DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 1

Computer Vision and the Creation of a Database of Printers' Ornaments

Hazel Wilkinson <h_dot_j_dot_wilkinson_at_bham_dot_ac_dot_uk>, University of Birmingham, UK
James Briggs <jimy_dot_pbr_at_gmail_dot_com>, Graphcore
Dirk Gorissen <dgorissen_at_gmail_dot_com>, Machine Learning Ltd.

Abstract

This article describes the creation of a database of over 1 million images of eighteenth-century printers' ornaments (or fleurons) using computer vision, and how the database was refined using machine learning. The successes and limitations of the method used are discussed, and the purpose of the database for research in the humanities is outlined. The article concludes with a summary of recent developments in the project, which include the addition of a visual search function provided by the Seebibyte Project.

1 Throughout the hand press period (roughly from 1440–1830), many printed books were decorated with printers' ornaments and ornamental type.[1] These have all but vanished from modern books, with the exception of the small number still printed on private hand presses. The term "printers' ornaments" encompasses wood- and metalcut blocks, where designs were cut by hand (see [Maslen 1974], [Goulden 1988], and [Ross 1990]) as well as cast metal blocks, and copies of hand-cut woodblocks made from molten type-metal (a process known as dabbing, recently investigated by [Mosley 2015] and [Bergel 2016]). Books were also decorated with fleurons, or ornamental pieces of type (also known as printers' flowers), which could be used individually for a tiny flourish, or arranged into larger, complex patterns. The history and use of fleurons has been documented by [Meynell 1923], [Baines Reed 1952], and [Ryder 1972]. Ornaments and fleurons were used to decorate title pages, the beginnings of chapters or sections (headpieces), blank spaces between sections (tailpieces), and initial letters (using blocks, arrangements of fleurons, or factotums).[2]

2 Printers' ornaments of all kinds are useful as evidence in bibliographical studies. Those cut by hand are unique, and although casts were reproducible, they could develop unique patters of wear or damage over time. This means that in the right circumstances ornaments can be used to identify when, where, and by whom a book, pamphlet, or other printed item was produced. Books and printed ephemera did not always include the name of their printer, or the date or location of printing. Sometimes these details were omitted to conceal the printer's identity, in the case of a controversial work or pirated printing. Most books include a date and location, but often the publisher's name(s), and the names of the booksellers who stocked it, are given instead of the printer's (the publisher was the book's financial backer and marketer). Pamphlets and leaflets are less likely to have complete production information than books. It was very common for a printer to put his or her name to some but not all of their output. A representative example is the eighteenth-century printer John Darby II, who put his name to approximately 230 books between 1707 and 1732, but probably printed over 1000 items during this period.[3] If the items signed by a specific printer contain ornaments, however, we begin to attribute unsigned items. If an ornament is uniquely identifiable (not a cast, or a cast with unique wear), and it can be found in multiple books with a known printer, we can infer that it belonged to that printer. When it appears in an item with no known printer, then, we can attribute the item to the printer who owned the ornament. When making printer identifications in this way, it is good practice to find as many examples as possible, since occasionally printers lent each other ornaments, or shared printing jobs with one another. When several ornaments known to have been owned by a single printer appear in an un-signed item, however, we can identify the printer with some certainty. If an item is undated, or the location of its production is unknown, the identification of the printer can usually offer some useful information. Fleurons can also help in printer identification, using a slightly different method. It is normally

impossible to distinguish between multiple casts of the same fleuron, but the arrangements into which they were combined can be identifiable, because more complex arrangements were sometimes kept as standing type and used in multiple books. When this occurred, they are good indicators of when and where an item was printed, and can potentially date items very exactly, as shown by [Wilkinson 2013]. It is possible that the identification of a book's printer, or the date or location of printing, can also help us to identify unknown authors, or confirm or refute suspected authorship.

