Google Books user needs research

Dr Megan Gooch and Dr Frankie Wilson, May 2024

Contents

Summary

The Bodleian Libraries is embarking on a second phase of digitisation as part of the Google Books Library Project. To inform decisions about hosting platform, this research investigated user needs and behaviours regarding Google Books, alongside a review of usage data and information about the platform approaches taken by other Library Partners.

The focus groups (20 participants from Oxford and 6 from Harvard) illuminated five distinct use cases for Google Books in a Higher education context.

Close Readers primarily utilise physical books but use Google Books as a proxy to quickly checking references without disrupting their writing 'flow', or previewing content before obtaining the physical copy. They prioritise faithful reproduction of the physical book's layout and page numbers, and need items to be in SOLO (library 'catalogue') if they are to come across its existence within their workflow.

Digital Discoverers prefer online books for convenience and accessibility. They prioritise ease of access via their normal workflow (SOLO) and require faithful reproduction of physical books for referencing alongside ebook functionalities like highlighting and annotating.

Specialist Data Miners are proficient in text and data mining, prioritizing access to vast volumes of data from various sources. They require detailed metadata for understanding data provenance. Accepting imperfections in data quality, they prefer XML or JSON formats and expect tools for handling multimedia content. Reproducibility and citability are paramount for their research.

Data Extractors share similarities with Specialist Data Miners but lack advanced technical skills, relying on provided tools for text and data mining. They prioritise transparency in data provenance and tool functionality, seeking ease of use, speed of downloads, and features like named entity recognition. They require access to underlying data and metadata via user-friendly tools.

Picture Pickers seek high-quality images for teaching, presentations, or publications. They distinguish this activity from content seeking and like Digital Bodleian for image searches. While they may use Google Books images, they prioritize high-resolution options and desire a unified GLAM catalogue for streamlined access to images across institutions.

Close Readers and Digital Discoverers are the majority use cases, with only very small numbers of users of Google Books engaged in text and data mining or using pictures.

The usage data showed only 0.5% usage of Oxford's Google Books was via the Bodleian Server.

The survey of Google Books Library Partners had responses from 14 of the 35 currently active partners, predominantly from the US with some representation from the UK and Europe. The responses revealed a varied landscape of platform choices for presenting the digitized texts, with the most common being linking from the library catalogue to Google Books; Utilising HathiTrust; Offering access via request or API (often via HathiTrust); and combining multiple methods.

The choice of platforms was influenced by drivers such as preference for linking directly to the resource from the library catalogue; desire for independence from Google; importance of collaborative developmental approaches; need for control over content delivery and digital preservation; and ensuring consistency of access across all digitized material.

Background

The Google Books Library Project partners with libraries across the world to digitise physical books so they can be searched and accessed online. The Bodleian Libraries is a foundational partner on this project - in the 2004-09 initial phase approximately 335,000 out-of-copyright books from our collection were digitised and made available to the world via Google's platform, and were hosted on a Bodleian Libraries server and made available via SOLO.

As part of the Bodleian Libraries Strategy (2022-2027), we are now engaging in a second phase, running until 2026, to digitise approximately 140,000 additional books. The digital documents landscape has evolved in the last 14 years, and there are now a number of potential platforms we could use to provide access to the digitised versions of our books. The Bodleian Google Books project team need to decide which to choose.

The Bodleian Libraries take an explicitly user-centred approach to our work. We know that the existing Google Books were downloaded from our server 408,000 times last academic year, but we do not know what the motivations of these users are, what they are doing with the content, how well accessing them individually via SOLO fits into our users' workflows, what barriers and frustrations they experience in using this resource, and what they would like to be able to do with the resource that they currently cannot. Without such information, the project team cannot make a user-centred decision about the most appropriate platform(s).

This is a gap in the body of knowledge about information-seeking behaviour and use of digitised texts: only one of the global Google Books partner libraries has undertaken user research to address these questions. The Bodleian Libraries is an acknowledged leader in HE libraries in applying a user-centred approach. Therefore, a secondary aim of this research is to add to the body of knowledge by sharing our findings both with Google Books partner libraries, and more broadly within the sector.

Research design

This research took a mixed-methods approach to develop a rich understanding of user needs for Google Books content presentation.

1. Focus Groups

The first strand used 90-minute semi-structured online focus groups to explore the context and workflows of existing users of digitised text content, and their preferences for different types of presentation platform. Participants for the focus groups were recruited through three paths:

(1) A pop-up survey on SOLO asking 'What is the main reason you are using SOLO now?' for a chance to win a £50 Bodleian Shop voucher. Those who answered to the screening question indicating they may have come across Google Book content were asked if they were willing to be contacted about participation in the project. If they indicated they agreed, they were sent full participant information and a link to another screening survey, asking specifically about Google Books use. Only those who had used Google Books were in scope for invitation to a focus group.

(2) We know that using Google Books for text and data mining is a particular use case, so recruited participants who are engaged or about to be engaged in this work at University of Oxford, from a variety of disciplinary areas and career stages.

(3) To explore user experiences of different platforms we needed to recruit participants who use such platforms. Via the Google Books Library Partners network, we identified three institutions willing to assist in recruitment of their Google Books users: Bayerische Staatsbibliothek (BSB) in

Germany (using IIIF viewer for whole collection), Harvard University Library (using HathiTrust) and University of Virginia Library (using IIIF viewer for a small sub-collection), both in the USA.

Unfortunately, BSB had to withdraw prior to recruitment of participants. In addition, University of Virginia was unable to recruit any participants, despite their best efforts.

All the focus groups were run by the Oxford research team in all aspects other than recruitment of participants, using the same questions (see Annex A).

2. Usage data

The second strand planned to analyse the existing data about the use of the Bodleian's Google Books to investigate access location, type of device used, and access route.

