From provider to partner: how digital humanities sparked a change in Gale's relationship with universities

The past decade has seen huge growth in the teaching and research of what is broadly called digital humanities (DH). Increases in computing power and data availability have seen a rise in individual researchers and research groups working on digital scholarship projects in the humanities, arts and social sciences.

This article shows how publishers of traditional digital archives have adapted to the increasing prevalence of DH amongst their traditional customers. The success of this adaptation depends entirely on the relationship with the academic community, and Gale has seen a shift from being a provider of products to a partner, trusted to help libraries, scholars and institutions achieve their objectives.

As a leading global provider of digital archives, Gale is well placed to review the current state of DH research and teaching, and this article will discuss significant academic events that have brought scholars, librarians and students together, and the lessons learned for institutions around the world looking to expand into DH.

Finally, the article looks at how working to understand the common challenges and barriers to DH research and teaching has pushed many archive publishers to re-evaluate traditional archive publishing and enable new and innovative ways to explore the past.

Keywords
Digital humanities; Gale; digital archives; primary sources; libraries; academic publishing

CHRIS HOUGHTON

Head of Digital Scholarship
International Gale Primary Sources

SARAH KETCHLEY

Faculty Affiliate Instructor
University of Washington

An introduction to digital archives and Gale

Over the past 20 years digital archives have become an essential resource in university libraries. A natural evolution of microfilm and CD-ROM archive access, a digital archive usually provides cloud-hosted access to vast amounts of primary source material accessible through a web interface. For researchers around the world, a digital archive democratizes access to many of the world's leading research, national, private and public libraries, making access to material available on their desktops that would previously have necessitated a visit to the library in question. There are several private companies currently digitizing large archive collections, including Gale, ProQuest, Adam Matthew, Wiley and EBSCO, and archives will generally be available for institutions to subscribe to or purchase. With recent increases in technological capabilities, there are numerous large regional/national open digitization projects, including Europeana, the Digital Public Library of America (DPLA) and the Hathi Trust, that provide access to large digital archive collections, as well as smaller collections being digitized at the institutional level.

Gale's digital archive programme began in 2002 with *Eighteenth Century Collections Online* (*ECCO*), one of the most ambitious digitization projects of the time. *ECCO* provides digital versions of the 18th-century texts catalogued in the *English Short Title Catalogue*[1], and gives researchers desktop access to 'every significant English-language and foreign-language title printed in the UK between the years 1701 and 1800'.[2]

The following year saw Gale publish *The Times Digital Archive*, and in the 16 years since, we have published hundreds of digital archives containing over 250 million pages of often unique, difficult-to-access documents from six continents, covering nine centuries.

Digital archive publishing is becoming more prevalent as the technology for digitizing historical artefacts gets cheaper and more ubiquitous, and many universities, museums, libraries and other research institutions have digitization projects of their own. At the other end of the digitization spectrum, there are a number of commercial publishers working at a global scale, digitizing significant publications or national library collections; Gale, ProQuest, Adam Matthew, Alexander Street Press, EBSCO, Wiley and Brill among them.

Most commercially available digital archives are full-text searchable, meaning that any researcher can search for a word or phrase and theoretically find it anywhere in a document. This functionality requires a process of optical character recognition (OCR), which is used to transliterate large corpora of documents. For publishers of historical archives, this requires running OCR software over scanned images, converting the letters on the page into machine-readable text, and capturing their position on the page. As a result, users can search for words and the underlying OCR text will identify the page, and location on the page, of each matching term in the archive.

A changing relationship: archives as infrastructure

Gale was founded in 1954 as a publisher of directories and reference titles and is now part of Cengage Learning, one of the world's largest educational publishers. Within Cengage, Gale used to be known as Library Reference', a name which accurately reflected the traditional business model for a publisher of primary source archives working with libraries to provide reference material. In the early 2000s, around the world, large national consortia such as Couperin and DFG operated as de facto buying groups, negotiating discounted access to products for their members. In the UK, Jisc purchased the archive with funding it received to make the archives freely available to all UK higher and further education institutions.

After the 2008 financial crisis and subsequent funding cuts to higher education, the ability of many consortia to offer this kind of large investment in archives drastically reduced, and publishers found themselves having to change business models and operations to deal directly with university libraries on a much more regular basis.

