LIBRARY ANALYTICS AND METRICS

Using data to drive decisions and services
Every purchase of a Facet book helps to fund CILIP’s advocacy, awareness and accreditation programmes for information professionals.
LIBRARY ANALYTICS AND METRICS

Using data to drive decisions and services

Edited by
Ben Showers

facet publishing
Dedication
To Jennie, Willow and Rowan
# Contents

List of figures and tables ......................................................................................... xi

Contributors ............................................................................................................. xiii

Acknowledgements .............................................................................................. xxiii

Introduction: getting the measure of analytics and metrics ......................... xxv
  Library analytics .................................................................................................... xxv
  The streetlight effect ............................................................................................ xxvi
  Learning analytics ............................................................................................... xxvii
  About this book ................................................................................................... xxviii
  Chapter summaries .............................................................................................. xxviii
  Analytics and metrics: a brief note on definitions ........................................ xxix
  References ........................................................................................................... xxx

1 Library data: big and small ................................................................................. 1
  Chapter overview ................................................................................................. 1
  Defining big and small data .................................................................................. 2
  Small and big data in libraries ............................................................................ 3
  CASE STUDY 1.1 The potential of data to inform personalized
  recommendations at the Open University Library ........................................... 5
  Richard Nurse
  CASE STUDY 1.2 Library ‘big data’: developing a shared analytics service
  for academic libraries ......................................................................................... 14
  Ben Showers
  Chapter conclusion .............................................................................................. 20
  Big and small data: further resources ............................................................... 21
  References ........................................................................................................... 21
2 Data-driven collections management ...............................................23
Chapter overview ........................................................................................23
The collections turn ......................................................................................24
Managing the local collection .....................................................................24
Managing the 'national' collection ...............................................................27
CASE STUDY 2.1 Building an analytics toolkit at the Harvard Library
Kim Dulin and Carli Spina ..........................................................................28
CASE STUDY 2.2 Collection management analytics: the Copac Collection
Management tools project ...........................................................................35
Shirley Cousins and Diana Massam ............................................................35
Chapter conclusion ......................................................................................44
Data-driven collections management: further resources .........................44
References ......................................................................................................45

3 Using data to demonstrate library impact and value .........................47
Chapter overview ......................................................................................47
Does library use have an impact on student success? ............................47
Taking analytics seriously ..........................................................................48
The ethics of impact ....................................................................................50
CASE STUDY 3.1 Library impact data: investigating library use and student attainment (University of Huddersfield) ................................................51
Graham Stone ............................................................................................51
CASE STUDY 3.2 Retention, student success and academic engagement
at Minnesota (University of Minnesota) ....................................................58
Shane Nackruder, Jan Fransen, Kate Peterson and Kristen Mastel ...........58
CASE STUDY 3.3 The Library Cube: revealing the impact of library use on
student performance (University of Wollongong) ....................................66
Brian Cox and Margie Jantti .......................................................................66
Chapter conclusion: from knowing to showing! ....................................74
Library impact and value: further resources .............................................75
Notes ............................................................................................................75
References ......................................................................................................76

4 Going beyond the numbers: using qualitative research to
transform the library user's experience ..................................................79
Chapter overview ......................................................................................79
Qualitative research and the user experience ..........................................79
Qualitative research and emerging user needs ........................................81
A mix of skills and methodologies ...........................................................81
CASE STUDY 4.1 Utilizing qualitative research methods to measure
library effectiveness: developing an engaging library experience ...........82
Lynn Silipigni Connaway, Erin M. Hood and Carrie E. Vass .................82
CASE STUDY 4.2 Ethnographic techniques and new visions for libraries....96
Donna Lanclos
Figures and tables

Figures
1.1 Improving the data in your recommender system ................................................8
1.2 RISE user interface ......................................................................................................9
1.3 A simple relevancy algorithm ....................................................................................10
1.4 Example of the kind of data delivered by the LAMP ‘ugly prototype’:
loans activity for a specific subject, broken down by gender ..........................18
2.1 Initial wireframes for dashboard .................................................................................32
2.2 Modified Harvard library explorer dashboard ..............................................................32
2.3 ‘Bubble’ visualization wireframe for the Harvard library explorer
dashboard ...........................................................................................................................33
2.4 CCM tools results visualization: the graphical presentation of a result set showing the number of documents held by each contributor within the search results ..........................................................37
2.5 Extract from York collection profile: data exported from CCM tools
used to generate locally tailored graphs .................................................................41
3.1 Correlation between electronic resource usage and student grades ......................71
3.2 Correlation between borrowing and student grades ...............................................72
4.1 Emerging groups and human sources ......................................................................88
4.2 Emerging groups and digital sources .......................................................................89
4.3 Educational stages and databases ...........................................................................90
4.4 Educational stages and decision ..............................................................................91
4.5 Educational stages and contact .................................................................................91
4.6 Part of the ground floor in Atkins Library, Spring 2012 (photo by M. McGregor) .................................................................98
4.7 Students working at a whiteboard ............................................................................99
4.8 Student on sofa while another draws on the whiteboard ............................................99
4.9 Student referring to textbook while another uses the whiteboard .......................100
4.10 Whiteboard and large computer screen near the sitting area (Photo by M. McGregor) .......................................................................................................................100
4.11 Flexible study space being used by students (Sketch by M. McGregor) ................101
4.12 Touch-screen tables .................................................................................................102
xii  LIBRARY ANALYTICS AND METRICS

4.13 Group study rooms ................................................................. 102
4.14 Café-style seating space outside the library café ................. 103
4.15 Highly configurable space .................................................... 103
4.16 Zoning use diagram: eating zones ........................................ 104
4.17 Zoning use diagram: study zones .......................................... 105
4.18 Zoning use diagram: talking zones ....................................... 105
4.19 Zoning use diagram: eating and studying zones ................. 106
4.20 Zoning use diagram: studying and talking zones ............... 106
4.21 Zoning use diagram: studying and laptop zones ............... 107
5.1 Web metrics tools and methodologies considered by this case study ...... 121
5.2 ‘Hits’ received by the UK Government Web Archive hosted by the National Archives (National Archives, 2012) ........................................................ 122
5.3 ‘British Museum’ and ‘Terracotta Army’ as seen through Google Trends .... 124
5.4 Network diagram of URL citations between five CHIs ................ 131
5.5 Social network of the UK’s most-visited CHIs ......................... 134

Tables
1.1 Survey results for the value of different types of recommendations ........ 11
3.1 Dimensions of usage .............................................................. 53
3.2 Demographic data examined in Phase 2 .................................... 54
3.3 Discipline variables .................................................................. 55
3.4 Weighted average marks (WAM) for undergraduate students (Faculty of Business) against annual electronic resource usage frequency (hours) ............................................................................. 72
4.1 Educational stages and number of participants in each stage .......... 85
5.1 Visitor numbers and Alexa Traffic Rank for five CHIs ................ 123
5.2 Prominent social media of the five CHIs .................................... 125
5.3 Estimated number of results for organizational names in Google ...... 128
5.4 The number of inlinks and linking domains for the five CHIs according to three different SEO sites ................................................................. 130
5.5 Impact of websites on social media platforms ............................ 132
5.6 Total Twitter followers and centrality of the five CHIs ............... 134
Contributors

Seb Chan
Seb Chan is currently Director of Digital and Emerging Media at the Cooper-Hewitt Smithsonian National Design Museum in New York. He has been responsible for the Museum’s recent digital renewal and its transformation into an interactive, playful new museum, re-opened after a three-year rebuilding and re-imagining. Prior to relocating to New York, he was Head of Digital, Social and Emerging Technologies at the Powerhouse Museum in Sydney, where he led teams responsible for the Powerhouse’s pioneering work throughout the 2000s in open access, mass collaboration, web metrics and digital engagement, as well as large-scale Australian cross-agency projects.

Seb also led a parallel life in electronic art and music organizing, curating festivals and international touring, and was founding editor-in-chief of a long-running music magazine.

Ian Chowcat
Dr Ian Chowcat has over 20 years’ experience in using digital technology in learning and teaching. For many years he was an Associate Lecturer for the Open University, teaching distance courses in humanities and technology, including the entirely online MA in Philosophy. In the late 1990s he helped pioneer the training of OU Arts tutors in online teaching and research. He worked as an e-learning facilitator for Ufi/learndirect and the National College for School Leadership, and from 2005 to 2008 was Director of the South Yorkshire e-Learning Programme, then the largest regional e-learning programme in the EU. Since 2008 he has been Head of Learning Innovation for Sero Consulting Ltd, working on many assignments for Becta, Jisc, and local and regional authorities, covering e-learning, digital inclusion and digital information projects, as well as project managing the establishment of
both of Sheffield’s University Technical Colleges. A former civil servant in the Department for Education, he has a doctorate in Philosophy. He has taught at the University of Sheffield in both the Philosophy department and the management school, and helps to facilitate a MOOC in modern and contemporary American poetry run by the University of Pennsylvania. He lives in Sheffield and is married with one daughter.