However, for technical reasons, it was not possible to install Google Analytics on the Bodleian's local hosting server, so the only data available was number of book views and number of PDF downloads by month.

Data about the Bodleian's books hosted on the Google Books platform was accessed via Google's library partner interface, called GRIN, though this was also only the number of full book views.

3. Library partners survey

The third strand was a survey of Google Books library partners, to investigate which platform they use to make their content available, why they chose this platform, and what feedback (if any) they have had from their users about the platform. The 'JISC Online Surveys'-hosted survey was distributed via the Google Books Library partner mailing list (which excludes Google, and only includes library partners). The survey questions are presented in Annex B.

Participants

1.1 Oxford focus groups

2,974 people completed the SOLO survey. Of these, 1,672 answered 'To find a book that is online' or 'To find a book in whatever format' and so were presented with the additional information and asked if they were willing to be emailed information about participating in an online focus group. 368 agreed.

Following receipt of the participant information, 85 people completed the second screening survey. 68 (80%) had used Google Books via SOLO, for the following reasons:

- 85% needed to read the book(s) for their academic research / degree studies.
- 6% can't remember, but they would have wanted or needed to read the book(s).
- 7% needed to read the book(s) for their personal research / school studies / college studies / work.
- 1% wanted to read the book(s) for their interest.

The text and data mining participants were recruited via mailing lists related to digital scholarship (including the Digital Scholarship @ Oxford mailing list and the Oxford Internet Institute postgraduate student mailing list). An additional 2 participants were recruited by personal invitation as the researchers were known to the research team as active in text and data mining.

20 participants were able to attend the three focus groups:

- All had used our existing Google Books – 5 extensively; 5 a moderate amount; 10 once or twice. Twelve had used our Google Books as they needed to read the books for their academic research / degree studies; two for their personal research interests (academic research but unaffiliated with an HE institution); six for text and data mining.
- Five were undergraduates (1st year Japanese; 1st year History; 1st year English Language & Literature; 3rd year History & Economics; 4th year Japanese & Chinese).
- Two were taught postgraduates (MSt Portuguese; Part-time MSt Literature and Arts).
- Five were DPhils (1st year Philosophy; 4th year Social Science; 4th year History of Art: Medieval Spain; 1st year Asian and Middle Eastern Studies: East Asia; 3rd year Medieval and Modern Languages: French).
- Four were researchers (Archaeology; Data Science; Digital Humanities; English Language & Literature)
- One was a retired professor of modern history.
- Two were Library Card Holders (artists' books; 19th Century Theology)
- One was an alumna with no current relationship with the Libraries (Post-Doc Medieval Studies @ Cambridge).

1.2 Harvard Focus group

The Harvard Library Head of UX and Digital Accessibility sent an email to their recruitment pool of 800 students, from which there were 34 volunteers. Individuals were invited to the focus group from across the breadth of subjects and levels. Six were able to make the scheduled time (17:30 – 19:00 GMT):

- All had used HathiTrust extensively.
- One was an undergraduate (1st year Biology).
- Two were taught postgraduates (3rd year English; 2nd year Languages)
- Two were PhDs (1st year History; 1st year Public Health)
- One was a post-doc (History)

Use cases

The data gathered through the four focus groups was analysed and synthesised into case studies, which describe five different use cases for Google Books in a Higher Education setting.

Case 1: Close Readers

Figure 1 Close readers like to read physical books, but will use digital books as a proxy. Monk reading, Oxford, Bodleian Library MS. Holkham misc. 21

Just a proxy

Close readers use Google Books as a proxy for the physical book for book for specific purpose or context. They have used physical version of the book in the past, or plan to use the physical version in the immediate future.

The physical version is used for reading the content, and Close Readers appreciate the haptic qualities of holding a "*real book*", reporting that they get deeper understanding and retention reading a physical book compared to onscreen.

Google Books is used for reference when Close Readers are writing, usually to quickly check citations and footnotes. This is particularly useful as accessing it at their computer doesn't disrupt the 'flow' of their writing (even to the extent where they have a physical copy on their own bookshelves 'somewhere').

Close Readers also use a Google Books version in advance of obtaining the physical version, to "*get a sense of whether it is worth going to the library or buying the book*", or to check the exact details so they can request the correct copy from offsite storage.

If it ain't in the 'library catalogue', it doesn't exist

In order for Close Readers to discover that a Google Books version exists, it must be in the 'normal' library catalogue, otherwise they do not come across it as part of their workflow. They will search on SOLO (or equivalent library catalogue) for the physical book, and so come across the digitised version, whether it is on the same record or the immediately adjacent record.

Faithful reproduction

Because Close Readers are only using the digitised version as a proxy for the physical version, it is absolutely essential that it is a true 'photograph' of physical book. Google Books is appreciated for this, including preservation of the page numbers, layout and all end papers of the item. Close Readers would prefer a somewhat higher quality of 'photograph', but have amused tolerance for folded corners, pages at a slant, and people's thumbs. They are only frustrated by missing parts of a page or where the text is too difficult to read (e.g. too faint; curved/compressed by tight binding; fuzzy due to movement at time image was taken).

Using and keeping

Close Readers would like to have the option to download either each chapter individually (as takes up less space and they often need only one part), or the whole book (to create their own virtual library). If forced to choose, they would prioritise chapter-by-chapter.

When Close Readers are using a Google Book to check a citation or footnote, they are willing to scroll to what they hope is the correct page, but they greatly prefer being able to search by keyword as it is not only quicker, but also works even if it turns out they don't know what the correct page is. They expect to be able to do this using the 'search' function of Acrobat Reader, or Control+F. This only works if the PDF has OCR (optical character recognition) applied, and Close Readers assume that this is a function of all PDFs.