3 Printers' ornaments and fleurons are potentially very useful as bibliographical evidence. They are also neglected resources for the study of graphic design and book production. Their role in the reading experience during the hand press period has received very little attention. Samuel Richardson's novel *Clarissa* (1748) is the only book to have attracted serious attention for its fleurons, in particular from [Barchas 2003, 119–22], and [Toner 2015 , 68–95], largely because Richardson himself was a printer of particular creativity. Despite their potential as evidence and their interest for research, printers' ornaments are seldom used because they have been difficult to locate. Unlike illustrations, the presence of ornaments is not routinely noted in library catalogues, and they have not been tagged in digital repositories. To identify printers using evidence from ornaments and fleurons, it has been necessary to search books page by page. The mass digitisation of books has made searching page by page more convenient, but it is still a time-consuming task. Occasionally scholars like [Maslen 1974], [Goulden 1988], and [Ross 1990] have produced catalogues of the ornaments belonging to individual printers, but these have been the work of years, or decades.

4 *Fleuron: A Database of Eighteenth-Century Printers' Ornaments* was created to speed up the process of locating and studying printers' ornaments. In 2014, Gale Cengage announced their intention to make their content available for data mining projects. Gale Cengage are the publishers of Gale Digital Collections, which includes the repository of digitised eighteenth-century books, Eighteenth-Century Collections Online (ECCO) (http://www.gale.com/primary-sources/eighteenth-century-collections-online/). Together, Parts I and II of ECCO contain over 33 million page images from over 150,000 titles (in 200,000 volumes). ECCO has an emphasis on books in English, but also includes items published throughout Europe and America, and in many languages. All 33 million page images (2TB of data) were supplied to Cambridge University on a hard drive. It was previously possible to download only 250 page images at a time from ECCO, so the provision of 33 million page images in a single location presented new possibilities for the location and extraction of printers' ornaments. This paper describes how a team of researchers used computer vision and machine learning to create a database of more than 1 million printers' ornaments from the 33 million page images supplied by Gale Cengage.

5 Computer vision has been applied to printers' ornaments before. In 1997, the library of the University of Lausanne and the Université Paul Valéry, Montpellier, created Passe Partout, a searchable database of printers ornaments, described by [Corsini 2003] (http://bbf.enssib.fr/consulter/bbf-2001-05-0073-010) It is based on software produced by the Federal Polytechnic School of Lausanne. Passe Partout is built around archives of locally-provided data, provided by research groups working on specific printing offices, predominantly in Switzerland. The images can be searched using a program called T.O.D.A.I, or Typographic Ornament Database and Identification, developed in 1996 and 2000 at the University of Lausanne, by Stephane Michel and Heike Walter, under the supervision of Josef Bigün. T.O.D.A.I uses orientation radiograms to match ornaments within the database. It is documented by [Bigun 1996], and [Michel 1996]. Passe Partout uses relatively small datasets of 200 DPI jpeg images. It is enormously useful for studies of specific printers operating in Enlightenment Switzerland (particularly in Geneva and Lausanne).

6 Recently a team at Oxford University, headed by Giles Bergel and Andrew Zisserman, developed a program called ImageMatch for Bodleian Broadside Ballads Online, a catalogue of single-sheet printed ballads of the hand press period. 900 seventeenth-century ballads were scanned by the Bodleian Library in Oxford for the project. Using these high quality scans, Bergel, Zisserman, and their team employed a bag-of-words approach to develop ImageMatch, which allows image searching of the woodcut illustrations in the Ballads, as outlined by [Bergel 2013].

7 The data from ECCO presents a very different problem from the data used by these two previous projects. There is much more data in ECCO, and it is of lower quality. The images used in Passe Partout and Bodleian Broadside Ballads Online are direct scans of books (in greyscale), whereas the images in ECCO are secondary scans of the Eighteenth-

Century Microfilm Set (in black and white). The original microfilms were made decades ago in libraries around the world. To identify and extract printers' ornaments from the images contained in ECCO required a different approach, on a bigger scale.