**Lynn Silipigni Connaway**

Dr Lynn Silipigni Connaway is a Senior Research Scientist at OCLC Research. She leads the User Behavior and Synthesis Activities at OCLC Research and is the Chair of the ALA ACRL Value of Academic Libraries Committee. Lynn held the Chair of Excellence position at the Departmento de Biblioteconomía y Documentación at the Universidad Carlos III de Madrid and collaborated with the faculty on user-centered research. She was a Visiting Researcher in the Department of Information Studies, University of Sheffield and a Visiting Scholar at the Royal School of Library and Information Science, Copenhagen, Denmark. She has received research funding from the IMLS in the USA and Jisc and the Arts and Humanities Research Council in the UK. She leads OCLC Research in the digital visitors and residents project and is the co-principal investigator of an IMLS-funded project with Rutgers University to investigate the possibility of seamless collaboration between knowledge institutions such as libraries and the Social Q&A (SQA) community. Lynn is the co-author of the fourth and fifth editions of *Basic Research Methods for Librarians* (Libraries Unlimited) and has authored numerous other publications. She is frequently an international and national speaker on the topic of how individuals get their information and engage with technology, and on the assessment and development of user-centred library services. Prior to joining OCLC Research, she was the Vice-President of Research and Library Systems at NetLibrary, the director of the Library and Information Services Department at the University of Denver, and on the faculty of the Library and Informational Science program at the University of Missouri, Columbia.

**Shirley Cousins**

Shirley Cousins is the Copac Service Manager, dealing with the day-to-day running of the Copac service as well as being involved with planning, managing and participating in a range of project activity, including the Copac Collection Management Tools project. Shirley has been at Mimas, the UK National Data Centre at the University of Manchester, since 1995. She has worked on the design and development of the Copac service since its inception, as well as having more recent involvement with Zetoc development.
activity. Her background is in information retrieval, doing research and lecturing in the Department of Information and Library Studies, University of Wales Aberystwyth.

**Brian Cox**

Brian Cox has worked in a wide variety of roles, including privacy, copyright, records management, quality, program management and planning. He started his professional career in privacy, where he worked in both policy and compliance for the Office of the Federal Privacy Commissioner. He then moved to the University of Wollongong (UOW), where he dramatically reduced copyright risk through a mixture of education and audits, underpinned by Copyright Policy and Guidelines. He overhauled the Library’s operational data – transforming it from an unreliable, highly varied, complex and difficult to navigate structure, into highly reliable and easily accessible data. He also played a key role in transforming planning and performance reporting, where through championing project management and a much simplified but more coherent planning and reporting structure, he was able to greatly simplify the Executive’s task of strategic stewardship. Brian first developed his vision for the Library Cube when he started working with quality, where with the support of the Library Executive he worked with the University’s Performance Indicators Unit to transform his vision into reality. Since then, Brian has been seconded to the Peer Learning Unit at UOW, where he made great contributions to improving processes, delivering real and significant ongoing savings for that unit.

**Kim Dulin**

Kim Dulin serves as Associate Director for Collection Development and Digital Initiatives at the Harvard Law Library and Director of the Harvard Library Innovation Lab. She has been an academic law librarian since 1988. In addition to her experience as an academic law librarian, Kim has served as a practising attorney and an adjunct professor of law. Kim has a JD from the University of Iowa College of Law, an MS from the University of Illinois Graduate School of Library and Information Science, and a BA from the University of Iowa.

**Jane Finnis**

Jane Finnis is the Chief Executive of Culture24 (WeAreCulture24.org.uk). She is an experienced strategist, producer and communicator with a deep understanding of audiences, digital tactics and tools. She has worked in the
Jan Fransen is the Service Lead for Researcher and Discovery Systems for University of Minnesota Libraries in the Twin Cities. In that role, she works across Libraries divisions and with campus partners to provide library systems that save researchers’ and students’ time and improve their access to the materials they need to get to their next steps.

Jan’s previous library experience was as an engineering liaison librarian, working with faculty and students in her alma mater, the College of Science and Engineering at the University of Minnesota. Before becoming a librarian in 2008, Jan had a career as a programmer, trainer and writer, with a focus on helping people use technology to be more effective in their jobs.

Erin M. Hood MLIS is a Research Support Specialist at OCLC Research. She received her undergraduate degree in Religious Studies from Otterbein College and her graduate degree in Library and Information Science from Kent State University. She is currently working on the Digital Visitors and Residents study with the University of North Carolina, Charlotte, funded by Jisc. She also regularly contributes to other OCLC Research studies, including Cyber Synergy, the User Behavior Report, and WorldCat Survey Report. She previously worked on the IMLS-funded study, Seeking Synchronicity, in combination with Rutgers, The State University of New Jersey, and the Digital Information Seeker Report.

Margie Jantti

As the Director, Library Services at the University of Wollongong (UOW) Australia, Margie Jantti provides leadership and direction for library services spanning six onshore campus locations and guidance for offshore library partnerships in the UAE and Asia. She is an active member of UOW policy and governance committees. She is the Deputy President of the Council of Australian University Librarians (CAUL), and for a number of years chaired the CAUL Quality Assessment Advisory Committee. She contributes to the CAUL and Council of Australian University Directors of IT (CAUDIT) Leadership Institutes.

Key outcomes at UOW Library for the past five years include: significant restructuring of the organization to extend capacity to support the research
community, the creation of the Library Cube (an enterprise reporting system focusing on the impact of library resource usage and students’ academic performance), and revitalized approaches to organizational performance monitoring and reporting frameworks.

**David Kay**

David Kay is a UK-based consultant at Sero Consulting (www.serohe.co.uk), which he co-founded in 2004 after over 20 years working in systems design for education and library management. David began investigating the potential of activity data and analytics in connection with Higher Education libraries in the Jisc TILE project (2009). He has subsequently been involved in Jisc’s ongoing examination of those concepts in the MOSAIC demonstrator project, in the synthesis of the 2011 Activity Data program (http://activitydata.org) and in the LAMP project (http://jisclamp.mimas.ac.uk/2014/01/so-what-do-we-mean-when-we-say-analytics).

He has co-authored articles in the CETIS series on analytics and the whole institution and on legal and ethical considerations (2012) and an analytics White Paper for Innovative Interfaces in 2013 (www.iii.com/sites/default/files/Innovative_Conversation_Analytics.pdf). In 2014 he presented a keynote speech at the ARL Assessment Conference on the role of the library in unravelling the cat’s cradle of activity data.

**Anra Kennedy**

Anra Kennedy is responsible for the development and delivery of Culture24’s web publishing, content strategy and partnerships. This includes editorial, research and data aggregation services for clients, including the BBC, Europeana and the Centenary Partnership. An ex-teacher, she is also a trustee of Kids in Museums and GEM (Group for Education in Museums).

**Naomi Korn**

Naomi Korn is an IP Consultant specializing in copyright, rights management and licensing for clients from across the UK’s public sector (www.naomikorn.com). She has been running all of ASLIB’s copyright training for a number of years.

Naomi is the Chair of LACA (Libraries and Archives Copyright Alliance) and a founder member of the MuseumsIP Network. She sits on the Executive Committee of the Museums Copyright Group and on the Advisory Panel of the Public Catalogue Foundation. Much of her work includes lobbying and advocacy on behalf of museums, libraries and archives.
Naomi has a prolific publishing portfolio covering an eclectic range of copyright and licensing topics, which includes multiple briefing papers, blogs, tools and templates. She also regularly writes a copyright surgery column for the *Managing Information* journal and other publications.

**Donna Lanclos**

Dr Donna Lanclos has been the Library Ethnographer at the J. Murrey Atkins Library at University of North Carolina (UNC) Charlotte since 2009. Her research interests include the nature of information seeking behaviour, as well as the relationship of the needs of instructors and learners to the physical and virtual spaces of academic libraries, and in higher education generally. She has done research at University College London, as well as in and around the UNC Charlotte campus. Donna’s training is in folklore as well as socio-cultural anthropology, and she earned her PhD in Anthropology from the University of California, Berkeley, in 1999. Her research on the folklore of primary school children in Northern Ireland was published in 2003 by Rutgers University Press as *At Play in Belfast: Folklore and the Formations of Identity in Northern Ireland*. She has published several articles in LIS, IT, folklore and anthropology journals.

**Sejul Malde**

Sejul Malde is Research Manager at Culture24, and works on a range of projects, both UK and European wide, all of which at their core aim to help cultural organizations, and the wider sector, become better at understanding and engaging audiences digitally. In particular he co-leads Culture24’s collaborative action research projects, Let’s Get Real. He has significant experience in both the commercial and cultural sectors, and is interested in exploring opportunities for more meaningful audience engagement through cross sector learning and collaboration.

**Diana Massam**

Diana Massam joined Mimas, the UK National Data Centre at the University of Manchester, in 2007, as a project manager working to deliver new and existing services for libraries, researchers and end users. While contributing to the work of the Copac Collection Management Tools project, she is also currently involved in managing a project researching metadata enhancement techniques and their impact on users, and working with Arthritis Research UK developing content for their website. Before joining Mimas she was a senior library manager at Manchester Metropolitan University for several
years, where she managed the delivery of information literacy services as well as a busy business library. Diana has been a member of the UKSG Education Subcommittee since 2009.