Close Readers using Google Books as part of their writing workflow want to be able to copy-paste quotations from the book – including substantial passages. Their experience is that this is easier from PDFs than 'ebooks' and they expect all PDFs to facilitate this.

Close Readers want Google Books to be available to them as OCR PDFs, because they offer the most faithful 'book-like' experience and the basic functionality they need without needing to download any 'special' software. Although enhanced ebook functionality (e.g. ability to highlight text, make annotations, add notes and links) might be useful for some, it is not essential for Close Readers.

Summary

Close readers:

- Primarily use the physical copy and use Google Books content as proxy for physical book, to dip into for a specific purpose

- Prioritise discovery via their normal workflow
- Would not use if it is not in SOLO/library catalogue
- Second priority is that the Google Books version is a faithful ‘photographic’ reproduction of the physical book, including page numbers and layout
- Want to be able to download both individual chapters and the whole book
- Want OCR of text to support keyword searching
- Want to be able to easily copy-paste quotes
- Would find enhanced ebook functionality (highlight, annotate, add notes, add links) useful but not essential

Case 2: Digital Discoverers

Figure 2 Digital discoverers prefer digital books. Sketch showing the route traversed by Wm. Macgregor. Oxford, Bodleian Library D44:8 (2)

Digital first

Digital Discoverers prefer reading online books to physical books, either as a first preference or the only way they will read a text.

The main driver for this is convenience, as online books are instantly available from a computer. Digital Discoverers appreciate the ability to access their texts even when the libraries are closed, or when they are away from Oxford. Student Digital Discoverers do not have time to wait for a physical copy that is in high demand to be returned to the library, so appreciate instant access to the digital version. They also do not want to make a special trip from their accommodation to the library when in "essay crisis" or to quickly dip into a book as wider/background reading.

A subset of Digital Discoverers are print-disabled, and therefore need the accessibility features of an electronic version, including ability to change text size, change background colours, use a screen reader, all of which are impossible with physical books.

For Digital Discoverers, Google Books are merely one of the many types of online book they come across, and they are both unaware and uninterested in the fact that they are digitized versions of physical books in the Libraries' collection.

If it ain't in the 'library catalogue', it doesn't exist

As with Close Readers, most Digital Discoverers come across a Google Books via the 'normal' library catalogue, as checking this is part of their normal workflow. They will search on SOLO (or equivalent) for the book they want and filter (either explicitly or by looking for the green dot for e-resources) for an online version, and so come across the Google Book.

Through this method, the Harvard Digital Discoverers observed that many of the items they were looking for were available on the same platform (HathiTrust), and so started to search this platform for items relevant to their research.

Faithful reproduction with added bells and whistles

Although they are not using the digitised version as a proxy for the physical version, Digital Discoverers still appreciate that a Google Book PDF is a true 'photograph' of physical book, because the references they are following are often to the physical book, and it also means they don't have to deal with the challenges of citing an ebook.

They expect PDFs to have OCR applied and may be confused by PDFs that are only images, as it is outside their experience. OCR is essential for use with screen readers.

Digital Discoverers expect ebook functionality such as being able to highlight passages, annotate the document, and add notes and links in the same platform so they are 'all together'. Some may have the technical skills or know how to achieve this using add-ons or other programmes.

Like Close Readers, Digital Discoverers would like to have the option to download either each chapter individually or the whole book, and be able to copy-paste quotations from the book – including substantial passages.

Summary

Digital Discoverers:

- Prefer reading online books to physical books and use Google Books content as a way to get an ebook
- Prioritise discovery via their normal workflow
- Would not come across a Google Books to use if it is not in SOLO
- Want to be able to download both individual chapters and the whole book
- Need a 'photographic' reproduction of the physical book for page numbers and layout.
- Want OCR of text to support keyword searching and screen reader use
- Need to be able to easily copy-paste quotes and want enhanced ebook functionality (highlight, annotate, add notes, add links)

Case 3: Specialist Data Miners

Figure 3 Data miners use book data as a vast corpus. Grant of mining rights, Oxford, Bodleian Library MS. Ch. Suffolk 1221

Everything, everywhere, all at once

Specialist Data Miners have the technical expertise to work with text and data corpora at scale, building their own tools and transforming data. They want as much data as possible and are not overly concerned as to the accuracy of that data. The breadth of data is their primary concern, as they are engaged in 'distant reading' of texts or images within the books, and want to make statistically significant conclusions from their data: "*I get twitchy if I'm not getting everything*". In their search for data volume, Specialist Data Miners will visit multiple corpora and digital resources, relying on serendipity and library resource discovery systems such as SOLO.

Awareness of the shape of novel corpora such as Oxford's Google Books corpus can lead Specialist Data Miners to form new research questions. "*the nineteenth century is an ocean of untapped data*".

Provenance matters

Specialist Data Miners are particularly concerned with understanding the shape and origins of the data they acquire, due to their use of multiple sources. They have a good understanding of copyright restrictions as they apply to the Google Books content (and its derivatives within HathiTrust) and other sites such as Project Gutenberg and Internet Archive. They realise there are limitations in how the Google Books corpus has been created, especially regarding the library partners being primarily drawn from the USA and Western Europe. In an ideal world, these researchers would like more metadata, especially around technical processes such as camera specifications and post-processing: "*I've never been unhappy to have too much metadata*"

In order to address primary research issues of provenance, collections bias, reproducibility, and modern decisions affecting the nature of the data, Specialist Data Miners want documentation to accompany digital research datasets, which contains information on data provenance, dataset formation, characterisation, processing and metadata. They also need to know the accuracy and reliability of this metadata so they can make informed decisions.