The problem of extracting printers' ornaments programmatically from scanned pages can be approached in different 8
ways, each with their relative trade-offs. The approach which was taken for Fleuron, which is outlined in this paper, is a morphological one. A series of morphological operations (filtering, dilution, erosion, etc.) was applied to each image, followed by a series of heuristics to filter out those connected components that are deemed to be printers' ornaments. This was possible because ornaments do not occur randomly on the page, and they have particular size and shape constraints. For example, ornaments depicting capital letters have a fairly constrained aspect ratio (roughly square) and can be found at the left-most margin of the text. Other ornaments are typically horizontally centred with the text and surrounded by white space. This helps us to distinguish them from illustrations and blobs of text that end up glued together as a result of the morphological operations. Examples of these normal ornaments are shown in Figure 1.

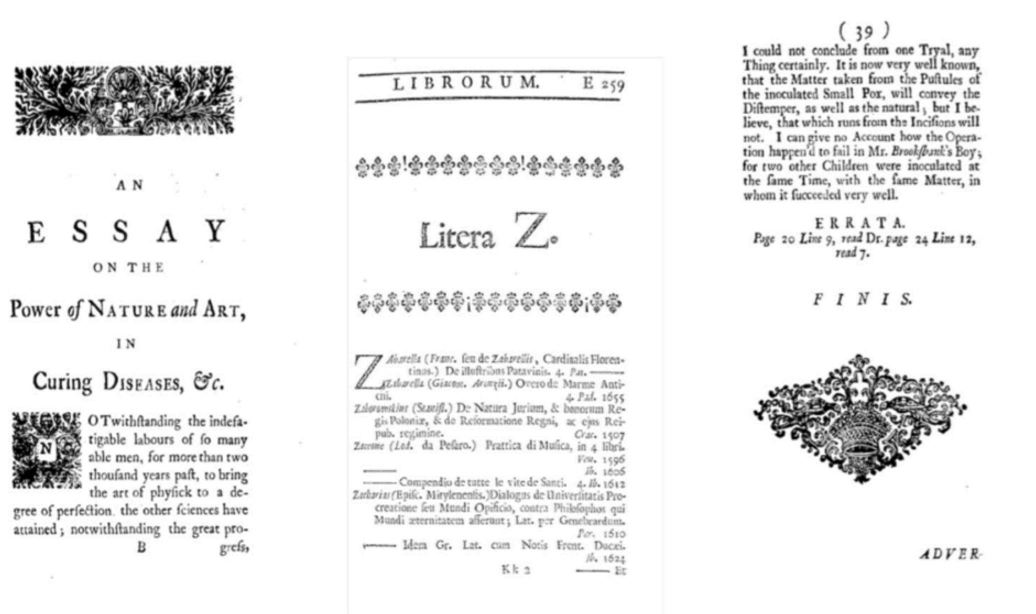

Figure 1. Three examples of typical printers' ornaments. 1a: A headpiece and factotum; 1b: Two rows of fleurons; 1c: A Tailpiece

Each page can be treated independently, so the processing can happen in parallel and the implementing code can 9
make use of all the cores on a machine. The processing was implemented with OpenCV, using the following steps. Illustrations of the process are shown in two different examples, in Figures 2 and 3.

1. Preprocessing: Conservatively cleaning up the scanned image, removing small artefacts introduced by the scanning process.
2. Threshold the image to black/white such that all white pixels are a 1 and all black pixels are 0.
3. Apply a series of open and closing morphological operators in order to remove small (white) speckles and close small (black) holes.
4. The contours of the image are found, and they are all closed; isolated contours with a bounding box area of less than 50 pixels are removed.
5. Of the remaining blobs, a rough estimate of what are just lines of text is performed, and these are removed.

6. Heavily dilate the image. This will cause ornaments that are made out of many different small separate elements to be joined together as a whole. Note this has as side effect that the letters in the text will be glued together as well. Something has to be filtered out again later.
7. Remove small, negligible contours.
8. Loop over all remaining contours and decide for each one whether it is an actual ornament, a full page illustration, a blob of glued together text, or something else. This decision is made based on a set of heuristics which were chosen to be optimistic. It was deemed preferable to minimise the chance of type two errors (false negatives) at the cost of more type one errors (false positives), in order to ensure we captured as many ornaments as possible.
9. If we are not sure whether the blob under consideration is an ornament or not, we try to break it up into smaller pieces using the original (undilated image) as a guide and recurse.
10. If, after recursing, it is still not clear what the object is, we treat it as an ornament.
11. Finally, for each ornament we extract the bounding box, merge any overlapping bounding boxes and write the corresponding coordinates to a json file.