**Kristen Mastel**

Kristen Mastel is an outreach and instruction librarian at the University of Minnesota. She received her Master of Library Science from Indiana University, and her undergraduate Bachelor of Arts in speech communication from the University of Minnesota–Morris. Her research areas of interest include instruction, information literacy, outreach and instructional design. Kristen is Past President of the Minnesota Library Association.

**Shane Nackerud**

Shane Nackerud has worked at the University of Minnesota–Twin Cities for 16 years, currently as Technology Lead for the Library Initiatives. In his current position he is working on finding new ways of integrating library content and open textbooks and resources into courses and curricula, and assisting faculty and researchers publish open scholarly content using library publishing tools and services. Shane’s research interests include library use assessment, libraries and e-learning, resource integration, academic publishing and web design. He received a BA in history from Augustana College in 1994 and his MLS from Indiana University in 1995.

**Richard Nurse**

Richard Nurse is Library Services Manager: Digital Services Development at the Open University Library Services and is responsible for managing library websites, developing funding opportunities and managing the delivery of key strategic library projects. Current projects include the development of the Open University Digital Archive to preserve and showcase material from the collections of the Open University archive, and the Library Futures project, which is carrying out research into student needs of library search, procuring and implementing a next generation library management system, and developing a Digital Skills Passport. Richard has also been involved with the LMS Change project and the Jisc Library Analytics and Metrics Project as a member of the Community Advisory and Planning Group.

Richard has been at the Open University since 2009 and previously spent more than 20 years working in public libraries managing library IT services, bibliographical services and heritage services. His research interests are in activity data, digital libraries, linked data and library technology in general.
**Kate Peterson**

Kate Peterson is the Undergraduate Services Librarian at the University of Minnesota–Twin Cities. She also works with Writing Studies, First Year Writing and Orientation. She has previously worked at California State University–Long Beach, St. Cloud State University and Capella University.

**Mia Ridge**

Dr Mia Ridge’s PhD in digital humanities (Department of History, Open University) focuses on historians and scholarly crowdsourcing. Mia has published and presented widely on her key areas of interest including user experience design, human–computer interaction, open cultural data, audience engagement and crowdsourcing in the cultural heritage sector. While at the Science Museum, Mia held the first ever museum mashup competition, helped the Science Museum’s Centenary Icons poll hit the front page of the BBC News, and organized the release of over 200,000 collections records as open data. Her edited volume, *Crowdsourcing our Cultural Heritage* (Ashgate) was published in 2014. She is Chair of the Museums Computer Group (MCG) and a member of the Executive Council of the Association for Computers and the Humanities (ACH).

**Carli Spina**

Carli Spina is the Emerging Technologies and Research Librarian at the Harvard Law School Library. She has published and presented on a variety of topics related to technology, librarianship and literature. Her interests include the use of technology in libraries, accessibility, user experience and teaching.

**Graham Stone**

Graham Stone is Information Resources Manager at the University of Huddersfield, UK, where he manages the library’s information resources budget, including acquisitions, subscriptions and APC payments. He also manages the University Repository and University Press. Graham has been involved with a number of Jisc projects including the Library Analytics and Metrics Project (LAMP), Huddersfield Open Access Publishing and the OA best practice pathfinder project. He is co-author of OAWAL (Open Access Workflows for Academic Librarians) and TERMS (Techniques in E-Resources Management) with Jill Emery of Portland State University, UKSG Publications Officer and an *Insights* editorial board member. He is undertaking a doctorate at Huddersfield investigating the viability of the University Press as an OA publisher.
David Stuart
Dr David Stuart is a research fellow in the Centre for e-Research, King’s College London, and an honorary research fellow in the Statistical Cybermetrics Research Group, University of Wolverhampton, from where he gained his PhD in webometrics. He is the author of two previous books published by Facet Publishing, Facilitating Access to the Web of Data (2011), and Web Metrics for the Library and Information Professionals (2014).

Elena Villaespesa
Elena Villaespesa works as Digital Analyst at Tate where she is responsible for managing digital analytics, which includes producing website, mobile and social media metrics reports and co-ordinating surveys to inform decision making, with the aim of improving user journeys across different digital platforms. Elena is also undertaking a PhD at the School of Museum Studies at the University of Leicester. Her research is about how museums can measure the impact and value of their social media activities.

Carrie E. Vass
As a Research Assistant at OCLC Research, Carrie E. Vass MA assists Senior Research Scientist Lynn Connaway on the activities: ‘Cyber Synergy: Seeking Sustainability through Collaboration between Virtual Reference and Social Q&A Sites’ and ‘Visitors and Residents: What Motivates Engagement with the Digital Information Environment?’. She also assists Associate Research Scientist Ixchel Faniel on the ‘E-Research and Data: Opportunities for Library Engagement’ activity. She received both her BA and MA in communication studies from the University of North Carolina in Charlotte. Carrie’s personal research interests include issues of power in popular culture, using a critical cultural lens to reveal, oftentimes, oppressive ideologies.
A book is a team sport. This is particularly the case with a book like this one, which relies so much on the examples and case studies provided by experts and practitioners from across the world. The case studies provide not just the in-depth and rich practical material for the book, but have also helped to guide and shape the overall feel of the book. So much of the credit for this book also belongs to the contributors.

I am particularly grateful to my publisher, Facet Publishing, both for the commissioning of this book and the stellar work of my editors, Jenni Hall and, most recently, Damian Mitchell. Not only have they reviewed the chapters as I drafted them, but they have managed to corral and encourage me so that, eventually, I handed in a final manuscript.

I’d also like to thank my colleagues in Jisc and the wider information, library and cultural communities. Their work, ideas and enthusiasm have inspired me in my work and directly led to the idea for this book. There are too many names to mention, but the innovative work taking place in the library and cultural heritage communities at present makes me confident that these communities will not just survive in the new information environment, but actively thrive.

Finally, I want to thank my family: specifically my wife, Jennie, for giving me the space and time I needed to write and for keeping everything going while I was furiously typing; and my children, Willow and Rowan, for providing the occasional interruption and light relief. And also my extended family – my mum, dad and brother – who are always there when needed.
Introduction: getting the measure of analytics and metrics

It seems that almost every aspect of our lives and the world around us is on the cusp of being transformed by the potential that data and the analysis of that data hold for the services and products we use and activities we undertake. Businesses and services are adopting analytics to help drive more informed decisions, to gain a better understanding of their customers and users and to make sense of the ‘big data’ created by all those interactions and actions. The potential appears limitless: from healthcare to education and from government to business.

Similarly, individuals are increasingly using analytics to help improve their performance and understanding of themselves. The ‘quantified self’ captures data from activities as diverse as running and sport, through to sleeping and general well-being. These popular apps and services enable the collection and analysis of data to help improve performance in whatever it is you’re trying to achieve, whether running, sleeping or productivity at work.

The aim of this book is to explore the potential of analytics at an institutional and organizational level: how analytics can unlock a better understanding of your users, inform decision making and help drive new services.

Library analytics

Libraries, along with archives, museums and galleries, find themselves ideally placed to exploit the full potential of analytics.

Libraries, and the cultural sector more generally, have long been familiar with the potential of statistics and data for informing everything from service development to measurement of impact and value (both locally within the institution and nationally – and even internationally). The variety and scope of the data collected and generated by libraries and organizations such as
museums and archives is significant: transactional data on catalogue searches, item check-outs, log-ins to online resources and services, swipes through the entrance gates; manually collected statistics on space usage, student satisfaction, external visitors to the library. The applications of the data are equally varied and overlapping, including management functions (collections development and management, usage statistics), impact (demonstrating value, benchmarking, improving learner outcomes) and improving services and meeting user requirements (recommendation services, collections management/development).

While this diversity in sources and applications is indicative of the importance of data to organizations like libraries, it also highlights the multi-faceted processes and practices for collecting and analysing the data. These practices are often unique to the local institution and its library and reflect both the accessibility of the data in its local systems and the specific uses and types of data that benefit that particular institution and its users. These local variations and challenges would by themselves be sufficient to make this a difficult landscape to traverse, but there are also significant external factors that conspire within the analytics space. Such complications include data access and ownership, formats and standards, privacy and ethical implications.

Maybe more critically, libraries and other institutions are beginning to question exactly what it is that they are measuring in the first place. There is a need to be clear about what is being measured, and why. Otherwise there is a very real risk that our measures become too simplistic or, worse, that we are simply measuring the wrong things: ‘we look away from what we are measuring, and why we are measuring, and fixate on the measuring itself’ (Crease, 2011).

The streetlight effect

Have you heard the parable of the man who lost his car keys? Walking from the office to his car in the dark, he fumbles for the keys to open the car door, but drops them somewhere in the gutter. The light in the gutter is poor and he searches on the pavement, where the light from the street light is brighter and it’s easier to see.

He ends up walking home, unable to find his keys.