These fundamental changes to the university sector in many major markets prompted significant adaptation in operations. Now, rather than dealing with a centralized funding body, commercial publishers had to have relationships with individual institutions. For Gale, this meant a significant investment in customer-facing operations and a shift in focus to not just working with libraries but understanding what they needed to be successful.

'publishers were now doing more than ever to speak with, and understand the needs of, the primary users of archives'

Digital archives are a significant investment for any university, and the majority of libraries communicated the need to provide evidence of academic support for an archive within their institution before countenancing a purchase. This need for academic support meant that, with the agreement of the library, publishers were now doing more than ever to speak with, and understand the needs of, the primary users of archives – the teachers, scholars and academics who could use archives like ECCO or *The Times Digital Archive* in research or in the classroom.

These relationships with academics would naturally become significant. By understanding the research topics, needs and objectives of scholars, archive publishers were able to ensure two things: that they digitized archives and created new products where there was a desire for the material and that digital archives were presented in ways that would best support research and teaching needs.

As the 2010s progressed, Gale started to see the beginnings of an evolution in this relationship, especially in universities that had not traditionally purchased archives. The feedback we were getting was that many archives were now seen as crucial to the work of a humanities department; we would receive orders for digital archives because an academic was moving to a new institution and purchasing the archive was a condition of their move. This evolution played out in the way that libraries purchased. In many markets around the world, evidence started to build of a move away from traditional end-of-year purchasing, as libraries found new ways of funding these purchases.

Gale's relationship with libraries was changing again, and we would find ourselves supporting significant capital bids for investment in digital archives. Allied to this was significant growth in new markets such as China, where in 2015 we developed the Gale Scholar programme to help ambitious Chinese institutions quickly develop digital libraries on a par with the top universities around the world, a programme that is now being exported globally due to its popularity.

First contact with digital humanities

Supporting 'gather data and analyse'

More and more, institutions were reflecting a need to not just search and retrieve documents as had been commonplace, but also to gather data at scale and analyse it. Digital archives started to respond by creating cross-searches of multiple archives, by incorporating analytical tools into archives and by facilitating access to OCR.

For Gale, these developments took hold in 2013 and 2014 with Artemis Primary Sources, one of the first major archive cross-searches (later rebranded as Gale Primary Sources[3]), which not only allowed the potential search of hundreds of millions of pages of primary source content, but included analytical tools to allow users to look at them through a different lens. Significantly, users now had the ability to download the OCR for individual documents, which they could then combine with OCR for other documents into a corpus ready for analysis.

'requests to access the underlying data of archives, both metadata and OCR, for the purposes of text mining'

Examples of digital scholarship

The other change that was happening during the early 2010s was that we were starting to see requests to access the underlying data of archives, both metadata and OCR, for the purposes of text mining.

From the start, we were keen to agree to these requests, but it began as an ad hoc process, often taking many months. For example, in 2011 Gale was contacted by Dr Michaela Mahlberg, PhD, supervisor for Kat Gupta, who was studying at the University of Nottingham. Gupta was interested in getting access to the OCR for *The Times Digital Archive* for the years 1908–1914 while researching their monograph, *Representation of the British Suffrage Movement*.[4] Focusing on *The Times*, Gupta's monograph, 'uses corpus linguistics to examine how suffrage campaigners' different ideologies were conflated in the newspaper over a crucial time period'.[5] Gupta was able to extract certain sections of the newspaper and mine them for references to 'suffrage', 'suffragism', 'suffragette' etc in order to identify the prevailing attitudes to the movement, amongst other conclusions.

Being exposed to scholarship really helped Gale to understand how academics were using archive data, and whether they were using digital archives as we had envisioned or were making new applications. We would connect with researchers who were using metadata (such as word counts[6]) that we had never considered making widely available as the basis of their research.

Making data available

Collaborating with researchers to provide data for these projects and many others like them really helped to give an understanding of DH and the DH community. In 2014 Gale made the decision to move the data provision from an ad hoc process to something more structured, and Gale became the first humanities publisher to make underlying OCR and metadata available to customers through text and data mining (TDM) drives. Subsequently, most commercial publishers of digital archives make their OCR text and, in some cases, metadata available to researchers.

