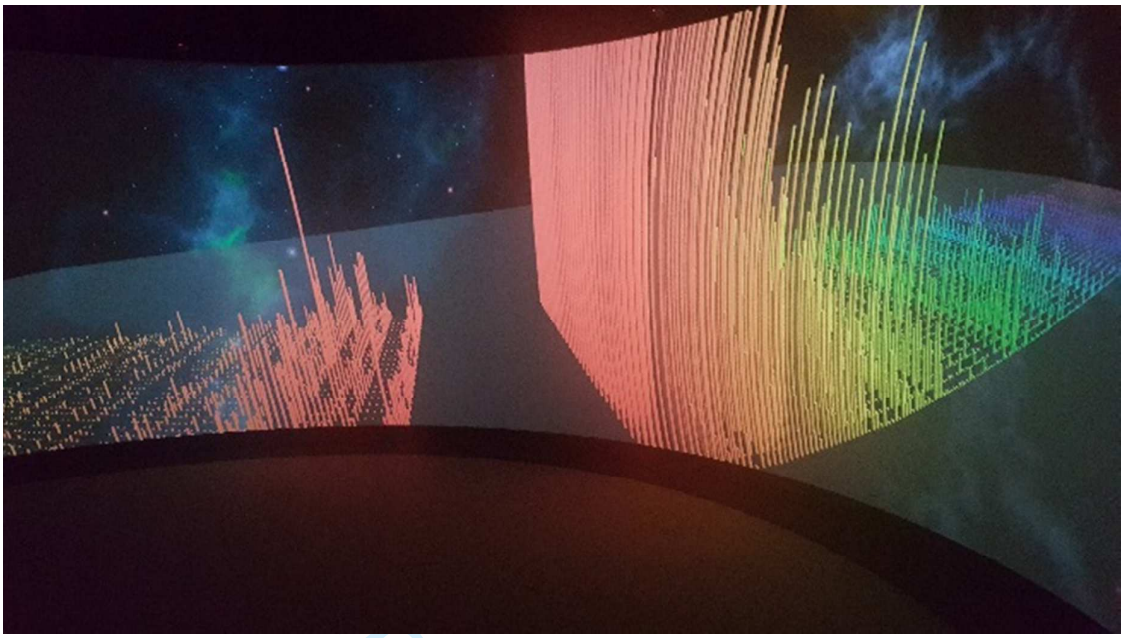

Figure 9. Database Usage Visualisation

Figure10. Free Fly Camera view

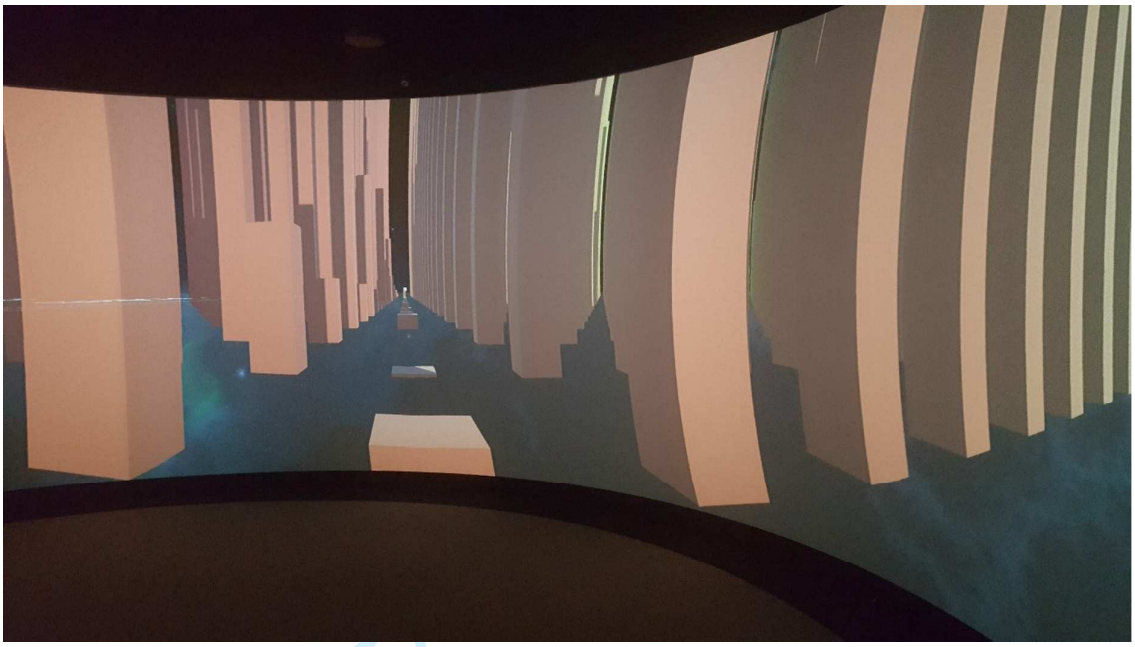

Figure 11. Ground View Creating an Immersive Experience

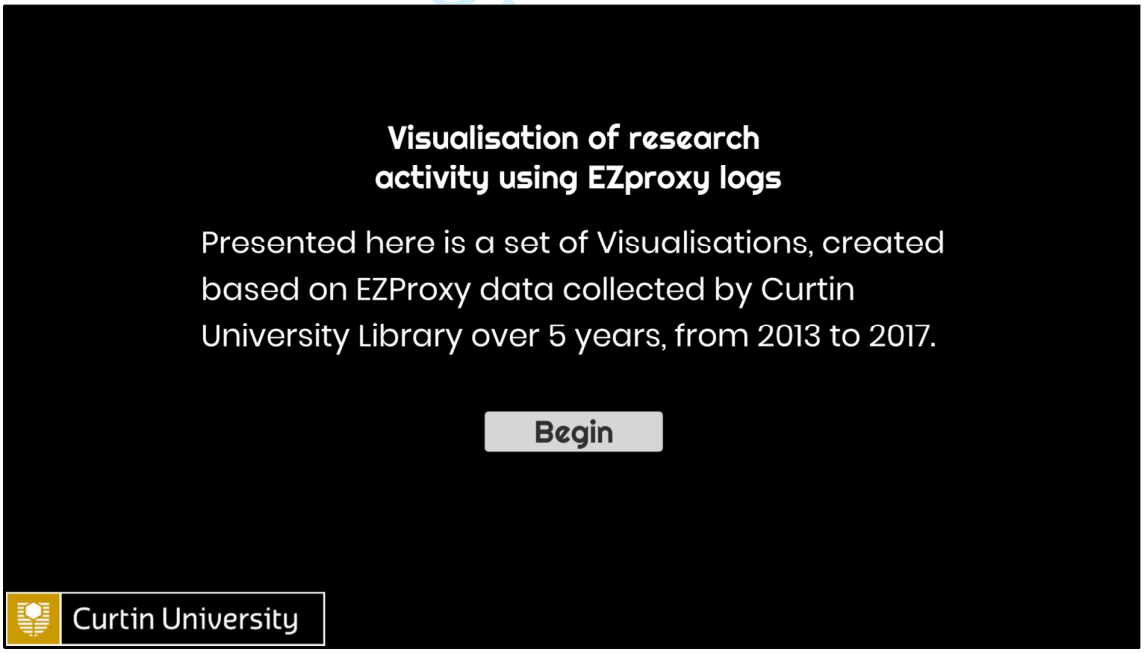