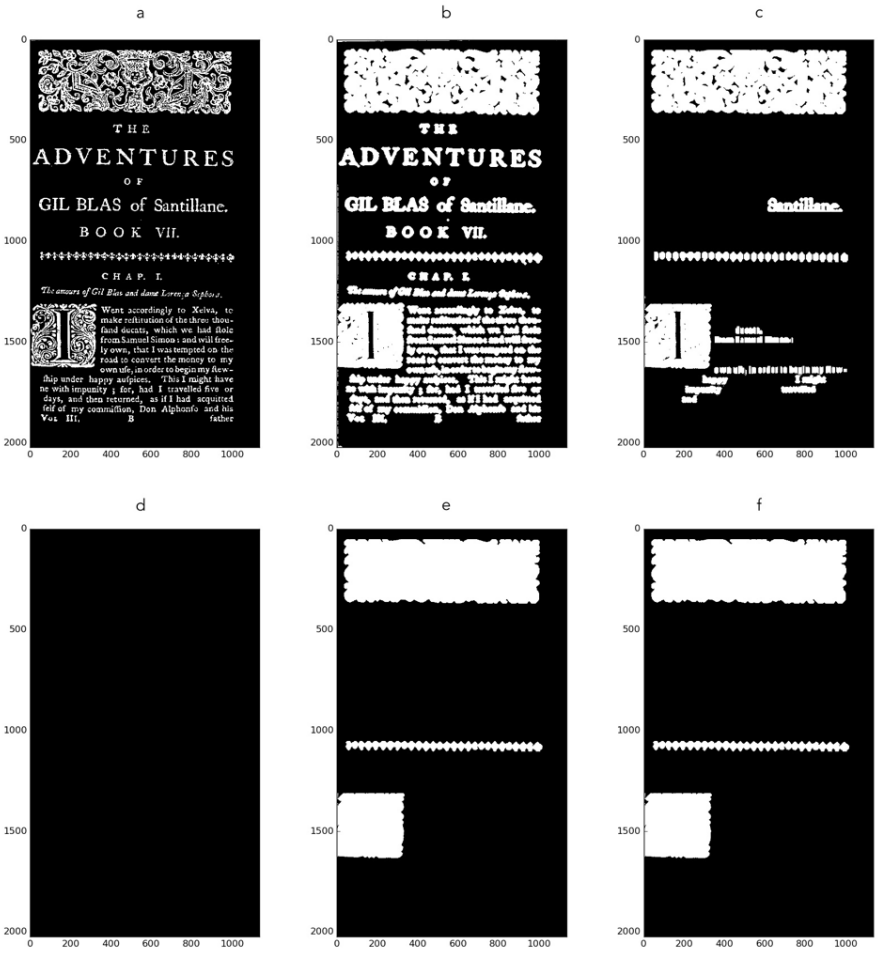

Figure 2. Illustrations of the steps described: a) thresholding to black and white; b) removing speckles and holes; c) removing text; d) heavily dilating the image; e) small elements of ornaments are joined together; f) loop over remaining contours and decide whether each is an ornament; recurse if necessary.