This development helped to crystallize Gale's relationship with the DH community. For the first time, we knew exactly who was using our data, and had the option to remain in contact to understand how they were using it.

'Gale became the first humanities publisher to make underlying OCR and metadata available'

Discovering common barriers

Demographics around DH soon became apparent; there was a relatively small set of core practitioners: researchers who were creating projects, writing code and were comfortable with managing large data sets. However, most scholars and institutions around the world that were interested in working in DH would often find the path to successful projects barred by a few common barriers:

1) ***Access to relevant data in an optimized format***
   Bringing together a significant corpus of data for analysis often involved insurmountable challenges: finding the data in the first place; combining data from disparate sources; cleaning the data to prepare it for analysis. The time and technical skills needed to undertake these processes were proving to be an obstacle for many. Figure 1 shows a typical research process, based on academic insight.[7] Many researchers were telling us several of the research steps could each take up to 80% of the allotted project time. Cleaning data and creating exploratory tools were proving to be extremely time-consuming activities.

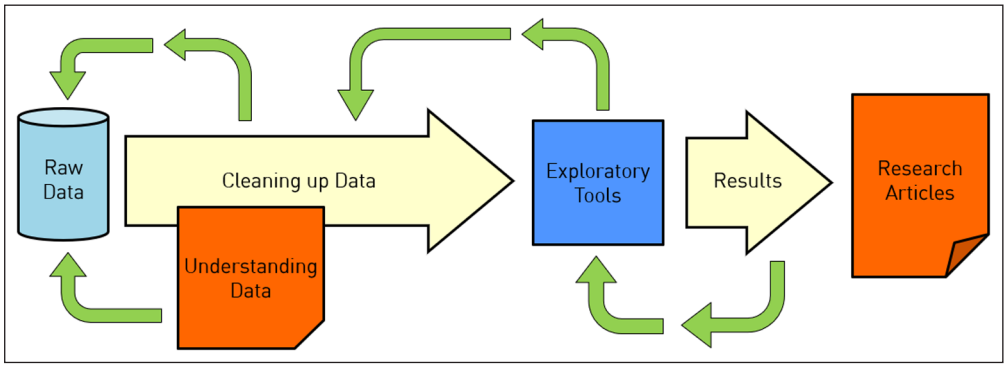

Figure 1. A typical digital humanities research process
Based on COMHIS Collective, University of Helsinki, Text and Data Mining Eighteenth Century based on ESTC & ECCO, BSECS Conference 2017, Oxford. [Slide 5]

2) ***Hosting data***
   This challenge occurred frequently when universities purchased the TDM drives. For many Gale digital archives, the OCR and metadata equates to several terabytes of data, which makes it sometimes problematic or expensive to host locally. Anecdotal evidence suggested that some university bureaucracies made it hard for researchers to get access to the data that they had purchased, if the university was able to find the server space to mount the drives at all.

3) ***Tools to analyse data can be challenging***
   The analysis of large corpora of OCR or metadata text typically requires a degree of coding proficiency. Experienced DH practitioners can be coders, while we would often see data analysis from academics who might consider themselves traditional

humanists but had taught themselves some basic coding. The message that came across strongly was that this need for coding often acted as a barrier to teaching DH in the undergraduate classroom, and to wider DH take up, as it required a significant time commitment.

Developing a solution

In consideration of these barriers, it soon became apparent that there was an opportunity for us to develop a solution to support the existing digital humanities community and help to spread its skills and insights beyond the core practitioners.

Gale started building a cloud-hosted text and data mining platform in 2015, and in 2018 released Gale Digital Scholar Lab,[8] the first (and currently only) product to combine the broad range of archives available from Gale with powerful text mining and natural language processing (NLP) tools. Developing the Lab proved to be a significant process, featuring several redesigns. At every stage, we made sure to solicit input from scholars to ensure that the product would deliver for established DH practitioners and those looking to break into DH.

Designed to overcome the three common barriers, take-up was strong immediately, with initial customers in China, Singapore, Australia and UAE, spreading to Europe and the United States. Libraries identified the Lab as a tool to help them support DH in a relatively low impact way, with vast archives of data optimized for use, simplified cleaning, and tool customization that did not require any existing coding knowledge.