Figure 12. Welcome Screen for the Digital Story

Figure 13. Background information about the Project

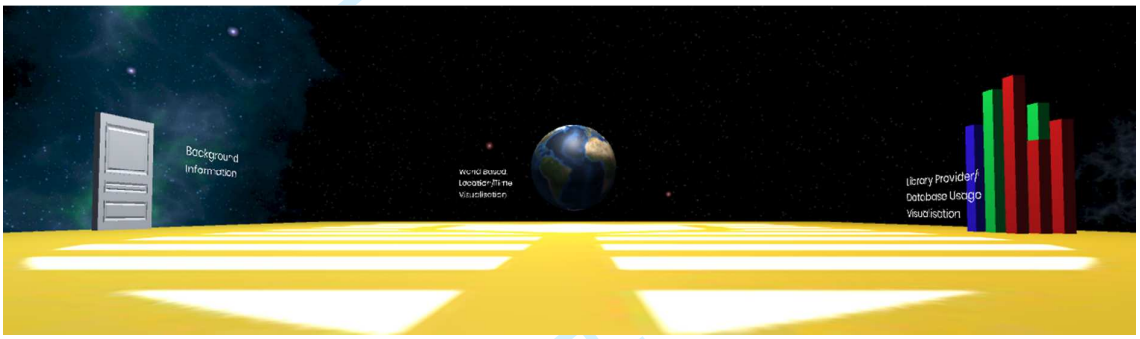

Figure 14. Main room enabling the user to move between visualisations.