Two specific provenance needs of Specialist Data Miners are the disambiguation within the Google Books corpus of different copies of the same text; and a publicly available record of changes in the Google Books corpus due to reprocessing (where copies of a text are replaced with an 'improved' one).

Imperfection is acceptable

Specialist Data Miners can afford to accept error rates (such as "*scruffy transcription*" and low-resolution images) as they have the skills and tools to download, refine and modify data to create a dataset from different data sources that is regularised and interoperable. They expect to match data, both in terms of formatting and also ontological matching, particularly when working with multiple languages or writing systems. They use their skills in coding (in languages such as R and Python) to create these datasets as well as the tools with which to interrogate them. They would prefer data to be in XML or JSON (or both) for their ease of use, but they are able to work with a variety of file formats and transform them into the file types they need.

Books are multimedia objects

Specialist Data Miners are not just concerned with text, but also with images, different informational forms such as maps and musical notation, as well as with book page elements such as title pages and indices, and the interrelationships between elements on a page or within a book. They apply computational methods to extract these at vast scale.

Reproducibility

As researchers working at scale, the citability and reproducibility of data is crucial to Specialist Data Miners, both in terms of understanding the contexts which created the corpora of book data they collect, and also how they can cite the corpus they create.

Specialist Data Miners, as creators of datasets and tools with which to interrogate book data, would like to be able to input their tools and research back into their source datasets to improve functionality and usability for other researchers.

Using and building systems

Whilst they often access data by 'traditional' download or API, Specialist Data Miners are amenable to accessing data via a trusted research environment or HathiTrust Research Centre's data capsule, if it means they have access to more data, a reduced environmental impact of high-performance computing, and the ability to have a URI (uniform resource identifier) or PID (persistent identifier) for a dataset.

Although they have the skills to build their own tools and transform data, Specialist Data Miners are happy to use systems that exist, notably HathiTrust Research Centre, but want to be able to perform some top-level searches to restrict data by date ranges.

Summary

Specialist Data Miners:

- Use Google Books to build a corpus for text and data mining
- Build their own tools to interrogate their data
- Use a variety of sources and tools to create their datasets, including Google Books, HathiTrust and HathiTrust Research Centre, as well as subject specialist resources
- Prioritise volume - everything everywhere, all at once
- Second priority ease of bulk download
- Value good quality metadata, especially around understanding the provenance of the data. This includes both MARC and technical metadata
- Can live with lower resolution and lower accuracy OCR as long as they have a high volume of data

Case 4: Data Extractors

Figure 4 Data extractors need some tools to data mine. Honey extractors, Oxford, Bodleian Library MS. Rawl. G. 98

Tools required

Data Extractors are researchers interested in working with book data at scale like the Specialist Data Miners. However, they don't (yet) have the advanced skills of the Data Miners, and rely on systems for text and data mining provided by text and data mining (TDM) aggregators and platforms. If the platform tools do not meet their needs, they are unable to build their own tools instead. These researchers typically search, sort, filter and refine data and use tools to refine hypotheses and the available data. With experience and skills development, they may become Specialist Data Miners.

Corpus provenance

Data Extractors are concerned with the provenance and formation of their dataset, in order to address primary research issues of provenance, collections bias, reproducibility, and modern curation decisions affecting the nature of the data. They also need to understand what decisions and mechanisms are happening with the ready-made tools they use on the platforms.

Data Extractors rely on the metadata they are provided in corpora as tools and clues or 'guardrails' with which to understand and estimate the reliability of their data. They need transparency, and ideally documentation, about the factors that impact the reproducibility of the data (e.g. IP address).

TDM platforms

Data Extractors use a range of platforms and data sources and compare and contrast both the data, and the tools provided.

As Data Extractors are dependent on tools built by others, they are critical of the TDM platforms in both senses: in the need to understand the provenance of the data and decisions which led to the creation of the tools, and of the limitations in functionality which impact their research. They are particularly sensitive to new versions of platforms and tools, and have strong opinions regarding what they consider as improvements and deteriorations in functionality.

Issues of importance to Data Extractors include: ease of use without coding skills; speed of downloads; features for different languages; reliability of advanced search functionality (date range, country search, language search); quick looks at data (thumbnails); ability to answer their research questions. Data extractors take pains to be aware of the limitations of the functionality available to them. "*Knowing your tool is important*".

Due to this extensive experience with different TDM platforms, Data Extractors have a clear conceptualisation of their ideal features:

- Access to the textual data via API or copy-paste
- Named entity recognition, in particular of people, places and dates would be a "*game changer*"
- Easy access to the underlying data and metadata via a simple tool

They can be frustrated with their tools, but do not expect all digital texts to be digital editions (a layer of scholarship for a text which goes well beyond a library finding aid or published work).

Tools to aid learning

Data Extractors may use the Ngrams functionality within Google Books, and the utility of the tool, and other tools such as the HathiTrust equivalent, as a teaching resource for digital humanities and computational thinking was echoed by Specialist Data Miners, thanks to its easy-to-use online functionality requiring no coding skills or downloads.

Data Extractors use Ngrams as a diagnostic or interrogative tool to suggest and test hypotheses, benefitting from the fact it works across different languages. However, this would be a precursor to the main research methodology (usually a statistics study).

Summary

Data Extractors:

- Use Google Books and other platforms as a corpus and set of tools for text and data mining
- Need tools to be provided for them, and want to understand how these tools work and were created
- Are critical of their tools, questioning both their creation, and their usability
- Want volume of data: text and images
- Want to be able to copy and paste, or access texts via an API
- Need to understand the provenance of the data and the tools

Case 5: Picture Pickers

Figure 5 Picture pickers seek digital images from books, like this famous image from the Codex Mendoza Oxford, Bodleian Library MS. Arch. Selden. A. 1

Activity-based

Picture Pickers are different from other Use Cases, which are person-based, as they are activity-based. Picture Pickers can be anyone who is looking for very high-quality images to illustrate their teaching presentations, conference presentations, or written publications.

