Gale Digital Collections: Ray Abruzzi Interviewed by Luisa Calè and Ana Parejo Vadillo

Ray Abruzzi, Luisa Calè, and Ana Parejo Vadillo

This interview addresses the commercial dimensions of the nineteenth-century digital archive. Luisa Calè and Ana Parejo Vadillo ask Ray Abruzzi, Vice President and Publisher for *Gale Digital Collections* at Gale, about the company's origins, its commercial approach to digital collections, and the challenges of digitization. In the context of the open access movement, the architecture of participation, and crowdsourcing, Abruzzi discusses how the company works with academic partners and interfaces with other digital libraries and platforms.

What were the founding principles of Gale, and how has its mission changed?

Gale was initially founded to answer a question: 'What are the local businesses in my area and how do I contact them?'. In the process of trying to answer this question, local Detroit businessman Fred Gale Ruffner realized that no comprehensive sources for this information were available — and then he set out to collect and publish this information himself. From this was born the *Encyclopedia of Associations*. Recognizing a talent for gathering and publishing information, further directories and encyclopedias of this nature followed. The move into the academic space was gradual. Drawing on his interests in literature and history, Ruffner commissioned authors and scholars to create the *Dictionary of Literary Biography* and *Contemporary Authors*. From there Gale went on to publish numerous reference and encyclopedic works under imprints including Charles Scribner's Sons, Macmillan Reference USA, Schirmer, Twayne, Thorndike Press, and others. Gale also began publishing microfilm via Primary Source Media, Research Publications (Inc.), Scholarly Resources, Lost Cause Press, and working with NARA (US National Archives) and the Library of Congress as a distributor. Gale eventually went digital, of course, and so we started to aggregate journal content alongside the secondary sources we were publishing — and that is when we began to digitize our primary sources collections, which we previously had on microfilm. We first published the archive of the *Times of London* (*TDA*) and followed this with *Eighteenth Century Collections Online* (*ECCO*). But Gale has not moved too far from that original idea today — creating resources that help answer questions. Whether it is digitized archival materials, databases that synthesize primary and

secondary sources for students, or journal aggregations, Gale's primary mission is to help people answer questions and pose new ones. Our official credo is: we believe in the power and joy of learning. We are enriching relationships and advancing the way you learn and research. We are touching the lives of millions of people globally every day — one person at a time.

With our archives (*Gale Digital Collections*), this mission is expressed by providing content — and now the 'data' behind that content — in ways which help students and researchers pose and answer questions against 250 million page-images of primary sources spanning seven centuries.

How do you see Gale in relation to other digital libraries?

Gale is a 'digital library' when viewed in the aggregate of its component products, but Gale publishes up and down the educational chain, from early literacy education products like *Miss Humblebee's Academy* to hardcore research collections like *State Papers Online*. As such, this material is not all discoverable in one place — that would not make sense for users or for their needs, and hence, we are not quite a 'digital library' in that way — perhaps we are 'many' digital libraries, each one suited to the needs of a particular set of users.

With *Gale Digital Collections* (*GDC*) we partner with over three hundred libraries, archives, commercial publishers, and other content repositories around the world to create unique digital collections which meet the needs of students, faculty, and researchers, while satisfying the collection principles of academic libraries. *GDC*, especially given that *Artemis: Primary Sources* does unite all (almost all!) of this content on one research platform, can be considered one of those many Gale digital libraries.

Acknowledging that Gale is a commercial publisher, *GDC*'s major difference from other digital libraries, public and commercial, is in the way we present our content. As the body of open access material grows and as our competition in the commercial space attempts to displace Gale as the leading humanities publisher, our driving principle is to present content in ways that tie directly to educational and research outcomes — to seamlessly deliver content in ways that match the workflows and goals of our users. Making content available is not enough; making it available to 'me' in ways that suit 'my goals' is the key to value and to making archival digitized content not just available but engaging. Useful.

From Google to other articulations of the cloud library, the open access movement advocates a digital library without paywalls. How does Gale engage with these initiatives?