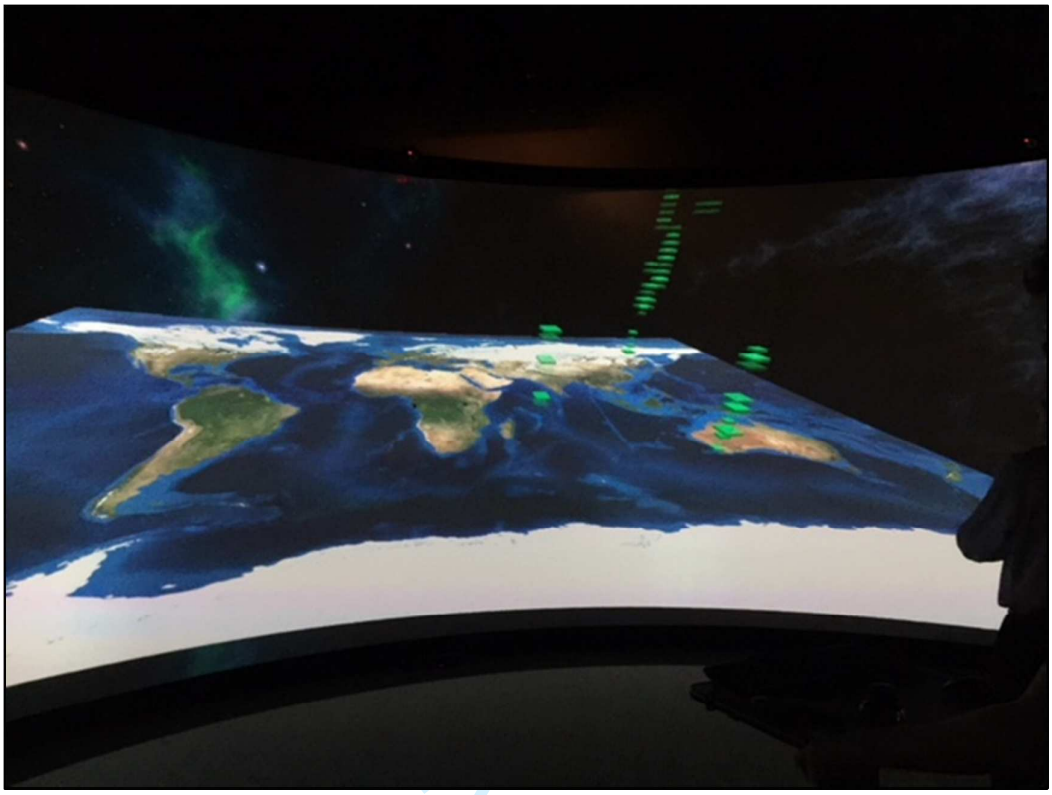

Figure 15. An example of the Global Visualisation on the HIVE's Cylinder screen

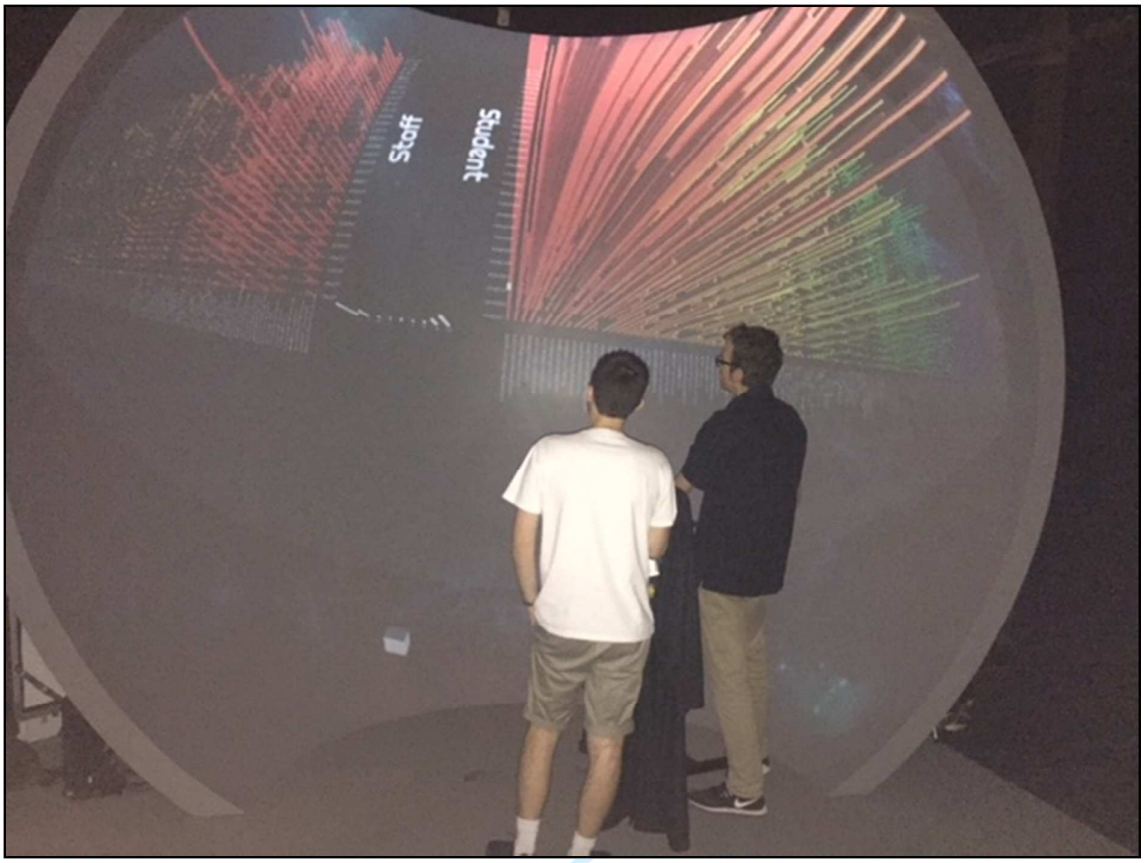

Figure 16. An example of the Database Usage Visualisation prototype displayed on the