The implications of the streetlight effect (Freedman, 2010) are that we often look for answers where it is easiest to find information and data. The result is that we often end up focusing on the information and data that we find, rather than on our original questions. We are so busy searching in the light that we forget what we were looking for, or why it was important.

Much current work in libraries, archives, museums and galleries is looking to address this issue and make sure that we are asking the right questions in the beginning and finding new ways to expose and analyse the data that can
contribute to answering these questions – and, indeed, help to refine and improve the questions themselves.

Much of what will follow in this book is a record of this ‘analytics turn’: a renewed concentration on the questions that we ask and how they evolve as the data we collect forms part of a feedback loop, informing both service developments and the reasons for measuring what’s being measured, and improving the questions that we ask.

The challenges of getting analytics and metrics right are not insignificant, but their benefits to organizations like libraries and other cultural heritage institutions are compelling. This opportunity to begin measuring what really matters, is also one clearly recognized by the library community, but it is not unique to the library. The wider education and academic sectors recognize the importance of the right kind of metrics and analysis as a critical part of the services and systems they use and deliver.

For the libraries, the exploitation of learning and research analytics is likely to be an institutional priority for the foreseeable future.

**Learning analytics**

The potential and opportunities presented by the capture and analysis of data appear boundless. Nowhere is this sense of potential for analytics to transform felt more keenly than in the education sector.

In its 2013 edition the New Media Consortium Horizon Report for Higher Education (NMC, 2013) describes learning analytics as

> [the] field associated with deciphering trends and patterns from educational big data, or huge sets of student-related data, to further the advancement of a personalized, supportive system of higher education.

Put simply, learning analytics is concerned with understanding why some students may not be succeeding, what would contribute to their success and how and when interventions might be helpful. The vision is usually to create a more personalized and effective learning experience for students, and even for researchers. The benefits for learners are substantial, and they provide institutions with the opportunity to improve student satisfaction, as well as to enhance completion and retention rates. These are critical success factors for any academic institution.

Much of the current effort surrounding learning analytics is being put into assembling and organizing the disparate departments and services that might contribute to an institutional learning analytics strategy. This is no small task and represents significant institutional change in most cases. Furthermore, much current discussion is around how learning analytics can move beyond
a simplistic approach to learning, to look at performance beyond the confines of the classroom.

The library has a clear role to play in this larger analytics picture, contributing both its data and analytics experiences and its leadership and expertise, in effectively collecting and analysing data for the benefit of students and in delivering more effective and efficient services.

**About this book**

This book will provide libraries and cultural heritage institutions with an overview of some of the main themes surrounding analytics and the development of metrics. Each of the major themes is accompanied by a series of short, practical case studies describing the development of services or outlining current research and practice in that area. It is hoped that the book will be of use to both managers and library directors in helping them to think about the challenges and implications of analytics in their library or institution, as well as to practitioners who are currently working with analytics or want to learn more. This is ultimately a practical book: you should be able to read the case studies and apply some (or all) of their content to your current role and your library or institution.

**Chapter summaries**

1. **Library data: big and small**
   This chapter explores the definitions of these increasingly popular terms and provides a clear understanding of the differences between them and of the kinds of opportunities that they present to libraries and cultural heritage institutions. While big data captures much of the headlines, it is of little use if we can’t get the ‘small data’ of our systems and services up to scratch.

2. **Data-driven collections management**
   This chapter delves into some of the developments currently taking place in the library sector to exploit the potential of analytics so as to help drive informed decisions about the purchase of materials, usage and collections management and opportunities to extend the impact of the library into new domains.

3. **Using data to demonstrate library impact and value**
   Analytics are increasingly being used to uncover new insights and demonstrate new types of value and impact for libraries and their institutions.
This chapter explores some of the current opportunities that institutions are exploiting through the use of analytics, and the innovative services and tools they are developing.

4 Going beyond the numbers: using qualitative research to transform the library user's experience
While much of the buzz around data and analytics is inevitably about the quantitative ‘big data’, the role of qualitative data in informing decisions is critical. This chapter explores the many ways in which institutions and researchers are capturing this kind of data and the kinds of insights it is providing.

5 Web and social media metrics for the cultural heritage sector
This chapter explores the potential of web analytics for cultural heritage institutions. The increasingly social nature of the web, and in particular the sharing and discovery of content and resources, makes this a critical area for any cultural institution to understand.

6 Understanding and managing the risks of analytics
This chapter explores the legal and ethical risks of analytics and provides best practice and practical examples for how they can be met and managed.

7 Conclusion: towards a data-driven future?
A peek into the future: given the current work and developments that are taking place within cultural heritage institutions and organizations, how might such developments change the cultural landscape over the next five to ten years? What might a data-driven future look like?

Analytics and metrics: a brief note on definitions
Before we go any further it seems worthwhile to pause briefly and explore the two critical terms that will be used throughout this book: analytics and metrics. I will not provide an in-depth discussion of the terms but, rather, make sure that we all begin with a similar understanding of the terms. This is also a useful way to introduce some of the complexities and controversies of the two terms.
Analytics
Analytics is the *discovery* and *communication* of meaningful patterns in data (Wikipedia: en.wikipedia.org/wiki/Analytics). Importantly, analytics is about analysing data to uncover information and knowledge (discovery) and using these insights to make recommendations (communication) for specific actions or interventions. The term ‘actionable insights’ is often used specifically to describe the kind of information that analytics should provide: information that leads directly to an action or actions.

The communication aspect of analytics is often done through visualizations: taking complex patterns of data and representing them in a visually meaningful way that informs specific actions. A good example of a well-known analytics service is Google Analytics, which provides analysis on website data, such as the number of people visiting a site, where they come from geographically and so on, and delivers this analysis via different visualizations on a web dashboard.

Metrics
In the context of this book and of its use in the analysis of data, metrics means the criteria against which something is *measured*. A more formal definition is provided by the Oxford English Dictionary as: ‘a system or standard of measurement; a criterion or set of criteria stated in quantifiable terms’ (OED, 2001).

If you are a salesperson, the number of sales you make in a given period or the amount of positive feedback you receive may be the specific metric against which you are measured. By using that data as a benchmark you can make an informed decision about your performance as a salesperson. Metrics can also be applied to much more complex areas, such as bibliometrics, which seeks to apply mathematical and statistical methods to the analysis of research outputs and literature to measure citations of articles, for example. Bibliometrics is increasingly used to explore the impact of a specific field of research, and even the impact of individual research papers.

As with the streetlight effect described above, there is often an inevitable bias towards the types of data that can be captured rather than the data that *should* be captured. Indeed, the increasingly controversial topic of journal article impacts and how they are measured provides an example of the potential problems that an inappropriate application of metrics can have on education and, in particular, in research.

References
Freedman, H. David (2010) Why Scientific Studies are So Often Wrong: the
Streetlight Effect, Discover, July/August 2010.
CHAPTER 2

Data-driven collections management

Chapter overview
Data-driven collections management is not a new concept for libraries. From reading lists to user recommendations, through to library management systems and the careful analysis of data in spreadsheets and other systems, using data to inform collections management and policy is a key part of curating a library collection. It is therefore of little surprise that libraries and librarians have, over the past few years, become increasingly interested in exploring more sophisticated and joined-up ways of taking advantage of library transaction and management data to help drive more informed and open approaches to decision making on collection management.

This chapter explores some of the most recent and innovative examples of how libraries are refining their collection-management processes by creating tools and applications that can utilize data to make more informed decisions about a wide range of collection-management decisions. Case Study 2.1, from Harvard University Library, explores the creation of a library analytics toolkit and dashboard with a primary focus on collection data. Here the data and visualizations are aimed both at supporting the librarians in their everyday decision making and at enabling users to see how collections have changed over time. Case Study 2.2 describes the work of the Copac Collections Management (CCM) tool and the development of a prototype shared collections management service for UK academic libraries – a service that will enable both local holdings analysis and comparison across other Copac research and specialist libraries in the UK:

- CASE STUDY 2.1 Dulin, K. and Spina, C., Building an analytics toolkit at the Harvard Library (Harvard University), p. 28
- CASE STUDY 2.2 Cousins, S. and Massam, D., Collection management analytics: the Copac Collection Management tools project (Mimas, University of Manchester), p. 35
The collections turn

Libraries increasingly find themselves in a double bind. While budgets are reduced or remain static, user demand for access to the library, its services and content continues to grow. Further compounding this situation are the changing expectations and demands on space from users, meaning that large physical collections need to be rethought and space has to be reconfigured to meet the changing demands of users.

At the same time, e-books and digital monographs present libraries with their own challenges. In the public sector especially, the delivery of e-books continues to be problematic, as the technical and legal issues blight the user experience and compromise the library’s ability to deliver them to users (see the UK government’s independent review of e-lending in England’s public libraries for an insight into these issues – Sieghart, 2013). And, while students and researchers have largely embraced digital access to journal articles, the same is not true for the scholarly monograph. Research suggests (Jisc, RLUK and Ithaka, 2012) that academics require the hard-copy text for long-form reading and in-depth research, especially within the humanities, social sciences and mathematics.