Gale makes our metadata available to many organizations, from library discovery services to Google, DPLA, OCLC, CRL, and others. There is no illusion, here, to be clear. Gale is a for-profit company, engaged in digitizing archives to drive revenue goals. That mercenary bit out of the way, we work closely with the academic community and we offer several value propositions:

- Funding: we digitize content that libraries and archives often cannot (lack of funding, infrastructure, expertise, etc.).
- Preservation: we pay for conservation — analogue content is given back in better physical condition than when we encountered it.
- Surrogates: users can now use facsimiles as opposed to analogue content, unless direct access is appropriate, saving wear and tear on the documents — particularly newspapers and unique manuscripts.
- Access: through Gale, content from 'your' archive is now available simultaneously around the world, enabling new levels of simultaneous research and scholarship, which extends to students.
- Metadata and cataloguing: Gale provides detailed metadata and improved bibliographic records back to every archive we work with.
- Editorial: Gale works with library partners around the world, not working with just what is available or on hand in one institution or region or country, but actively matching up archives from around the world to create globally useful resources, saving time and travel budgets for the academic community.
- Discoverability: beyond Gale's own marketing and sales efforts to publicize collections, Gale also works with all of the major discovery engines to ensure that interested parties can find relevant content.

What economic models do you think will be needed or can be implemented to sustain nineteenth-century digital archives and the labour underpinning them? If it is important that such resources be open to all, who should pay?

Gale believes — well, *I* believe, to be precise — that many of the most useful and sustainable digitization projects undertaken have been funded by commercial entities like Gale simply because of the investment required and the resources available to put those investments to work. That is not a mark of disrespect to any of the great efforts that continue to grow in the open access world. As you know, even in an open access collection — someone, somewhere, paid to create that collection and someone needs to maintain that collection. It is not simply 'free' and you can see that a lack of ongoing investment or sustainable funding has had negative effects on some of the older open access digital resources. They are growing clunkier, outdated,

and some are simply disappearing, having been built on outdated technologies and due to the lack of investment and funding to support upkeep and updates. I also believe that the sheer focus of economic gain means that the end products Gale produces each year will be valuable to the academic community and sustainable as a commodity held by a business. We keep our interests alive and continually improve on them to make sure they are relevant and discoverable. Speaking plainly, this means that Gale:

- Delivers digital images, OCR, and metadata to all of our partners for their content.
- Often enables those partners to make those digital images immediately available for free, onsite. (We never seek to restrict access to the analogue materials.)
- Puts terms in place that allow every partner to make those images and metadata available for open access worldwide within reasonable timeframes.

Gale does not tie up digitized content in perpetuity, and we sign agreements that enable our partners to work with open access repositories after the exclusivity period, as they see fit. I do not know how to create a sustainable open access model that will not require some form of shared costing between the users and those who maintain the collection. There are many models in place but all of them have some type of funding that created them and will need some funding source to keep them maintained and relevant. I do not believe that 'free for everyone' is a working model for sustainable publishing. Continued investment is necessary to keep digital content not just 'alive', but available in ways that adapt to changing technologies — and changing methods of research.

What do you think are the most exciting emerging or potential practices around the nineteenth-century digital archive, and what social or technological barriers, if any, are hindering them?

The most exciting thing in the world of *Gale Digital Collections*, *NCCO* included, relates to textual analysis and data mining (TDM). Gale announced that the data for our *GDC* programme would be made available to libraries almost a year ago, and since then we have gone beyond that solution to improve the TDM tools available through our *Gale Artemis: Primary Sources* platform, including Term Clusters and Term Frequency. We are about to announce a new, cloud-based data 'playground' which will feed into the many ways digital humanities (DH) researchers want to access and analyse historical works as data.

Some of the barriers in DH right now are in expertise and infrastructure — it is not a light task for a library to grow a Digital Humanities Centre or other similar resource to support research and scholarship in this area. Training and investment are necessary to support DH activities in the

emerging fields of digital research, and not every library is supported in their efforts to build up resources in this area.

Digital knowledges

How do you determine what to digitize? What drives your different digitization projects?

Gale's background and expertise as a publisher uniquely positions us as a leader in the digitization of historical works. We work with advisors and editors from the research and curatorial worlds to identify collections of interest, and then align collections across repositories to fit curricula and to advance research. Our products are driven by the same drivers in academia — Gale looks for opportunities in what is new, what is interesting, and what is going to move ideas forward — the only difference is that our projects are commercial products. They are also *timely* products. Gale is there to support growing fields and emerging sections of research within the humanities, and we are 100 per cent committed to digital humanities as both a leader and a participant.

How do you create a product? (I am thinking of packages of different digitized nineteenth-century materials.)

The driving forces around creating *NCCO* modules were derived from the fields of research themselves. There are so many disciplines studying/flowing through the nineteenth century that at best we could only be 'provisionally comprehensive'. Gale recruited a board of eminent scholars and archivists to help us determine the key topic we wanted to cover and we sat down in a room for a few intensive days, and we came up with a plan.