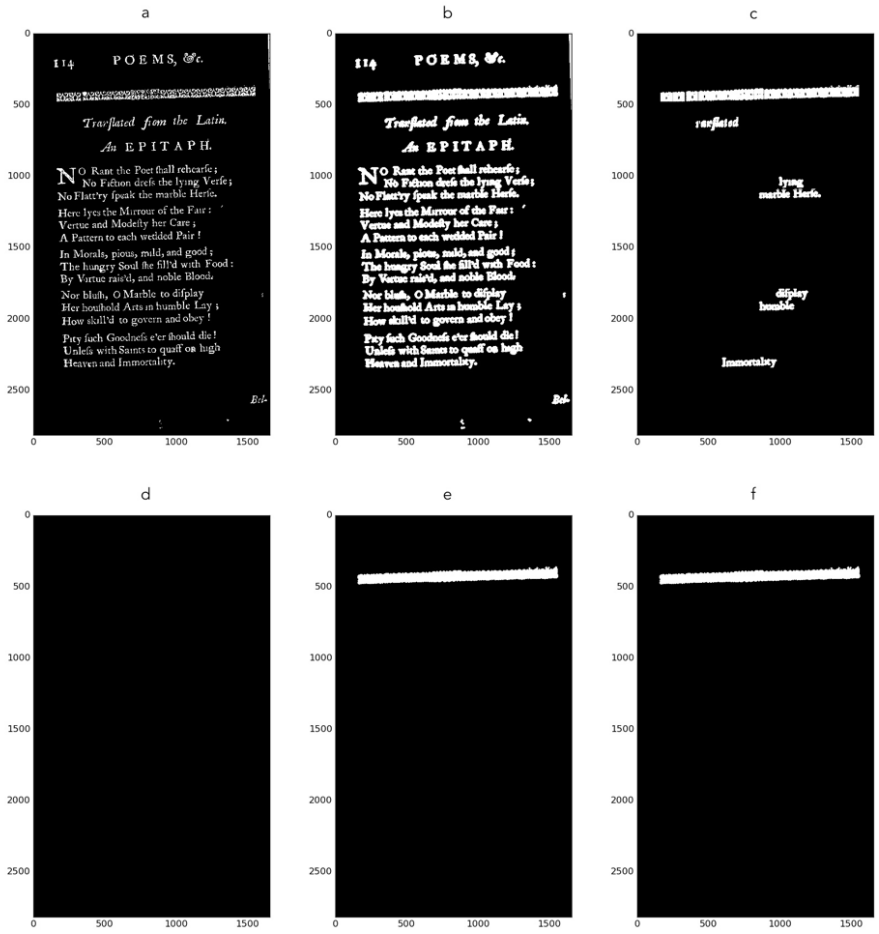

Figure 3. Another illustration of the steps described in Figure 2.

Each image contained in ECCO also has an associated metadata file, and the metadata for each extracted ornament was also written to the json file. The metadata is originally derived from the English Short Title Catalogue, so it tells us the book's title, author, date, publisher, location, format, and ESTC number (the unique identification code assigned to all books by ESTC). 10

The permissive approach that we adopted was necessary because of the presence in the dataset of images like those shown in Figure 4. In Figure 4a the decorated initial letter has no white space to the left side, because a tight binding has prevented the scanner from reaching the full margin of the page. It is also placed particularly close to the text on the right. Figure 4b features two arrangements of fleurons (which we need to capture), but also manuscript notes and significant shadowing. Figure 4c does not contain any ornaments, but it does contain several elements that might mistakenly be characterised as ornaments. It is from an almanac, where the frequent appearance of tables and charts present problems, as they disrupt the normal division between white space and content. This page also features the problem of show-through. The cheap, thin paper means that the text from overleaf is showing through to this side, creating a blurred area that may be mistaken for a non-textual element, hence the code's specification that an ornament has clear bounds. 11