For a long time in academic libraries it has been journals that have dominated discussion of academic print resources, while the book (or monograph) has been a somewhat neglected part of the collections management debate. This inattention is giving way to a sustained focus by libraries, researchers, funders and systems vendors, as there is an increasing realization that library print collections must be carefully and skilfully managed, space must be re-engineered for more social and collaborative uses and physical books must increasingly be seen as part of a larger collection, whether institutional/organizational, regional or even national.

There are the beginnings of a collections turn.

The collection challenges that libraries face, those technological, behavioural and economic examples described above, are the core drivers of this collections turn. Such a turn is defined by its use of data to provide insight into the entirety of the physical and digital collection as well as the space and the changing demands and expectations of users; it is a turn that emphasizes the whole collection. It also utilizes data to extend the overview beyond the local confines of the library and encompass an understanding of holdings elsewhere and of how the collective intelligence can help to inform local collection-management decisions. It is a repositioning of the local in the context of a national, distributed collection.

Managing the local collection

The collections turn that we are beginning to describe begins and ends with the local collection. Libraries must ensure that their collections continue to
provide users with access to relevant resources and to respond to changing demands. The collections turn is primarily a recognition that collections are not built to be great collections in themselves but are there to serve their users – to connect them with the content they need.

Libraries are responding by introducing a number of innovative approaches to local collections management that are reducing the ‘friction’ between the user and the collection, in terms of both more ‘formal’ collections management and acquisition approaches and more playful or ‘informal’ approaches to collections data.

**Patron-driven acquisition**

Patron-driven acquisition (PDA) is an acquisitions approach which is driven by the user. Rather than the library purchasing materials on the user’s behalf, and with a ‘just in case’ approach, PDA enables the user to trigger the purchase of material through the action of clicking on a catalogue link or similar. This ‘just in time’ approach has a number of distinct advantages for the library (for links to further research on the advantages of PDA see Stone et al., 2012):

1. **Cost effectiveness**: Only books or material that users actually want to access are purchased. In theory, the library isn’t purchasing anything until a user clicks on a link to a book that they want to read.
2. **Increased usage and circulation**: Not only is it the case that the material is purchased at the point of need (ensuring that the content is used at least once), but it also tends to be that PDA material has a higher circulation in general. What one user thinks is worth reading tends to agree with what others think is worth reading too.
3. **Collection balance**: Despite the fears, research suggests that PDA and other forms of user-driven acquisition help the development of a balanced collection

(Shen et al., 2011).

These shifting modes of acquisition are still relatively new developments. There are questions about the costs associated with these approaches, but fundamentally they demonstrate the way in which libraries are removing the barriers between the collection and its items and the user. Approaches like PDA enable users to directly affect the collections they use. But the circulation and management data used by libraries to make collections-management decisions, can also have a role in more playful approaches to engaging users with the collections.
Gamifying collections

The ubiquity of mobile devices means that we are comfortable – indeed we expect to be able – to interact between our physical presence and location and online applications and games. Additional layers are progressively being added to our actions and activities, with the aim of improving our experience of and engagement in those activities.

Gamification ‘involves applying game design thinking to non-game applications to make them more fun and engaging’ (http://gamification.org). While this is still a very new way to engage library users, a number of projects and companies are beginning to explore the potential for gamifying the library experience.

One of the best known in the UK is Library Game (http://librarygame.co.uk). Library Game uses library systems data – not necessarily the kind of data we might associate with an innovative, user-centred game – to collect participating users’ activity and transaction data. As a Library Game player borrows or returns a book that data is used by the game to provide a social element to the circulation process, showing what other users are borrowing, etc., as well as awarding points for different activities and moving the user up (or down) the scoreboard. Users can also leave reviews of books, create friends’ lists and see their borrowing history. Importantly, Library Game also provides that data back to the library so that it is able to analyse behaviours and borrowing patterns to inform both its services and collections, utilizing a depth of data not previously available.

Libraries are also utilizing existing applications and services, such as SCVNGR (www.scvngr.com), as a way to encourage students and users to undertake tasks within the library; specifically they are using them for library inductions.

It would be easy to dismiss such approaches as superficial and a passing fad. Yet this is increasingly how we interact with our environment, and these kinds of experiences are still at an early stage of development – they are only going to get better! Critically, they also point to a new way of engaging our users with the collections and services that libraries manage. This engagement is both about making the experience more fun and about using that engagement to generate intelligence and data that can be fed back to the services and collections to continually improve them. These approaches create a positive feedback loop where gradually more in-depth and richer data is generated that is used to further improve and refine services and collections.

These new and emerging approaches to local collections development, strategy and management become even more critical as the data begins to inform collections-management decision making beyond the local institution and library, at a sector, regional and even national level.
Managing the ‘national’ collection

As we have seen, a number of interesting developments are taking place in libraries to help transform local collections management and strategies. And, as library and collections data becomes more accessible, better curated, timelier and more accurate it begins to open up possibilities beyond the local library and enables decisions to be made on a regional or national level. This regional or national organization of collections describes the second part of the collections turn: the local collection is situated within and contributes to a broader regional, sector or national collection (or collections).

As Lorcan Dempsey et al. argue: ‘One important trend is that libraries and the organizations that provide services to them will devote more attention to system-wide organization of collections – whether the “system” is a consortium, a region or a country’ (Dempsey et al., 2013). This focus on the system-wide level is already becoming well established in the UK and elsewhere. The UK Research Reserve (UKRR) is an exemplar of the development of a national and collaborative approach to the challenge of retaining low-use print journals. The collaboration between academic libraries and the British Library enables its members to de-duplicate print journal holdings if the same title is held by three other UKRR members (including a copy in the British Library), ensuring long-term access to journal titles while allowing libraries to free up precious space and resources. In particular, UKRR has been successful in building a fabric of trust between participants. Mechanisms such as formal agreements and the inclusion of the British Library’s Document Supply Service have been key in developing trust. Reliable, comprehensive and timely data has helped libraries to be confident in basing local decisions on what is happening elsewhere nationally.

In the US the Maine Shared Collections Strategy (MSCS) is a collaboration between nine partner institutions working to broaden collection access across the state of Maine. MSCS is interesting for a number of reasons. Firstly, the strategy is aiming to focus on access to collections, preservation and resource sharing; it is not explicitly about disposal or de-duplication of materials in the same way that UKRR is. Secondly, the strategy also includes books, which until recently have largely been excluded from national or regional approaches to collections management. MSCS is beginning to address some of the challenges that the monograph presents to shared collections strategies, including being able to offer ‘print on demand’ services in relation to the millions of digital books available within the Hathi Trust collection. This approach may provide a way of exploiting digital copies while serving an ongoing preference for print, expressed by both researchers and students (Kay, Stephens and DeNoyer, 2014).

This increasing interest in the challenges that the transition from print to digital monographs presents to libraries and institutions is something that is
being explored as part of the National Monograph Strategy (NMS) in the UK. The project is exploring the potential for a national approach to the creation, collection, preservation and digitization of scholarly monographs. The project has outlined eight high-level ideas to address some of the challenges presented by the scholarly monograph, including a national monograph knowledge base, which would provide a comprehensive and open bibliographic and holdings database enabling the development of new applications and services for libraries, systems vendors, publishers and users. Much of what will underpin a national strategy for monographs in the UK will be based on accurate and timely data that can help to inform and drive decision making, both locally and system-wide. Further information on the NMS ideas and the strategy itself can be found on the project’s webpage: www.jisc.ac.uk/research/projects/national-monograph-strategy.

These examples of regional and national collection management each provide exemplars of how data is being used to help drive decision making at that system-wide level. This picture maps onto the two case studies below, which describe the innovative use of data locally to drive collections-management decisions and improve the user experience, and how local data can be aggregated to provide institutions with a national picture, helping inform local decisions in the context of a national collections landscape.

The two case studies also make it clear, however, that the data on its own is not sufficient. The data, intelligence and systems that underpin these services and strategies are tools that enable librarians to ask new questions – but do not necessarily provide easy answers. They enable libraries to explore new opportunities and librarians to develop their skills and knowledge so as to curate and manage collections that meet their users’ needs, as well as manage to balance the economic and technological changes that are also driving change.

Data-driven collections management can be transformative, as these case studies begin to describe, but it can achieve this only when the right skills and expertise are on the ground to ensure that data-driven insights can be turned into concrete actions.

**CASE STUDY 2.1**

**Building an analytics toolkit at the Harvard Library**

*Kim Dulin (Harvard Library Innovation Lab) and Carli Spina (Harvard Library)*

**Introduction**

In these days of data-driven decision making, libraries are starting to look for ways to use the data that they already collect to better serve their users while
maximizing their budget’s reach. At the Harvard Library Innovation Lab (http://librarylab.law.harvard.edu), part of the Harvard Law School Library, these two goals served as the jumping-off point for developing Haystacks, a library analytics toolkit. Haystacks visualizes library data in order to help provide insights into library activities, better inform library users and enable library staff to make smarter decisions in their day-to-day work. Over the last three years the Haystacks development team has conducted extensive research into data-collection practices used at Harvard University Libraries: the resulting tool is designed to search through and visualize collection data for over 12 million items in the Harvard Library’s collection. This case study describes the process we followed in creating the tool and offers some lessons learned for others who may be considering similar projects.