With *ECCO*, we had the Eighteenth-Century (now English) Short Title Catalogue (ESTC) as a bibliography to work from. For *NCCO*, the Advisory Board quickly realized that the Nineteenth-Century Short Title Catalogue (NSTC) was not enough (as it relates to the goal of this project), and that we would need to break down the archives into digestible but meaningful pieces. We also realized that building this resource would take several years (not an easy sell to a commercial publisher).

The noted scholars and bibliographers on the Advisory Board include:

Hilary Fraser
Birkbeck, University of London
Executive Dean, School of Arts
Director, Birkbeck Centre for Nineteenth-Century Studies

Tatiana Holway
Independent Scholar, Author, Researcher, and Editor
Specializing in nineteenth-century social sciences

Dominique Kalifa
University of Paris 1 Pantheon-Sorbonne
French Historian and Professor
Head of the Doctoral School of History and Director of the Centre of 19th Century History

H. K. Kaul
Founder and Director of DELNET- Developing Library Network (India)

Jerome McGann
University of Virginia
John Stewart Bryan University Professor
Founder and Co-Director of Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES)

Joris Van Eijnatten
Utrecht University
Professor of Cultural History
Chair of the section 'History of Culture, Mentalities and Ideas since 1500'

That plan put *NCCO* into over one hundred libraries worldwide in six months, with digitized content from over one hundred libraries and archives, and eventually amounting to almost thirty million pages of content never before available in digital format to students around the world. While *NCCO* is no longer an active programme — we are not planning new modules at this time — libraries continue to acquire this product, and there are over two hundred institutions with access to it around the world. *NCCO* greatly expanded the resources available to thousands if not millions of researchers in nineteenth-century studies, and Gale had similar success in the next two years of *NCCO*, releasing twelve archives in all to reach that thirty million page mark.

Description

What is your philosophy regarding the process of digitization, for example, for your collection of nineteenth-century periodicals or nineteenth-century manuscripts?

Our philosophy goes back to Fred Ruffner — answer questions, solve problems, and build a resource that is more useful than the current resources available.

Is there a quality control of the product? What does that entail?

Gale has developed a sophisticated quality control system that allows for quick and accurate review of all images and metadata:

- All XML is parsed/validated against the Gale DTD.
- Fixed XML values are validated against lists of predefined control lists.
- Internal quality control operators view every image and metadata element.
- Image quality is inspected.
- Title level metadata is reviewed to ensure strict adherence to capture requirements.
- Every captured page element is reviewed against the source image for accuracy.
- Image coordinates in support of hit-term highlighting are spot checked for accuracy.
- Rejected pages are sent back for rework and then put back in the queue for review.

What projects are you currently working on?

The team behind *Gale Digital Collections* are developing a number of new products, some of which still have working titles, but here is the current view:

- Archives of Human Sexuality and Identity:
  - Part I: LGBTQ Activism and the HIV/AIDS Crisis, 1940–200X (March 2016)
  - Part II: Erotica: 18th and 19th Century Erotic Literature (tbd)
  - Part III: LGBTQ: Hidden Archives, 18th–20th Century (tbd)
- American Fiction, 1774–1920 (March 2016)
- US Declassified Documents Online (1 million pages, with an additional 5000 documents annually) (December 2016)
- Brazilian and Portuguese History and Culture Online: Oliveira Lima Library Monographs (March 2016)
- Early Arabic Printed Books from the British Library (1475–1900) (December 2016) — this is a revolutionary product in that we have managed to OCR early Arabic!
- The Telegraph Historical Archive, 1855–2000 (December 2016)
- China from Empire to Republic: Missionary, Sinology, and Literary Periodicals, 1817–1949 (March 2016)
- 19th Century British Newspapers, Part V (March 2016)

We just released two collections important to nineteenth-century studies: Crime, Punishment and Popular Culture, 1790–1920; and Smithsonian Collections Online: The Evolution of Flight, 1784–1991.

What is your most used archive?

Most used and 'most impactful' are different questions. The straightforward answer is the *Times Digital Archive*, but the most impactful — so far — has to be *NCCO* — we have changed the scope, depth, the very dimensions of research possibilities with this programme. References and citations to *NCCO* show up every month, though it is tricky to track; students (and even

those who should know better) cite the physical documents and not the database, so it is hard to measure citations to *NCCO* itself.

What were the challenges of digitizing nineteenth-century manuscripts?