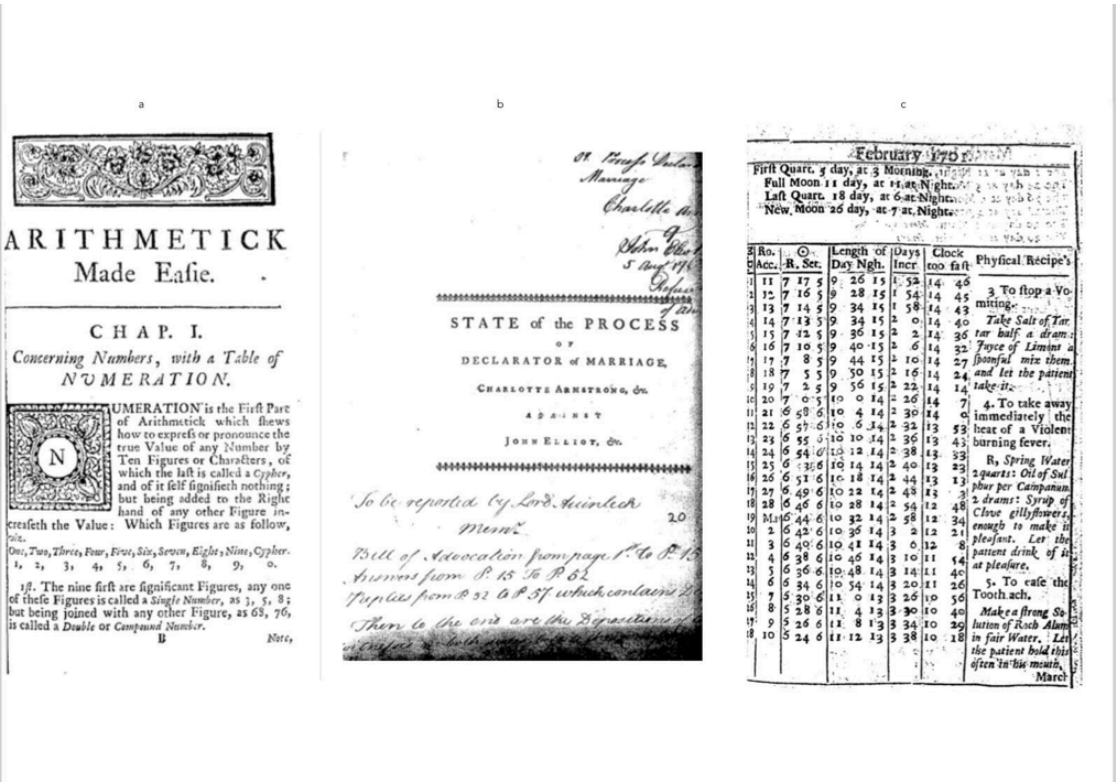

Figure 4. Problem images. 4a: A factotum close to the text and page boundary; 4b: Manuscript notes interfere with the image; 4c: A table, and show-through.

The next step was to run the program, which we named Fleuron, on all 36 million page images. This was undertaken in batches using the High Performance Computing cluster “Darwin” at Cambridge University Information Services. Each batch consisted of 50 books, and it required on average 6 hours for Fleuron to process a single batch on a single compute node of Darwin. It took five weeks to process all the batches using between 8 and 30 nodes in parallel at any given time. In total, 3,257,956 images were extracted from the pages. 12

Due to Fleuron’s permissive approach to image extraction, described above, there were a lot of type one errors in the extracted data set. Rather than try to improve Fleuron’s heuristics, a more reliable approach is to use the data produced by Fleuron to train a machine learning model that can review the extracted images and automatically detect which images are type one errors. A random set of 5,000 images was hand labelled as either “valid” or “invalid.” To represent the images, a bag-of-visual-words representation was used. RootSIFT vectors (see (Arandjelović and Zisserman (2012))) were extracted from all of the images in the labelled data set to train a visual word dictionary with 20,000 visual words. The RootSIFT vectors of each image were then matched to the nearest equivalent word in the dictionary to create a tf-idf vector that represents that image. With the images expressed as feature vectors a machine learning model can be employed. 13

The data set was split into 12000 images for training and 3000 images for testing, and linear SVM was trained on the data. It is more preferable in this case to classify an invalid image as valid than it is to classify a valid image as invalid. Therefore, the class weights were tuned in the SVM in order to trade off some recall, and gain precision. The resulting algorithm has 95.4% accuracy, 93.8% recall, and 99.5% precision when detecting invalid images. After applying the algorithm to the rest of the data set, 1,988,841 images were classified as invalid, leaving 1,269,115 images. 14

The resultant database was launched online in October 2016 at http://fleuron.lib.cam.ac.uk. It is hosted by Cambridge University Library. Although the full-page images collected in ECCO are kept behind a paywall, all of the content is in the public domain, so the images collected in Fleuron are freely disseminated for public use. The database still contains a number of type one errors that were not identified in the first round of removals. The images in Fleuron are currently searchable by keyword in two ways, by book or by ornament. The Book Search function allows searches by author, 15

publication place, publisher, and ESTC ID. The results can be sorted according to the following genres: history and geography; social sciences; general reference; law; fine arts; religion and philosophy; literature and language; medicine, science and technology. The Ornament Search function allows individual ornaments to be located by their size, and the results to be limited according to the criteria already outlined.