Distinct from content

Picture Pickers have a specific workflow when they are looking for images, completely separate from their persona as Close Reader, Digital Discoverer, Specialist Data Miner or Data Extractor. They regard looking for content and looking for images as completely separate activities.

They need relevant, meaningful and interesting images that is high resolution. They therefore search Digital Bodleian and similar resources. They may use Google Books images as they come across them, but would try to find higher quality image.

They do not wish to use such large files for the text of Google Books, as it would take up too much space on their computer, or they would need a more powerful computer to access them.

Cross-GLAM catalogue

Picture Pickers would like to have a single catalogue so they could search for images across all six Oxford GLAM institutions instead of repeating the search on their digital resources individually.

Summary

Picture Pickers:

- Can be any of the other personas when looking for illustrations for teaching, presentations, and publications
- Regard as completely separate function than seeking content
- Use Google Books content for pictures if they have no other source
- Use Digital Bodleian
- Want single catalogue across GLAM for pictures.
- Like very high-quality images for pictures; would not want such large files for text of Google Books

Functionality required by users

	Close Readers	Digital Discoverers	Specialist Data Miners	Data Extractors	Picture Pickers
Available via Google platform only	No	No	Partial	Partial	Partial
Link from SOLO to Google version	No	No	Not relevant	Not relevant	Partial
Host locally (PDF with OCR) and link from SOLO	**Yes**	**Yes**	No	No	Partial
Host locally (PDF with OCR) and linking from SOLO, with additional functionality to enable data extraction	**Yes**	**Yes**	Partial (limited corpus)	Partial (limited corpus)	Partial
Host in Digital Bodleian only	No	No	No	No	**Yes**
Host in Hathi Trust, exposing to Hathi Trust Research Centre, and link from SOLO	**Yes**	**Yes**	**Yes**	**Yes**	Partial

Holistic view

Users of Google Books do not do so in isolation from other formats or sources. A number of issues arose in the focus groups which were not directly relevant to any specific use case, but provide insight into the larger ecosystem within which users, of all types, are operating.

Users are confused and frustrated by having to learn the peculiarities of so many different platforms. They would like to see the same basic functionality across all platforms instead of having to discover which ones allow downloads, which allow copy-paste, which support additional functionality, or other functions.

Some users are concerned about the environmental impact of book format provision, with some favouring ebooks as they perceive not having to transport physical books as beneficial to the

environment, whereas other are worried about the environmental impact of server farms and cloud storage.

Oxford students would like a system-wide approach to book provision – one suggested prioritising the purchase of ebooks where the physical book is not in College libraries.

Whilst some users are well versed in the technologies available to them to support their use of information resources, and confident in their use, others would like the Libraries to provide such guidance and support, particularly signposting in a “did you know” approach – from browser extensions to OCR tools to how tablets can support PDF annotation.

Usage Data

Table 1: Book views by platform

Platform	AY2020/21	AY2021/22	AY2022/23
DBooks server[1]	453,781	428,409	408,041
Google	78,600,033	74,208,175	89,077,061

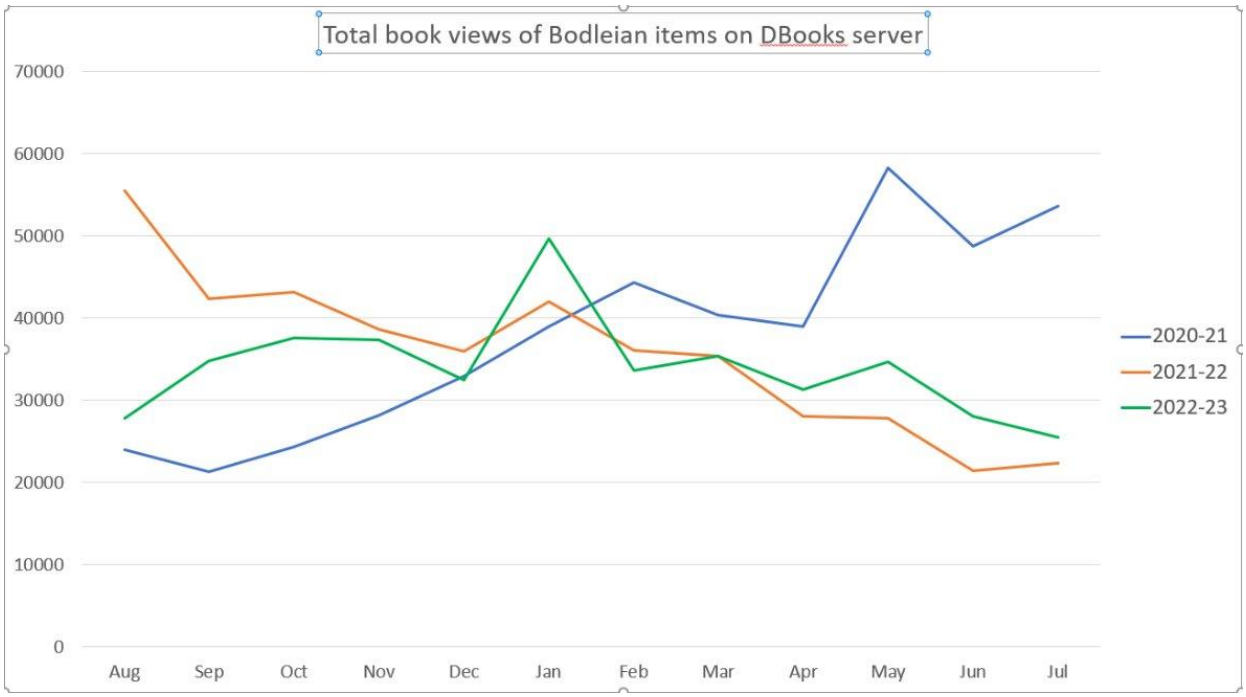

Figure 1: Pattern of use of Bodleian Libraries DBooks server over the academic year.