Challenges and implications

With a goal of continually developing the Lab to support the needs of the DH community and the wider academic community, a series of technical challenges and content implications arise.

Development challenges

For Gale, the Lab represents a new model for development. Unlike a digital archive, which is essentially a static product, the Gale Digital Scholar Lab iterates and develops in line with user feedback and market need. This development work is expensive and time-consuming, and like any large corporation with a varied product portfolio, this means competing to ensure that investment into the Lab is consistent.

'one extremely important facet is to make sure that there is input from academics around the world'

The requirement to understand the market is now stronger than ever, as Gale works to identify research trends, development needs and common issues in order to try and provide solutions where appropriate. As a publisher with a global profile, one extremely important facet is to make sure that there is input from academics around the world, especially non-English native speakers, a demographic currently under-represented in DH.

Product challenges

Developing a software solution often leads to challenges as myriad development paths become apparent, and prioritization is needed. Some of the immediate challenges include:

- ***increasing and improving tools***
  The Lab utilizes mostly open-source tools, which will be developed and expanded.
- ***increased outputs***
  By introducing tools to develop the kinds of outputs students are tasked to create in DH courses (interactive timelines, enhanced maps, scholarly editions, etc.) we can increase classroom efficiency.

- ***moving beyond TDM***
  Providing the ability to analyse the many non-text components of Gale's archives, including pictures, adverts and photographs.

- ***supporting classroom use***
  To fulfil one of the most common requests, Gale will partner with leading scholars to create material to contextualise the processes in the Lab and teach with it.

- ***non-English content***
  By introducing new non-English language archives and the training tools to mine them, Gale can further enable DH in non-English native countries.

Content implications

Giving users the ability to interact with digital archives in new ways by making OCR accessible, and through initiatives like Gale Digital Scholar Lab, has raised questions about archive publishers' existing data, with potential implications for future archive digitization.

The most obvious consideration involves the quality and accuracy of OCR. OCR has always been an imperfect process, relying as it does on software to interpret often unclear historical documents. Accuracy of OCR can depend on when the archive was processed (since earlier versions of OCR software are less accurate) and of the age and clarity of the original document.

Now that OCR is more visible than ever before, Gale is working with several academic groups to determine what choices there are to improve the quality of the underlying OCR in the digital archives. Rescanning or repeating the OCR process comes at a prohibitive cost, but we want to determine whether there are systematic or crowd-sourced solutions to such a significant issue across all of DH.

Similar questions exist around metadata. By giving scholars the ability to enhance metadata, it would greatly increase the number of research questions answerable through digital archives.

Probably the most common customer request involving the Lab is to make content hosted by the institution available for analysis through the Lab. This is a huge and complicated project, simply because of the vast range of types of content hosted by universities and libraries around the world, not to mention inconsistencies in OCR standards, metadata and document format. However, given that this is an obvious area of need for institutions, Gale is working on options to make it available in the Gale Digital Scholar Lab.

Supporting and working with DH

The increase in collaboration with the DH community is set to continue. Throughout the world, Gale is working to contribute to the community, increasing visibility for academics and software developers, with a goal of acting as partner, not solely a software provider.

Bringing academics in house

Since 2017 Gale has employed academics in the US as DH specialists with a brief to advise on development, support new customers to the Lab and help contextualize DH processes for research and teaching around the world.

Alongside their work for Gale, Dr Sarah Ketchley (University of Washington) and Dr Wendy Perla Kurtz (UCLA) teach DH courses in their institutions. Ketchley first offered introductory DH classes for undergraduates and graduates in 2015, integrating the Gale Digital Scholar Lab into her syllabus in late 2018. As a cloud-based platform, with no local software installation requirement, the Lab is an ideal platform for classroom use. The class featured 35 undergraduate students from 21 departments across campus, the majority with no prior experience in DH, working in teams to create and curate content sets, clean OCR text, and then analyse their collected research material using the digital tools incorporated in the

Lab. This structured workflow presents opportunities to teach digital project management, data curation, the process of creating OCR text and the challenges of working with it. The course was again being offered in the summer 2019 quarter in an entirely online format, for which the Lab is well suited. Its suite of digital tools generates in-depth discussions about the nature of text mining, qualitative vs. quantitative analysis and the types of research questions that can be asked and answered by topic modelling or named entity recognition, for example. In lieu of a final research paper or exam, students exported the results of their research and analysis in the Lab, including primary source document images, OCR text and visualizations, to build digital exhibits in Omeka and interactive narratives in StoryMapJS, both third-party applications to publish and visualize digital projects.