**Background research and project preparation**

The original inspiration for the analytics tool came from an earlier project called Checkout the Checkouts. Checkout the Checkouts generated a web page with information on the most frequently circulated items in the Harvard Library collection. Results could be further subdivided by the user’s school affiliation (i.e. Law School, Divinity School, Graduate School of Design, etc.). While this provided a nice snapshot of a specific type of library data, the team quickly realized that the tool would be much more powerful if it could incorporate other library data into a format that was similarly easy to navigate.

This idea grew into a proposal to build a library analytics toolkit. The proposal was submitted to the Harvard Library Lab, an on-campus organization that awards grants from the Arcadia Fund (www.arcadiafund.org.uk/) to support projects that ‘create better services for Harvard students and faculty and play a leading role in shaping the information society of the future’ (Harvard Library Lab Program Description and Guidelines, 2012). The initial proposal envisioned an open-source software package that would allow libraries to track and visualize various types of data, such as electronic resource usage and reference transaction statistics. The Library administration was particularly interested in a tool that would help it to track how resource expenditures translated to use.

The proposal was accepted and initial funding was granted, with the understanding that the first stage of the project would involve extensive interviews with staff across the Harvard Libraries to determine how statistics were collected and what information would be most useful. This stage would ensure that the product developed would integrate well with how that data was already being collected and used. The team collected information about how data regarding library space utilization, reference interactions,
acquisitions, circulation and electronic resource usage was gathered. This research revealed that various campus libraries and even different departments within the same library collected data in different ways. These differences often related to the varying types of services offered across the campus and the unique purposes for which each department collected data. We also found that acquisitions and collection data from various publishers often differed in format and availability, particularly for electronic resources. While the Counting Online Usage of Networked Electronic Resources (COUNTER) standards have helped to make recent usage information more compatible across electronic resources, we still found this data difficult to integrate with other collection and usage information, particularly for time periods prior to the advent of the COUNTER standards.

During the interview phase we looked beyond Harvard at how other libraries, archives, museums and similar institutions were collecting, using and sharing their own data. The work of other institutions influenced some of our early ideas about how the toolkit might be used. We looked at several existing web dashboards that appealed to us and seemed to match our needs, including North Carolina State University Library’s Dataviews Dashboard (now known as By the Numbers, www.lib.ncsu.edu/bythenumbers), the Indianapolis Museum of Art’s Dashboard (http://dashboard.imamuseum.org) and one then in development at the Brown University Library. These projects helped us to determine how data might be visualized so that it would be both useful for library staff and informative for library users. With this information, we focused on designing the toolkit as a dashboard.

This research stage culminated in a report (Spina, 2011) that laid out the goals for the project and identified which types of data might best integrate into the initial iteration of the toolkit. This document advised that the toolkit should take advantage of the data already available at the Harvard Libraries and that the design should be extensible so as to encompass additional data in later iterations. With this in mind, we focused initially on visualizing collection data, so that collection development librarians could immediately use the tool and so that library users could have a window into how the use of Harvard Library’s collections has changed over time.

Choosing a developer
From the beginning we knew that we wanted a tool to visualize multiple types of library data in a useful and somewhat unusual format. We did not intend the tool to be used only by librarians; rather, we wanted it to be useful to general library users interested in discovering not only what items were collected by the Harvard Library but also how they were used. This objective dovetailed with the Library Innovation Lab’s focus on user metrics (see for

We aimed for the dashboard design to display relevant metrics in a modular format and respond to user input. After reviewing the dashboard format and thinking about our users’ needs, we decided that we wanted to allow our users to dig more deeply into the data and manipulate it in ways that the sample dashboards did not allow. Based on our knowledge of advancements in data visualization, these features seemed possible.

To achieve this advanced data visualization, we decided that we needed to hire an outside specialist. We interviewed local data visualization companies, and found Rosten Woo, who had previously worked with our designer, to be the best fit (www.wehavenoart.net). Woo had never worked with library data but he was intrigued by the project and was interested in libraries in general. Once Woo had been briefed on the project he engaged the services of two colleagues, Sha Hwang and Rachel Binx, both data visualization specialists (http://postarchitectural.com, http://rachelbinx.com).

To help the developers understand the needs of the library community, Harvard librarians crafted use cases from which the developers could design the first iteration of the toolkit. Our primary use case featured a collection development librarian interested in seeing how the library’s collection had developed over time. Other use cases included a librarian interested in developing a research guide on a particular topic, a library administrator planning a weeding project based on circulation data and a library user interested in tracing the history of research on a topic by seeing when books were published. These use cases provided the foundation for development decisions and proved particularly useful in targeting future revisions.

We collaborated with the developers mainly via telephone or Skype meetings every few weeks. Over time the tool was modified extensively, and the final version appeared strikingly different than previous ones. Figure 2.1 below shows a very early version with modular components that look somewhat like a dashboard. Ultimately, we decided that this version did not synthesize and display data in the manner that we were hoping for, but it was useful to demonstrate the concept and to show how the available data could be manipulated.

The next version of the toolkit, shown in Figure 2.2, completely revised the way that the data was presented to make it easier to dive deeper into the information. Users were able to search by date and view the data retrieved in a graphical format. They could begin their search from within the Library of Congress Subject Heading (LCSH) hierarchy embedded in the tool. Search results could be instantly manipulated by clicking on the graphical interface. The current iteration, shown in Figure 2.3, looks entirely different from
previous ones. It reflects a wholesale revision based on information from the use cases and the additional data points, plus significant modifications in search and display. This version switched the visualization style to a bubble format, pulled individual item data into the display and deployed new navigational features as well as the ability to select and export data.

**Technical challenges**

The Innovation Lab staff and Woo’s team shared the work on the toolkit. We received supplemental funding from the Harvard Library Lab to add features after the first version was launched.
Library Innovation Lab in-house developer, Paul Deschner, did the back-end data-wrangling work and developed the API that allowed the data visualization experts to pull in Harvard data. Deschner’s work required him to gather data in three areas: e-download statistics, hierarchically structured Library of Congress (LOC) classification data, and LOC call-number supplementation for non-LOC-classed items. E-download statistics presented the greatest challenges, and automated solutions had to be developed so as to handle large amounts of source data; the data, in turn, was at times loosely structured and spread across ad hoc collected files with many different formats and data structures. The LOC classification data, which was bureaucratically difficult to obtain from LOC, needed to be restructured along hierarchical lines, and corrupt source data had to be fixed. LOC call-number supplementation required designing routines to match LCSH profiles and assign call numbers to items without them. Despite these challenges, Deschner described the work as rewarding, since all three solutions resulted in major new additions to our open data platform, which allowed easy access for developers who might want to build on the application in the future.

New challenges arose once the data was fed to Woo and his team. The massive data set needed to be scaled to an individual level that made sense for the use cases. There were many individual items to track (over 12 million objects, each with 10–30 fields) and the data was most interesting when the user was able to view it at the item level (e.g. it is probably more interesting to get the titles of books about ‘cacti’ than to know how many books about ‘cacti’ there are). In Woo’s previous work, it was not essential to drill down to such a granular level. Woo explained, ‘If you were doing population statistics, it’s enough to know a few general things about 12 million people (race, income, etc.) but here it’s as if we want to be able to drill down to know each of those 12 million people’s first name.’

A second challenge involved the peculiar nature of the LCSH, which are
arranged in a somewhat outdated taxonomical structure. Woo noted that it was quite a bit of work to make sense of this giant, sprawling, semi-proprietary taxonomy. The developers wanted to be faithful to the spirit of the actual library, in which 'each item gets one location', but also wanted to provide the flexibility that web users are used to, such as mixing and matching search terms of their own choosing.

Woo also found that the unique nature of library materials presented a steeper learning curve than he had faced in other projects. In every project a developer needs to understand the data and to familiarize himself with the conventions and nuances of a particular dataset. Library datasets have a large range of differences, and different disciplines use the materials in different ways. Resources in one discipline may have checkouts several orders of magnitude larger than those in other disciplines, but that does not necessarily mean that they are that much more important.

Conclusions, next steps and recommendations

While the toolkit is now in a form that achieves many of our goals, it remains a work in progress. In its current format, we anticipate that it will give library staff a view of how the collection is being used that was not available with previously existing tools. It will also help collection development librarians to identify trends in both acquisitions and use that will make it easier to predict and meet user needs. While we have primarily focused on creating a tool for librarians, we also hope that the toolkit will be a useful discovery tool for library users. It offers a new window into our collection and makes it easy to navigate through the holdings on a subject in a visual way. It also offers the public greater insight than ever before into how the Harvard Library collection is used.

Despite this success, there are still areas for improvement. In creating the tool, we worked with various sorts of data, including MARC records, LOC class numbers and LCSH, and COUNTER statistics, but we had hoped to include additional data, such as financial information to show dollars spent versus usage. This ultimately proved to be unworkable because of Harvard’s complex library structure and the desire to keep financial issues private. We ultimately determined that if we wanted to include this data in the future we could develop a spin-off version to be kept on a secure internal server.