16 The ornament search can be improved considerably by using content-based image retrieval. This will involve using images themselves as search queries and retrieving images from the data set that look identical or very similar to the query image. Like the machine learning filtering employed to remove bad images from the data set, image searching can also be done using the bag-of-visual-words methodology. By representing each image as a tf-idf vector, similar images should have a smaller distance compared to dissimilar images. Performing image retrieval like this is effectively a nearest neighbour search. This application will allow researchers to quickly find identical ornaments from different publications in the database, which can provide valuable evidence of the authorship and publisher of documents where these are unknown.

17 Fleuron has significantly improved and sped up the process of finding and browsing printers' ornaments of the eighteenth century. Fleuron offers an opportunity for book historians, art historians, and historians of graphic design to examine a wealth of ornaments with ease in one place. These miniature works of art have much to offer, and the Fleuron blog[4] documents some of the directions that may be taken by future researchers into the interpretation of ornaments. For scholars working on the history of printing, and particularly on printer identification, Fleuron will facilitate new bibliographical research. It is currently possible to browse all ornaments with a known association (from the imprint) with a particular printer, publisher, place, author, or year. This will facilitate the compilation lists of ornaments known to belong to a particular printer, or known to have been in use in a particular city. These can in turn help to identify the printer etc. of books where the imprint is lacking such information. As for solving the problem of an unknown printer where we have no such leads, this process will be greatly facilitated by the image-match function described in the previous paragraph, which could potentially enable the identification of unknown printers on an unprecedented scale. We conclude, however, with some necessary cautions on the use of Fleuron for printer identification. The identification of unknown printers using ornaments should be undertaken while bearing two facts in mind: first, that printers occasionally lent one another their blocks, and second that printers often shared jobs (for example, one printer printed sheets A–C of a book, and another printed D–E, though actual arrangements could be more complicated than this). Both of these facts mean that the presence of a single ornament belonging to a known printer in an unsigned book does not offer concrete evidence that the unsigned item (or all of the unsigned item) was printed by the individual who owned the ornament. The ornament could have been lent to another printer, or the printer to whom it belonged could have printed only one sheet of the publication, with a different unknown printer (or printers) responsible for the rest. If only one ornament appears in a given item, we can certainly speculate as to the identity of its printer, but for a definitive identification multiple examples from throughout the book are preferable. Likewise, when assigning an ornament to a printer, it is desirable to find more than one example of the ornament appearing under that printer's name, again to eliminate the possibility of shared printing and borrowed ornaments. Finally, because of the low quality of the images in Fleuron, it is not always possible to see all of the delicate details of the ornaments. Consultation of the original book will occasionally reveal very minor differences between ornaments that do not show up in Fleuron, and indicate the presence of a copy or a cast. Likewise, it is not always possible to determine, from examinations of the grain and texture of the impression, whether a given ornament on Fleuron was made from a woodcut, a metal cut, or a cast. The original document should always be consulted where possible, before printer identifications are published. Fleuron, then, is designed as a finding aid: its aim is to make printers' ornaments easily searchable, locatable, and browsable, and to direct researchers back to the original documents for further interpretation. Those interpretations, when grounded in bibliographical study, have the potential to open up thousands of new avenues for research into the history of printing, authorship, reading, and design.