Hmm. They were legion! The main challenge around digitizing nineteenth-century manuscripts is the condition of the source material. Manuscript content has many characteristics that need to be considered before digitizing. Deterioration, tight bindings, water/fire damage, and the presence of mould are some of the challenges found in this material. We also often find oversized materials, which can be time-consuming to capture due to the handling of such large items.

Another challenge is the quality of the existing metadata and bibliographic records. Depending on when the material was accessed by a library or archive, how much time and resources were available for cataloguing, and many other factors, manuscript collections can often have very sparse catalogue records. Gale spends a lot of time and resources — working with our library partners — just verifying what is and is not in a collection, as part of the digitization process.

A good example: Lord Chamberlain's Plays from the British Library (part of *NCCO: British Theatre, Music and Literature, High and Popular Culture*). We estimated the digitization of that collection about thirty weeks' time, assuming a certain number of documents per week, and that we would capture titles, authors, and dates from each play. We also accounted time for extensive conservation work. We *vastly* underestimated the amount of conservation needed. In reality, this took two years and we spent much more time and money, and it required more expertise and effort (qualified and quantified, in the end), than we ever planned. The end result, though, was the most beautiful and amazing record of nineteenth-century British theatre that has ever been created. In conjunction with the other component collections of that archive, it is the most amazing resource I have ever worked on . . . so far.

What technologies have you been using?

On the scanning side — capturing page images — we use three types of technologies based on the type and condition of the source material. Flatbed scanners are used for loose leaf, and book scanners with cradles that support the book are used to scan most monograph material. We also are big proponents of a third technology, which is using Kirtas/auto-book-turning capture. Our internal tests have shown that this equipment, when used in manual mode, is very gentle on the material while providing excellent quality scans.

On the conversion side we are using the latest version of ABBYY for OCR. We have worked closely with them throughout the years to tweak

their engine for the types of material we capture. We have also implemented a workflow solution for capturing names, places, and dates from handwritten manuscripts. This allows us to add a new degree of value to researchers, giving them the ability to discover and research this valuable content more easily. We have also seen the citation of manuscript materials increase since adding this new depth to the searchability of the material.

What has been the aim: visibility? accessibility? preservation?

The aim of products like *NCCO* touches on all three of those points, but to that list I would add another layer to the point around access. In the past, using a collection like Lord Chamberlain's Plays was the privilege of established scholars, faculty, and particularly promising graduate students. Getting access to a unique, valuable, and fragile collection of manuscripts was no easy feat. It took a certain level of 'clearance' from the institution, it took money to actually travel to that institution, and it took a lot of time to work through the boxes, folders, and documents to find what you were looking for — and then to do your actual research. All of that has been greatly sped up, and what's more, students and researchers of any level and at any point in their academic career can now access these documents. Access has been significantly democratized through commercialization — acknowledging again that not every school will own these collections.

Geographies of the nineteenth-century digital archive

Do you see different digital cultures (and economies) regarding nineteenth-century materials emerging? For example, the US, Europe, Asia, etc.

Yes, some countries certainly prioritize nineteenth-century studies over other fields of research, while others favour STEM content. Recently, we are seeing emerging interest in nineteenth-century studies in the Middle East and Asia, particularly for local-language content.

Is the nineteenth-century market global? How do you deal with the specificities of cultures and nations?

Yes, there is worldwide interest but it can be quite regional and specific, in terms of the materials they are seeking for research and analysis. That said, collections like *ECCO* and *NCCO*, and periodical archives such as *National Geographic* and *The Economist* travel quite far, in terms of interest and acquisition by libraries.

Architectures of participation

In 2004 Tim O'Reilly defined Web 2.0 as 'the architecture of participation', a system 'designed for user contribution'. What is the philosophy of Gale?

As a commercial publisher, Gale does not open its collections for user contributions. On our research platform, *Artemis: Primary Sources*, we do offer the ability to log in, as an individual or as a group (say, a study group) for project-based research, and users can add tags and annotations, save and share documents, etc. As we are publishing primary sources, we do not allow for users to add secondary analysis or commentary. I do not believe our customers would be keen for that, either. There are other outlets for analysis and commentary, of course — and, to be frank, very few use the available option to add public tags to our collections.

How does Gale interface with other platforms?

Gale Digital Collections are indexed on the various library products for discovery, and, in at least one case, are directly cross-searchable with other products from other publishers. I have been trying to bring other publishers to the table to discuss broadening the range of cross-searchable primary source collections but I have met with little success. Gale's door is still open for those publishers. It is my strong belief that researchers and students (not to put too fine a point on that distinction) would benefit greatly from being able to search related resources in a research atmosphere geared to primary sources as opposed to having them part of a much larger discovery service which often does not acknowledge the nuances of primary sources and is not well suited to this kind of digital content.