We expect to launch Haystacks to the Harvard Library community in the next few months. Once it has been officially launched, we plan to collect feedback and conduct usability testing to determine whether it meets the needs of the various target audiences. This process will give us the information we need to continue to modify and add to the existing tool. The Harvard Library Lab, which funded the work on the toolkit, will also fund
documentation, public distribution of code, future hosting and outreach. As with all of our projects, we will offer the code under an open source licence. We hope that other libraries will use our existing project as a starting place for their own data visualization projects, or will at least find inspiration in the way that we have combined our data to offer new insights into our collection.

CASE STUDY 2.2

Collection management analytics: the Copac Collection Management tools project

Shirley Cousins and Diana Massam (Mimas, University of Manchester)

The Copac Collection Management tools project (CCM tools) aims to deliver a collection management support service, providing tools that enable library staff to make more informed decisions around their collection-related activity, such as material disposal, conservation or digitization, collection assessment and development.

CCM tools is a Jisc-funded project which began in 2011 with a Research Libraries UK (RLUK, www.rluk.ac.uk) initiated collaboration between the White Rose Consortium libraries (the Universities of Leeds, Sheffield and York) and the Copac team at Mimas (www.mimas.ac.uk). We have worked closely with RLUK throughout and members of the RLUK community have played a key role, with most recently the libraries of the University of Manchester, the Royal College of Surgeons of England and Senate House Library, University of London, joining the project to take the work forward.

Copac brings together the catalogues of a growing range of UK and Irish research libraries, creating an increasingly valuable picture of library collections. We know from user surveys that some library staff make use of the service to assist their collection-management decision making; however, Copac is designed primarily as a public catalogue service, so it has limitations in this context. At Mimas we are enthusiastic about the potential for making the Copac data work harder, increasing the benefits to our contributing libraries and, more widely, to the library community. So, within the CCM tools project we have developed a new interface to the Copac data that provides library staff with a range of facilities to support exploration of the data at the collection level.

By enabling decisions to be made within the context of comparative analytics, the CCM tools support libraries wanting to benchmark specific collections within the wider national collection landscape and identify key areas of commonality or difference against other collections locally or
nationwide. The project has been driven by the collection management needs and the enthusiasm of the partner libraries; this has been a crucial element in the success of the project so far, grounding the technical development within the real requirements of library staff, with Mimas providing technical expertise to translate those requirements into practical reality. The initial interface-development activity was carried out alongside use-case development by the libraries, with ongoing iterative technical development as the partners worked on case studies. In this way the tools were formed to support the range of activities reflected in the early use cases, and honed through practical use.

The CCM tools pilot

The current CCM tools interface allows users to gain a valuable insight into the collections of our contributing libraries, with options for exploring collections broadly between institutions, or looking in detail at individual materials.

There are three search screens supporting:

- manual searching for small numbers of documents or individual items
- batch searching for large numbers of documents, e.g. for reviewing collections using a file of local record numbers exported from a library’s local catalogue
- keyword searching for collections in particular subject areas, by specific authors, etc.

Searches can include library or geographic limits, for example a limit to a particular region can improve understanding of a collection within the local context. Searches can also be tailored to show items held by a specified number of libraries, for example, to show just the commonly held items to assist making withdrawal decisions. It is also possible to apply varying degrees of results de-duplication in situations where the caution shown in the creation of Copac-consolidated records is not required.

Search results can be viewed as Copac records, but the main focus of the results display is the data visualization and the export options. Records loaded into Copac are de-duplicated, with matching records being formed into consolidated records that represent (as far as possible) all our contributors’ holdings of a particular document. This means that the Copac records contain valuable information about the number of libraries holding a copy of each document in the database. This holdings data underpins the data visualizations that allow the searcher to assess their own holdings within the search results in relation to those of other contributors. Figure 2.4 illustrates
the results of a search, expressed as a graph that shows the number of documents within the search results that are held by each contributor. The search results can also be exported in different forms for further local analysis.

In 2013 we released a beta version of the CCM tools to a wider audience comprising the RLUK member libraries, all of whom contribute their catalogues to Copac. This is allowing for testing by a much wider community and providing valuable feedback on the current functionality, as well as looking at new use cases within these libraries. Alongside this we are developing support materials and making provision for community support mechanisms.

The project is still in a pilot phase, but has been very well received, to the extent that some of the participating libraries are already beginning to use the CCM tools as part of their business-as-usual activity. Some of the range of activity being explored can be illustrated in the case studies our partner libraries have been undertaking.

### Case studies

We have collected a range of case study reports from users which demonstrate how the CCM tools can deliver added value and efficiency across a broad range of collection-management activities. These case studies are available on the CCM Blog and also via the CCM User Forum (http://copac.ac.uk/innovations/collections-management). Outputs from the tools enable library staff to take an analytical approach to making collection-management decisions which also have wider implications across a range of library processes and policy areas.

---

**Figure 2.4** CCM tools results visualization: the graphical presentation of a result set showing the number of documents held by each contributor within the search results
Participating libraries have used the tools in a variety of contexts to facilitate stock withdrawal, collection profiling and conservation or digitization work. The added value of the analytics that the tools provide has been demonstrated in many ways.

**Stock management**

Supporting decisions about stock withdrawal within the context of increasing pressure on space is a key benefit of the CCM tools. Despite the increasing availability and adoption of electronic materials, most libraries are under significant space pressure, in part due to the continuing acquisition of print monographs, but also through the need to support newly developing forms of teaching and research within the library. So this fairly clear-cut task is often the first area of experimentation for libraries embarking on their own use of the tools, providing clear statistical evidence and time-saving benefits.

By allowing libraries to check their own stock against the holdings of other libraries, and identify those titles which are rare or unique nationally, the tools provide a benchmark against which stock withdrawal or weeding decisions can be made, often in conjunction with data from other sources, such as usage statistics. Stock management tends to be a key driver in encouraging participants to try out the tools and we have reports of projects ranging in size from a brief assessment of 84 items in a library store to a major project resulting in the withdrawal of over 2,500 low-use monographs.

A good example of this activity is the work done at Sheffield (Ward, 2013) to withdraw low-use monographs from a document store so as to provide space for materials being relegated from elsewhere in the library. Sheffield was concerned that document withdrawal, necessary for local reasons, should not result in the loss of materials not widely held elsewhere. Based on local withdrawal criteria, a workflow was developed that used a list of ISBNs generated via the library management system to batch search CCM tools. From the results, a minimum cut-off of eight holding libraries was defined to permit withdrawal of an item. The data exported from CCM tools was then re-imported into Sheffield’s library management system to generate a pick-list for withdrawal of items from the shelves. This allowed 29 metres of shelving to be freed up, but with the assurance of retaining items that are less widely held.

The Sheffield work also illustrates how, by using the tools, libraries have been able to design far more automated workflows for stock-withdrawal activities than would otherwise have been possible. In other cases libraries report that they had previously performed manual stock checks on Copac when initiating stock weeds. This time-consuming task can be eliminated
when it is replaced with a batch item search using the CCM tools. Significant savings in staff time are identified in many of our case studies which can be quantified as direct cost savings to institutions. For example, staff at the University of Manchester established a workflow using the CCM tools to identify and process stock for withdrawal which took staff four to five days to complete. The team calculated that manually checking the same stock prior to withdrawal would have taken the same staff seven weeks to complete (Copac, 2012a).

The use of comparative statistical data to support stock retention decisions also enables managers to demonstrate clearly the rationale for decision making. This has had a positive impact on the level of support for such potentially controversial activity from users and academics, who can be reassured about continuing access at a local or national level. Not only do the graph and map visualizations help to clarify collection-related issues for stakeholders, but they also help the library staff involved to present a much more ‘professional’ case for the decisions they are taking.

We also have examples of libraries using the tools to support decisions regarding conservation and digitization, taking into account the rarity of individual items. The tools enable staff effort to be targeted on items which are identified as having maximum benefit to the collection, and therefore increase efficiency in this time-consuming area.

However, while facilitating decision making, the provision of data about other library holdings also flagged up the need for libraries to address their existing policies on stock retention and preservation. Decisions need to be made about the threshold for considering an item to be rare or unique, i.e. how many copies of an item held elsewhere make it a rare item? Even when they are identified as rare or unique, what are the policies for retaining such items? There are differences in approach even among the examples we have collected, including where regional policies may exist. These decisions also have a national dimension when considering the broader concept of the development of national monograph collection policies.

Collection profiling
Going beyond support for stock management activity outlined above, the participating libraries have also been keen to exploit the potential of the CCM tools in collection profiling and benchmarking. Various case studies have demonstrated the value of the tools in enabling libraries to deepen their understanding of their collections both individually and in relation to other libraries.

We have had reports posted to the CCM Forum describing activity around the profiling of specific subject areas, confirming areas of collection strength,
discovering previously hidden areas of significance, enhancing special collections and identifying heritage collections. All these case studies reflect the tools’ unique functionality in providing direct comparisons to other collections in an easy-to-use format.