[1] The Bodleian Google Books server

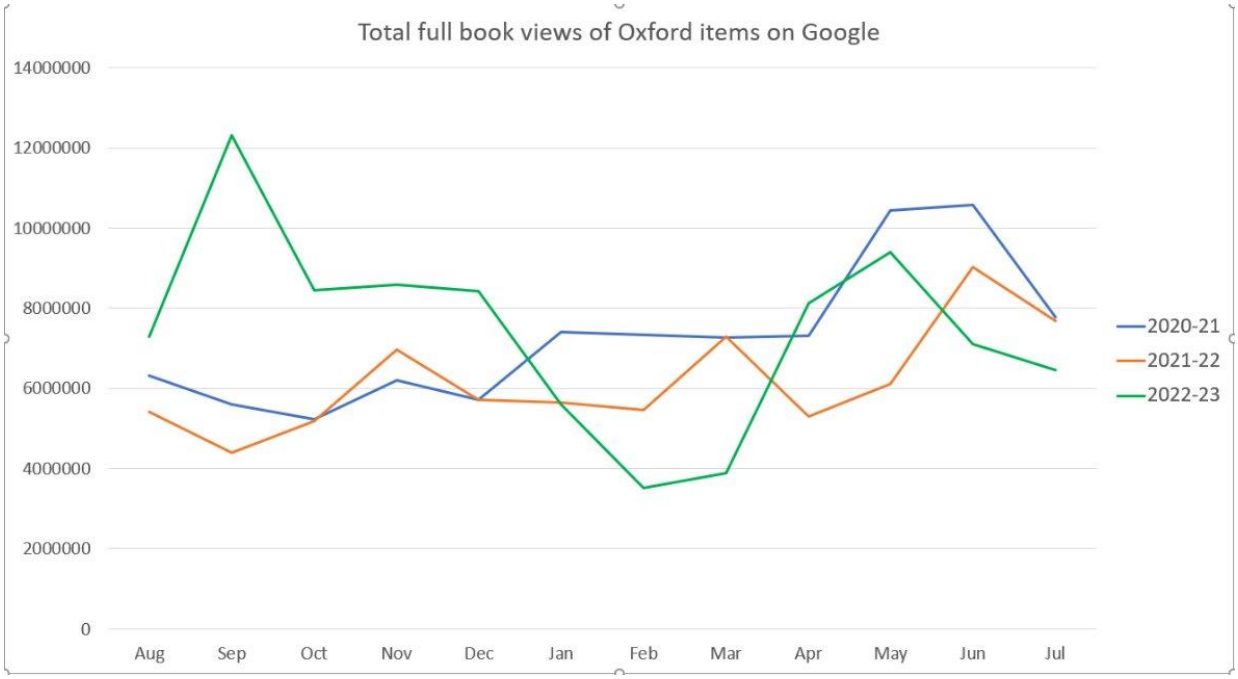

Figure 2: Pattern of use of Bodleian Libraries items on Google Books over the academic year.

Figure 3: Use of PDFs on DBooks server via Oxford Reading Lists Online (ORLO) Platform

Library Partners Survey

The survey was circulated to the Google Books Library Partner mailing list of c.75 institutions. We were informed by Google that c.35 Library Partners are currently actively scanning books and directly involved in the project at present. We received 14 responses. These comprised 9 US institutions, 1 UK institution (of the 3 UK partners, one of whom is Oxford), and 4 European institutions (Belgium, Germany and the Netherlands). Respondents from those institutions had a variety of roles and job titles including Librarian, IT manager, Software Developer Manager, Collections Management Librarian, Systems Librarian, Head of Digitisation, Repository Manager, Curator of Digital Collections, Head of Collections Development, and Metadata Manager.