Dino Felluga, in another contribution to this issue on the nineteenth-century archive, argues that the notion of 'crowdsourcing' is associated with 'outsourcing' and advocates a platform that encourages 'insourcing'. What are the constraints and possibilities of collaborative work at Gale?

Gale Digital Collections are resources for research, education, and other forms of scholarship — they are not ends but means. Outside of the actual collections — around them, if you will — Gale supports numerous research projects and has been involved in multiple grants, projects, and initiatives, providing content and expertise in support of various academic objectives.

Does it bother you that scholars use your collections but only quote from the text they are using? Do you fear in that sense invisibility?

Yes! I've had scholars look me in the eye and tell me that, even though they know it is incorrect, they cite the document and not the sources because it 'looks more scholarly'. This is the case even when I know very

well there is only one copy of a particular document in the world and this particular scholar has never been to see it! I try to explain that one metric librarians (those who make decisions about acquiring resources) measure a product by is the number of citations to the databases themselves. When they are not cited properly, that resource is losing value — and the acquisition of further resources at that library is less likely. We teach students to cite sources properly — it is baffling when the teachers themselves ignore their own instructions. (That is not to say this is always the case, to be sure.)

Digital change, transformation, and obsolescence

There has been a lot of anxiety about the obsolescence and unsustainability of our digital ecologies. How do you see the future of Gale given the current pace of technological innovation?

Gale has a number of measures in place to ensure sustainability and access — even in a world without Gale. First, we keep multiple copies of the files and images associated with our archives in multiple locations (the LOCKSS principle). We keep several copies in offices in Connecticut and in Michigan; we ship copies of each collection back to the original/source institution; and, most importantly, we keep a 'dark archive' with PORTICO, who updates the files as needed to keep pace with new technologies. In the event of a Gale apocalypse, PORTICO would be able to provide the content to our customers and non-customers alike. Additionally, our customers can purchase (at a nominal cost) hard drives of the files and images in the collections they own, and hold local copies for Text Analysis and Data Mining, as well as for preservation.

Without belabouring the point, preservation and access is an area where being a commercial publisher means we invest to keep on top of the archives — we continually improve the means of search and access, and we continually keep our products relevant, because they are commercial assets.

What technologies are you experimenting with? Which ones have you rejected and why?

We are particularly excited about TDM. We are currently developing a service, *Gale DH*, which we are hoping develops as follows. *Gale DH*:

- Hosts the data behind *Gale Digital Collections* in the cloud, where students, researchers, and faculty can access and/or download the full text and metadata from collections such as *ECCO*, *TDA*, *US Declassified Documents Online*, and the many other archival collections currently available on *Artemis: Primary Sources*.

- Subsets of relevant data can also be accessed or downloaded using a simple interface similar to an Advanced Search page.
- Supports API interactions, and the data itself is available in several formats to work with the most popular TDM software suites (R, Gephi, etc.).
- Users can also upload their own data sets into *Gale DH*, allowing research to flow across content providers and open data seamlessly.
- For novice users, Gale will provide data normalization software which enables researchers to manage disparate data sets within one tool or TDM interface/software set/language.
- Finished projects can be archived and accessed via a *Gale DH Commons* at no cost.

What are the emerging digital experiments that excite you today?

As far as other kinds of emerging technologies go, I am interested in tools related to concordance, and to content-based image recognition, which is used to properly identify uncaptioned photographs and other images which are not well catalogued and have no attached guides or metadata.

NCCO as a programme ended in 2013, but the new *Crime* archive would most certainly have been a continuation of *NCCO* if it had not. My team and I are working on several new product ideas that would also be in the same vein as the existing *NCCO* archives, and will relate directly to many research areas in nineteenth-century studies, and we certainly welcome ideas and suggestions from those engaged in all aspects of the field — librarians, curators, archivists, teachers, researchers, and yes, students. Gale's development process is rooted in collaboration with both our content partners and the end-users of our resources. As part of this process I spend most of my time visiting archives, speaking with librarians, and engaged with faculty and researchers, trying to imagine resources that will meet their various needs, and make them available in ways that will tie to their key outcomes. I truly do enjoy my work in this way, and what's more, when I read an article, book, or thesis based on research done using the *GDC* products, it is extremely fulfilling on a personal level. I consider myself quite lucky!