HIVE's Dome screen at the HIVE Intern Showcase presentation

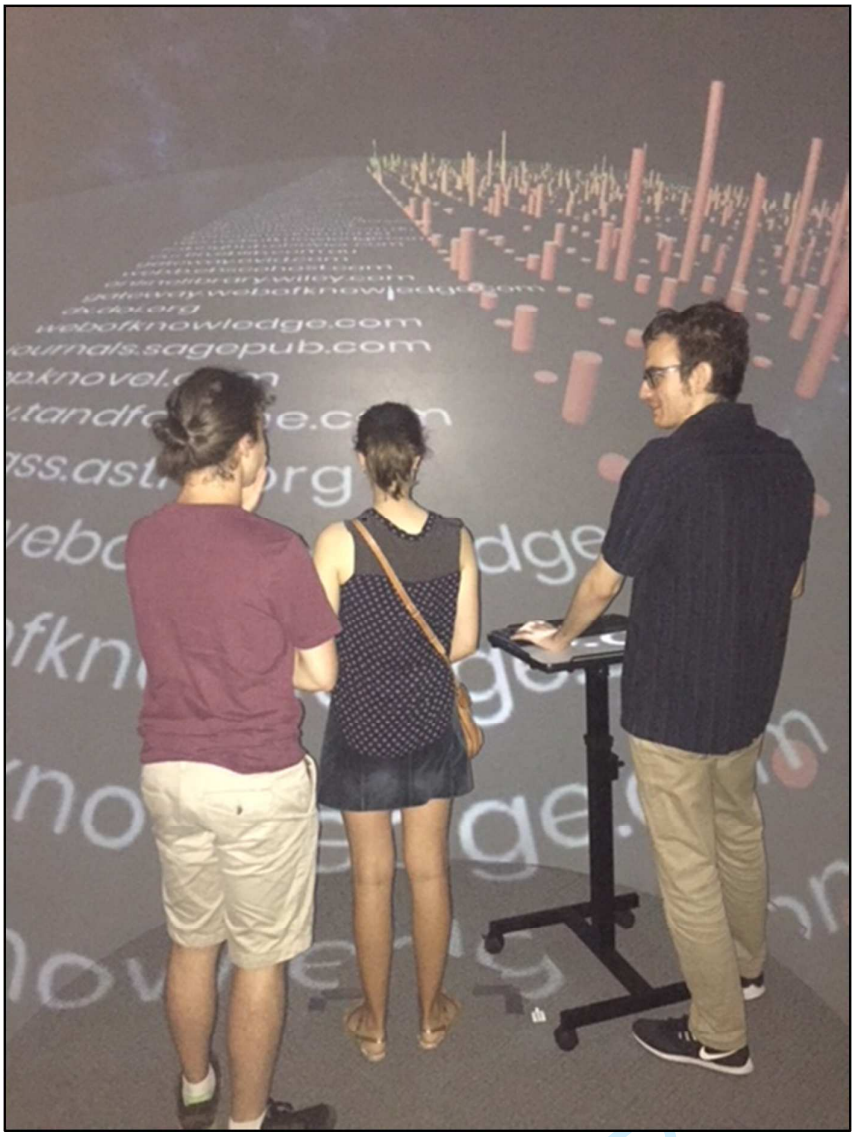

Figure 17: An example of an immersive experience moving though the Database Visualisation display on the HIVE'S dome screen at the HIVE Intern Showcase