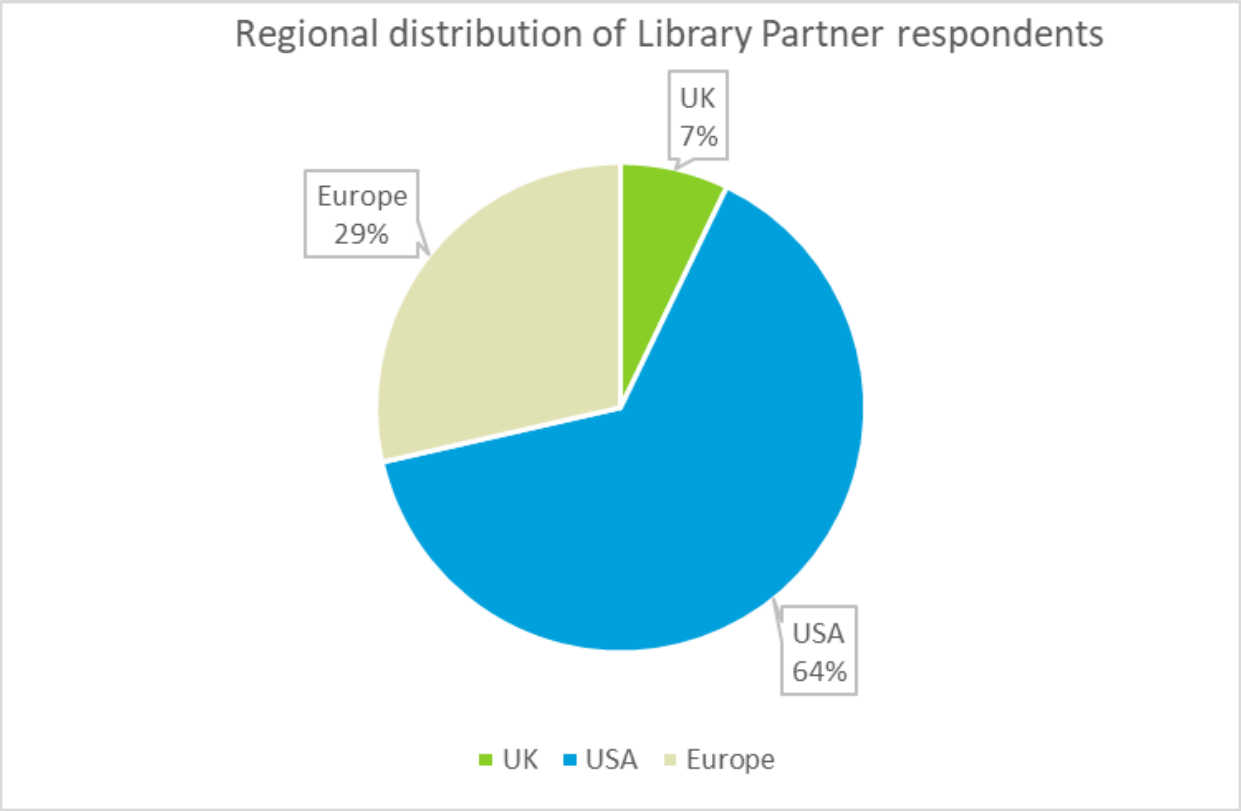

Figure 4: Geographic spread of responses to Google Library Partners survey

The institutions were asked which platform they used to present their Google Books data, and many had more than one option or route for the content to be discoverable. Of the respondents, 6 libraries used a link from their library catalogue which pointed to the Google Books platform, 6 used HathiTrust (all US institutions), 2 used a page turning software platform, 3 used a system based on IIIF, 6 allowed access to files upon request or via an API (of which 5 were through HathiTrust and one was through a bespoke interface), 1 connected their content to national digital infrastructure, 1 did not yet have a system in place, and 5 used a combination of these methods.

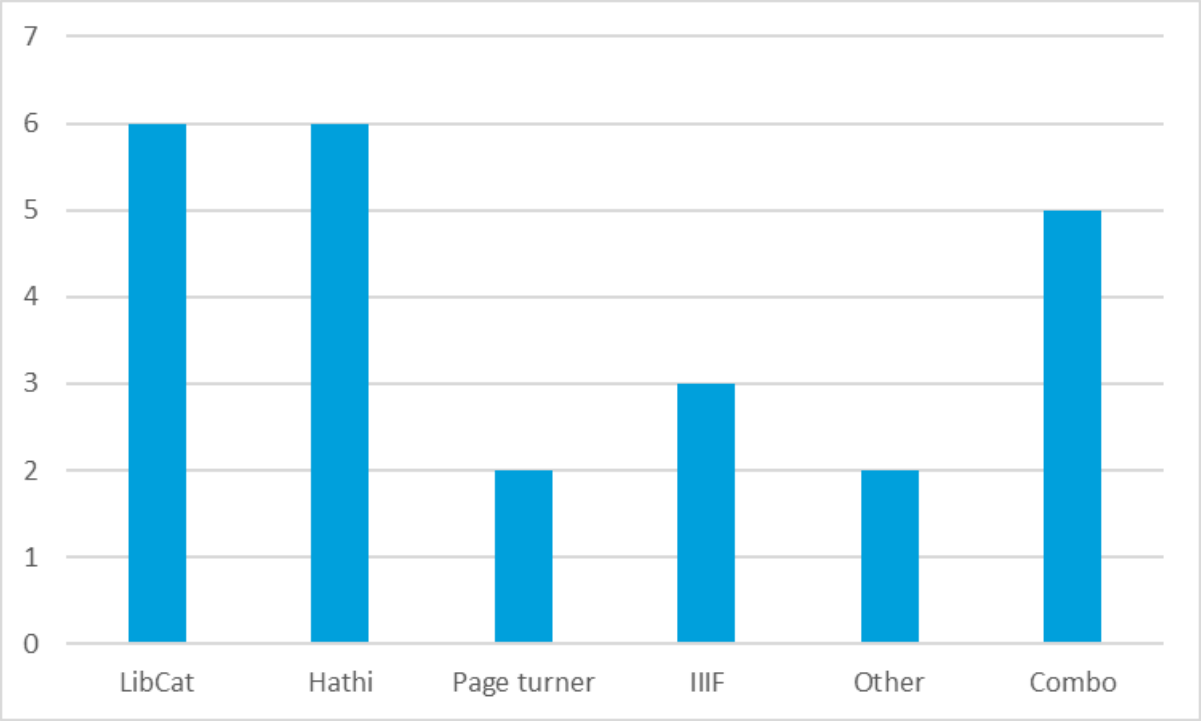

Figure 5: Platforms used for Google Books by Library Partner survey respondents

The reasons for choosing these platforms varied. A common thread was to create a system that was independent of Google, and another was working with other (US) Google Books Library Partners as a consortium to develop a shared solution, which developed into HathiTrust. The reasons for building solutions outside of Google included retaining control of content delivery, local ownership and stewardship, digital preservation, and to ensure consistency of access to all the library's digitised material through one platform. Respondents highlighted how important it was to link from their main library catalogue to the resource, whether this was on Google Books, HathiTrust, or an IIIF viewer. Three libraries had a function that allowed users to select only their institutional corpus of Google Books, through HathiTrust or within the library catalogue.