'In lieu of a final research paper or exam, students exported the results of their research and analysis'

Sensitivity to DH community ethics

As the relationship between commercial vendors and academia becomes more involved, we are acutely aware of the ethics of the DH community, namely the aspirations for data and software to be open and research to be freely available. Given the irreconcilable fact that our digital archives exist for universities behind a paywall, we are always working to make sure that the data is as available as possible, and several recent research projects have relied on Gale providing specific aspects of archives not commonly available. The Lab is a good example of Gale's desire to be as open and supportive as possible within the contractual boundaries of our agreements with source libraries, giving users the ability to export OCR, statistical analyses and visualizations at all steps of the workflow.

Development partnerships

One positive outcome of Gale's increasing visibility in the DH community is the increase in opportunities to actively work together with academics and research groups. In the pipeline are numerous joint partnerships exploring OCR correction, tool creation, development of pedagogy and many others. The ability to support and amplify innovative work is paramount for us, and the opportunity to (for example) develop cutting-edge tools to analyse Gale archives is too good to miss.

'increase in opportunities to actively work together with academics and research groups'

Evolving academic events

Publishers, vendors and other content providers traditionally sponsor and exhibit at academic conferences and other events. The focus on collaboration and openness driven by DH has prompted an added emphasis in events for Gale and, in the past year we have started to organize our own events to bring together academics, developers and librarians from around the world. In November 2018 Gale invited library directors from the leading Chinese universities to an 'Advanced Workshop of Digitization, Libraries and Digital Humanities' in Dali, China. Then, in December 2018, Gale Japan welcomed 65 scholars to 'An Invitation to Digital Humanities' at the Tokyo International Forum.

In May 2019, Gale brought approximately 100 European academics, librarians and students to the British Library to hear talks from a panel of distinguished international speakers for the inaugural Gale Digital Humanities Day.[9] The day was designed to provide insights into all aspects of DH, incorporating academic research sessions (Literature and Distant Reading and Computers Reading the News), teaching sessions (Digital Humanities in the Classroom), and sessions discussing institutional considerations (Institutional Support and Infrastructure for Digital Humanities). The international panel featured speakers from the US, UK, Netherlands, Japan and Australia, and one of the most striking points was how much similarity there was in approaches, methods and challenges.

These events all featured academic speakers to provide a forum for knowledge exchange. For Gale, these events launch partnerships and are as useful for our education as that of the academic community, with events often including associated focus groups.

Conclusion and future plans

Numerous projections of future job trends, including this from the World Economic Forum[10] (see Figure 2), indicate that analytical skills and the ability to work with large data sets will continue to grow in importance for jobs in the future.

Today, 2018	Trending, 2022	Declining, 2022
Analytical thinking and innovation	Analytical thinking and innovation	Manual dexterity, endurance and precision
Complex problem-solving	Active learning and learning strategies	Memory, verbal, auditory and spatial abilities
Critical thinking and analysis	Creativity, originality and initiative	Management of financial, material resources
Active learning and learning strategies	Technology design and programming	Technology installation and maintenance
Creativity, originality and initiative	Critical thinking and analysis	Reading, writing, math and active listening
Attention to detail, trustworthiness	Complex problem-solving	Management of personnel
Emotional intelligence	Leadership and social influence	Quality control and safety awareness
Reasoning, problem-solving and ideation	Emotional intelligence	Coordination and time management
Leadership and social influence	Reasoning, problem-solving and ideation	Visual, auditory and speech abilities
Coordination and time management	Systems analysis and evaluation	Technology use, monitoring and control

Figure 2. Comparing skills demand: top ten in 2018 vs. 2022

Increasing numbers of universities are turning to DH as a method of providing humanities and social science graduates with these desirable analytical skills, providing experience to help them thrive in a rapidly changing job market.

These changes in universities challenge publishers to evaluate their archives and the various ways to explore and interact with them. Changing the fundamental ways of using content requires real engagement with the academic and library communities on numerous levels in order to deliver solutions that address real-world problems.