In an environment within which institutions increasingly have access to the same electronic resources, many universities are recognizing that areas of their print collections are prime assets to be conserved and promoted as unique resources, enhancing the reputation of the institution. The CCM tools can help libraries to identify those areas of their collections which have regional or national significance, providing opportunities for marketing the library, as well as supporting applications for funding to conserve or develop materials in those areas. Previously, recognition might have come about through collecting expert knowledge and opinion from library and academic staff: a necessarily subjective approach, and one that entails the danger of losing knowledge as specialist staff move or retire. The use of comparative data from the tools enables what one library has described as ‘a sounder and evidence-based approach’ to collection analysis (Emly, Horne and Pindar, 2012).

The ability to collect clear statistical evidence is frequently cited as a key benefit in collection-profiling work. Such evidence can be used to confirm previously anecdotal or instinctive ‘hunches’ about a library’s collection strengths. For example, the University of St Andrews reported that:

The visual representation of the data, both the graphs and map, are instantly illuminating and supported our theory.

(Helen Faulds, Deputy Collections Manager, University of St Andrews Library)

**Combining data to best effect**

As familiarity with the tools has increased, so has the complexity of the projects which some participants have completed. A common feature of these projects is the additional value to be obtained by combining library analytics data from a range of sources with contextual collection data. This encompasses how the collection fits within the national context, as well as against specific competing or high-profile institutions, producing a detailed profile of a subject area within the library.

This approach is exemplified by work done at the University of York (Elder, 2013) on building up a detailed collection profile in one specific area by bringing together a variety of statistics from different sources; these related to features of the collection such as usage, age profile, stock acquisition rate), etc. York performed a gap analysis, identifying stock held in common by the top-listed non-copyright libraries, which are therefore shown to have the
strongest collections in a specific subject area. This information was subsequently used to produce a list of titles for potential purchase to strengthen the coverage for a newly developing subject area at the University of York (Figure 2.5).

The initial collection profile report was very well received, being seen as a valuable tool for working with academic departments in developing collections within their subject and generating interest from other subject liaison staff wanting to do similar work in their own area. This activity is now embedded as part of the York ‘5 Year Content Strategy’ and the reports are encouraging and facilitating discussions between the library and academic departments in regard to evaluating stock selection and use, thus contributing to more positive collaborative relationships.

Other potential uses for complex collection profiles include providing support for funding bids or identifying collaborators for joint venture, as suggested by participants using the tools in increasingly creative ways.

**Going forward**

As the participants have gained familiarity with the tools and the CCM tools project has progressed, they have been finding new ways to explore the collections and developing new use cases as they gain a sense of what is possible. This has not only provided concrete information about collections, but has enabled the participants to carry out trial collection research in areas that were previously too costly, or simply not possible. This increasing understanding of the tools’ potential has also broadened the focus of interest

![Figure 2.5](image-url)
from relatively clear-cut issues, such as withdrawals, into diverse areas: from improved stakeholder relationships, through library marketing, to identifying potential collaborators in collection-development activity or grant applications. To take the CCM tools into full service, further work is needed in a number of areas. Technical development is ongoing; the collection analytics are dependent on the quality of the record consolidation within Copac, which in turn is dependent on the underlying record quality. At the time of writing we are developing a new Copac database with, among other things, enhanced record de-duplication; but this will never be perfect and it is important to understand the limitations of the data. However, the match process errs on the side of caution, and so it will tend to over-emphasize rarity, which is reassuring in terms of withdrawal decisions, while the CCM tools de-duplication facilities provide users with control over the degree of de-duplication desired for a particular search.

Alongside this, another future activity is the gathering of further evidence for a business case as we work towards moving from a pilot project towards a full service. An important next step is to explore use cases and requirements for the wider library community, including libraries that are not contributors to Copac. We have had interest from a range of different libraries outside the pilot group and we need to look at how we might best support this broader community. Many of the existing facilities will work for any library, for example the ISBN batch search or the export of data for local evaluation; however, we will be working with non-contributing libraries to ensure that we understand their requirements fully and provide them with the most effective support.

Even at this early stage users can see significant benefits in the type of collection data the tools can supply. In our recent survey of trial users (the results are available on the CCM tools blog http://copac.ac.uk/innovations/collections-management) there were comments across areas such as ‘speed’, ‘benchmarking’, and support for withdrawals. Other comments related to the importance of having concrete data to go beyond the ‘anecdotal’ understanding, in particular for working with other stakeholders. One user valued

Increasing the amount of data and knowledge about the library collection that I can report on to senior management.

Another commented that

Understanding the profile of our collections in comparison to other UK collections has been politically helpful.

The CCM tools will not provide ‘the answer’ to everyone’s collection-
management questions, but this type of data can provide valuable supporting information. And, as the York case study illustrates, collection analytics can be used effectively alongside data from a range of other sources to provide a compelling case, which could be adapted in many different situations.

Alongside the development of practical collection-management support tools, there has been a second thread running throughout the CCM project. This has involved a dialogue with the community exploring shared approaches to collection-management activity. The CCM tools initially grew out of a concern for the managed withdrawal of materials as libraries seek to free up space. In many research libraries checking for ‘last copies’ is an integral but time-consuming part of the stock-editing process. The CCM tools can assist with the practicalities of the process, in terms of the speed and ease of gathering data for informed decision making. However, the provision of fast and easy access to new collection analytics places into sharp relief the need to carry out this collection-management activity within a broader context.

Within the research libraries, the issue of stock withdrawal raises important questions about the extent to which libraries should consider the needs of the wider research community beyond the individual institution. There is a growing sense of the importance of this wider context and the need to ensure that the broader regional and national document collection is not diminished by local decisions about collection disposal. That which is redundant in one library may prove to be rare and valuable more widely. This is an issue that the project has explored through community consultation, where there was an emerging consensus over the need for collaboration to ensure that researchers continue to have access to the full range of materials through a distributed National Research Collection of Monographs (Copac, 2012b). More recently, project members have been involved in the Jisc National Monographs Strategy project and the development of the project’s strategy roadmap (Showers, 2014) exploring problem areas across the monograph landscape. In the long term we hope that Copac and the CCM tools can offer valuable collection analytics to facilitate local collection-management activity across a range of areas, as well as being a focus for a regional and national approach to sustainable collection management, supporting collaboration between institutions and sharing of materials, protecting the UK research collection into the future.

**Acknowledgements**

A major factor in our success so far has been a highly engaged, supportive and effective Project Board. We are indebted to Mike Mertens from RLUK; Brian Clifford and Michael Emly from the University of Leeds; Gary Ward from the University of Sheffield; Ruth Elder from the University of York;
Chapter conclusion

This chapter has described the ways in which data is helping to inform decision making by libraries both locally (which, in the case of Harvard, is a multi-library institution) and within a region, sector or nation, as in the case of the CCM tools. What both case studies describe is the increasing convergence of these two types of activity: increasingly, the decisions that libraries make at a local level will inform and be informed by those events and decisions happening across libraries at a sectoral and national level. Any distinction between the local library’s data and that of the consortium, region or sector will erode. While this will inevitably present libraries and institutions with challenges – as innovations such as patron-driven acquisition have done – it will also provide libraries and the users they serve with exciting new opportunities. The idea of a national collection becomes a very real possibility.

However, we must be careful not to become too complacent in the ‘collections turn’ we are describing here. While analytics provides us with a clear picture of what people have done (i.e. borrowed a particular book), it doesn’t provide us with any data on how they did it or what the experience was like. If libraries and cultural heritage institutions really want to begin transforming their services, and the experience of their users, new forms of data and insight will need to be explored. This exploration of new types of insight will be examined in our next chapter, on demonstrating the impact and value of the library through analytics.

Data-driven collections management: further resources

If you’d like to find out more about the work described in this chapter, and access further reading and inspiration, below are additional resources for individuals and institutions interested in how data and new data-driven approaches can inform collections management and development.

National/regional shared collections strategies and services

- The National Monograph Strategy includes a literature/landscape review that covers a lot of relevant material around collections management and national/international activity in this space, http://monographs.jiscinvolve.org/wp/2013/07/31/monographs-landscape-report/.
• More information on the Maine Shared Collections Strategy (MSCS) can be found at its website, www.maineinfonet.org/mscs.
• More information on the UK Research Reserve (UKRR) can be found at its website, www.ukrr.ac.uk.
• The Hathi Trust, www.hathitrust.org.

Library game examples
• Library Game, http://librarygame.co.uk

References

Case Study 2.1
Case Study 2.2

Copac (2012a) Case Studies Synthesis Report,

Copac (2012b) CCM Tools Project, Retention and Preservation: final report, July 2012,
http://copac.ac.uk/innovations/collections-management/reports.

Elder, R. (2013) Use of the CCM Tool for Gap Analysis at the University of York Library,

http://copac.ac.uk/innovations/collections-management/case-studies.

www.jisc.ac.uk/reports/a-national-monograph-strategy-roadmap.

Ward, G. (2013) Stock Withdrawal Workflow, Case Study 1: Sheffield University, CCM Tools Beta Trial: introductory event, London and Manchester, July 2013,