Only 1 partner (USA) had undertaken audience research, but there are plans in the Google Books Library Partner community to undertake audience research across institutions and countries. Through informal feedback, respondents noted that users liked links from the library catalogue, disliked the page turner viewer, and liked the links between objects on an IIIF-based system.

"*Our page turning viewer is hard to use in some cases due to scrolling and scaling. Folks can zoom in and out, but then switching to another page they want the same zoom. Since screens are all sorts of sizes, it's tedious to read page images over a format of text that flows as it resizes.*" US Respondent

HathiTrust institutions noted an improvement in speed and functionality due to the work of the HathiTrust Usability Working Group, User Experience Special Interest Group and User Engagement Task Force.

ANNEX A: Focus Group Questions

Introductions

Let's start with some brief introductions. As I said, I am <name> and I am <job role at Bodleian Libraries> and am involved in this research <reason>.

- <name of first participant> please will you introduce yourself with your name, your broad subject area and your career stage, for example doctoral student, professor etc.

Thank you. Now <name of second participant>

etc

Context

Thank you. Now we all know each other, we will move onto the discussions about your uses of Google Books.

- Firstly, I am interested in why you use Google Books digitized texts, starting with the specific piece of research or assignment you are working on. Who would like to go first?
- Why did you need to use a digitized book(s) for this?

[Depending on their answers, choose relevant follow up questions.]

Text and data mining

- What specifically (but briefly!) do you want to do with text-mined data?
- What kinds of data and/or metadata do you need for your research or study?
- Why did you choose Google Books data in particular?
- What other sources of data have you used or plan to use?

Content

- Was a Google Book your first choice?
- How did you come across it? Where are you 'go to' search options?
- Did you have any alternatives to a specific book? For example a different edition / selecting something else from the reading list / a different title …
- How do you typically work with information from books? (e.g. make notes on the whole book or chapter electronically/in a physical notebook ; just note a specific idea / equation / quote ; annotate the text physically / electronically …)

Non-textual content (e.g. images, physicality of item)

- What specifically (but briefly!) do you want to do with it?
- Why did you choose Google Books item(s) in particular?
- Did you have any alternatives to this specific book?
- What other sources have you used or plan to use?
- Does anyone else have a similar use case?

[Continue until all have spoken]

Requirements and preferences

Now let's move on to your requirements and preferences in working with digital texts.

- <name of last person to speak in previous section> you use Google Books for <text and data mining / the content / images etc>

[select relevant questions for each person]

Text and data mining

- In what formats do you need the data to be available?
- How would you like to filter the data for study? For example, author, date, language, place of publication, subject, genre …
- What is the preferred mechanism for accessing the data? For example, download individual files or filesets, API…

Content

- In what format would you prefer electronic books? For example PDF, interactive, downloadable to Kindle …
- What would you like to be able to do with an electronic text?

Non-textual content (e.g. images, physicality of item)

- If you could have your ideal format for digitized texts, what would it be like?
- What would it enable you to do?
- <name of other participant using for the same purpose> do you agree or do you have different needs and preferences?
- Switching to a different use case <name>

[Continue until all have spoken]

- Now you have heard everyone's views, does anybody want to add anything before we move on to the next section?

Platforms

Now let's move on to your experiences and preferences in the platforms for delivering digital texts.

<name of someone who hasn't spoken much>

[select relevant questions for each person]

Text and data mining

- What limitations have you found with using the <Oxford / Harvard> platform for Google Books?
- If you have used other text and data mined datasets, did they have the same limitations?
- Does the <Oxford / Harvard> platform for Google Books avoid limitations of other platforms?
- What kinds of documentation on data/metadata do you expect from a text-mining dataset?

Content

- What limitations have you found with using the <Oxford / Harvard> platform for Google Books?
- If you have used other electronic book platforms, did they have the same limitations?
- Does the <Oxford / Harvard> platform for Google Books avoid limitations of other platforms?

Non-textual content (e.g. images, physicality of item)

- What limitations have you found with using the <Oxford / Harvard> platform for Google Books?
- If you have used other digitized text platforms, did they have the same limitations?
- Does the <Oxford / Harvard> platform for Google Books avoid limitations of other platforms?

<name of other participant using for the same purpose> do you agree?

[Continue until all have spoken]

- As we reach the end of our session, does anybody want to add anything before we finish?

ANNEX B: Library Partners Survey Questions

Which institution do you work for?

What platform, viewer or system do you use to surface or share your Google Books content with your users? (select all that apply)

- Link to Google Books platform from library resource discovery system/catalogue
- PDF from library resource discovery system/catalogue
- IIIF viewer
- Page turning software
- HathiTrust
- Other
    - Please describe

Why did your organisation choose this viewer for its Google Books content?

Are you thinking about changing your Google Books viewer to another system?

- Yes
- No
- Not sure

If 'Yes': Which system?

- Link to Google Books platform from library resource discovery system/catalogue
- PDF from library resource discovery system/catalogue
- IIIF viewer
- Page turning software
- HathiTrust
- Other
    - Please describe:

Why?

Do you have an API for users to access your Google Books content?

- Yes
- No

If 'Yes': How is this achieved?

- Through HathiTrust
- Through bespoke interface
- Other
    - Please describe:

Do you have a publicly available method for your users to search or access only your Google Books content?

- Yes
- No

If 'Yes': How have you achieved this?

Have you done any audience or user research on how your readers use your Google Books content?

- Yes
- No

If ‘Yes’: If possible, are you able to briefly summarise the extent of the investigations and the results. Alternatively, you can send a report or document to megan.gooch@bodleian.ox.ac.uk

Have you received any feedback from your users about the platform, viewer or system you use?

- Yes
- No

If ‘Yes’: If possible, are you able to briefly summarise the extent of this feedback. Alternatively, you can send a report or document to megan.gooch@bodleian.ox.ac.uk

What is your role within your library?