DH is a fascinating, complex and exceptionally diverse field that is simultaneously challenging and enthusing Gale. It can feel like a bold move to expose metadata and OCR through platforms like *Gale Digital Scholar Lab* because it exposes us to questions about their nature, format and quality. However, by taking this step, Gale has begun numerous conversations with academics and institutions around the world about potential solutions for the long-standing problems of digitizing historical documents. The possibilities, even in the relatively small area of OCR correction and remediation, are incredibly exciting and we anticipate seeing substantive improvements in this area as we begin to partner with academics around the world on OCR projects.

> 'there is a huge appetite to research and teach DH, and for … solutions to evolve to meet the challenges faced'

In terms of software, the possibilities for developing software to support DH research and teaching are no less extensive and exciting. Every interaction with an institution confirms that there is a huge appetite to research and teach DH, and for Gale's solutions to evolve to meet the challenges faced.

Future developments will see *Gale Digital Scholar Lab* grow to support teaching through pedagogical support; include local content upload to allow researchers to ingest their own content to use in the Lab; and increase the range of tools to support as wide a range of analyses as possible. For Gale, there is real opportunity in developing closer relationships with academia to fulfil our primary aim of advancing knowledge through a detailed exploration of the past and promoting opportunities for this kind of research as widely as possible.

In the future, we will continue to work as closely as possible with academics while being respectful to the ethics of the DH community. By collaborating on building new platforms and pathways, we remain committed to advancing humanities scholarship and to amplify it by supporting the global academic community.

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: http://www.uksg.org/publications#aa

Competing Interests

CH declares that he is employed by Gale. SK has no competing interests.

References

1. "Help for Researchers," *British Library*, http://vll-minos.bl.uk/reshelp/findhelprestype/catblhold/estcintro/estcintro.html (accessed 20 September 2019).
2. *Eighteenth Century Collections Online*, Gale, https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online (accessed 20 September 2019).
3. *Gale Primary Sources*, Gale, https://www.gale.com/intl/primary-sources (accessed 20 September 2019).
4. Kat Gupta, *Representation of the British Suffrage Movement* (Bloomsbury Academic, 2015).
5. Kat Gupta, Mixosaurus, http://mixosaurus.co.uk/publications/ (accessed 20 September 2019).
6. Dallas Liddle, "Reflections on 20,000 Victorian Newspapers: 'Distant Reading' The Times using The Times Digital Archive," *Journal of Victorian Culture*, 17, Issue 2, (1 June 2012): 230–237, DOI: https://doi.org/10.1080/13555502.2012.683151 (accessed 20 September 2019).
7. Exploring ECCO: Key moments in 18th-century philosophical literature. Eetu Mäkelä, Vili Lähteenmäki, Antti Kanner, Ville Vaara. Never Mine the Mind? Symposium on Computational Approaches to Intellectual History and the History of Philosophy. Helsinki, 31 May 2017 (slide 2), https://comhis.github.io/assets/files/Never_Mine_the_Mind__COMHIS_Collective.pdf (accessed 20 September 2019).
8. "Gale Digital Scholar Lab," Gale, https://www.gale.com/intl/primary-sources/digital-scholar-lab (accessed 20 September 2019).
9. "The Gale Review," Gale, https://www.gale.com/intl/blog/2019/05/09/gale-digital-humanities-day-at-the-british-library/ (accessed 24 September 2019).
10. *Forecast of Jobs Report 2018*, World Economic Forum, http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf (accessed 20 September 2019).

Corresponding author:
Chris Houghton
Head of Digital Scholarship, International Gale Primary Sources
Gale, A Cengage Company, UK
E-mail: chris.houghton@cengage.com
ORCID ID: https://orcid.org/0000-0002-5544-8158

Co-author:
Sarah Ketchley
ORCID ID: https://orcid.org/0000-0002-0803-4147

To cite this article:
Houghton C and Ketchley S, "From provider to partner: how digital humanities sparked a change in Gale's relationship with universities", *Insights*, 2019, 32: 30, 1–10; DOI: https://doi.org/10.1629/uksg.482

Submitted on 09 August 2019 Accepted on 30 September 2019 Published on 16 October 2019

Published by UKSG in association with Ubiquity Press.