About Fleuron

18 Fleuron was launched in 2016; since then, a new collaboration has been established with the Oxford Visual Geometry Group (VGG). The VGG created Seebibyte, a suite of publicly available programmes for image analysis, including the

VGG Image Search Engine (VISE) (http://www.robots.ox.ac.uk/~vgg/software/vise/) and the VGG Image Classification (VIC) Engine (http://www.robots.ox.ac.uk/~vgg/software/vic/).[5] In 2018-19 the VGG indexed the Fleuron database, and have used VISE to create clusters of similar images from the database. Wilkinson has annotated the clusters to make them searchable, and to provide a dataset for future work on the database using VIC. The indexing of the database has also made it possible to remove the many type one errors that remain in the database (library stamps, diagrams, clippings of text, etc.) In 2020 the website http://fleuron.lib.cam.ac.uk was replaced with a new site at https://compositor.bham.ac.uk, and rebranded as “Compositor.” With this update, it is possible to perform image searches, supported by the VISE tool, and to search the database by keywords pertaining to the subjects depicted in the ornament. This advance will make the database easier to navigate, and will greatly improve the efficiency with which we can perform the research described at the beginning of this article.

Notes

[1] A Version of this paper was first given at the Göttingen Dialog in Digital Humanities, Georg-August-Universität Göttingen, 2016.

[2] A block with a central hole in which a type letter could be inserted.

[3] Darby's signed output is estimated from the *The English Short Title Catalogue* (*ESTC*), and [Foxon 1975]. When printers have been subjected to comprehensive study, their output has been found to be in the thousands rather than the hundreds. See [Maslen 2001].

[4] https://fleuronweb.wordpress.com

[5] Development and maintenance of VISE and VIC is supported by EPSRC programme grant Seebibyte: Visual Search for the Era of Big Data (EP/MO13774/1).

Works Cited

Arandejelovic 2012 Arandjelović, Relja and Andrew Zisserman. “Three things everyone should know to improve object retrieval.” *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, 2012): 2911–2918.

Baines Reed 1952 Baines Reed, Talbot. *A History of the Old English Letter Foundries* (1887), rev. by A. F. Johnson. London: Faber and Faber, 1952.

Barchas 2003 Barchas, Jeanine. *Graphic Design, Print Culture, and the Eighteenth-Century Novel*. Cambridge: Cambridge University Press, 2003.

Bergel 2013 Bergel, G., A. Franklin. M. Heaney, R. Arandjelović, A. Zisserman, and D. Funke. “Content-Based Image-Recognition on Printed Broadside Ballads: The Bodleian Libraries’ ImageMatch Tool.” *IFLA WLIC* (Singapore 2013). http://library.ifla.org/209/1/202-bergel-en.pdf

Bergel 2016 Bergel, Giles. “Printing Cliches” (2016). www.printing-machine.org/notes/2016/6/4/printing-cliches

Bigun 1996 Bigün, J., S. K. Bhattacharjee, and S. Michel. “Orientation Radiograms for Image Retrieval: an Alternative to Segmentation.” *ICPR ’96: Proceedings of the International Conference on Pattern Recognition* (Vienna, 1996): 346–50.

Corsini 2003 Corsini, Silvio. “Passe-Partout.” *Bulletin des bibliothèques de France* 5 (2001): 73–9.

Foxon 1975 Foxon, David. *English Verse 1701–50: A Catalogue of Separately Printed Poems*. Cambridge: Cambridge University Press, 1975.

Goulden 1988 Goulden, Richard. *The Ornament Stock of Henry Woodfall*. Oxford: The Bibliographical Society, 1988.

Maslen 1974 Maslen, Keith. *The Bowyer Ornament Stock*. Oxford: The Bibliographical Society, 1974.

Maslen 2001 Maslen, Keith. *Samuel Richardson of London, Printer: A Study of his Printing Based on Ornament Use and Business Accounts*. Dunedin NZ: University of Otago, 2001.

Meynell 1923 Meynell, Francis and Stanley Morison. “Printers' Flowers and Arabesques.” *Fleuron*, 1 (1923):1–45.

Michel 1996 Michel, S., B. Karoubi, J. Bigün, and S. Corsini. “Orientation radiograms for indexing and identification in image databases.” *European Conference on Signal Processing (Eupsico)*, (Trieste, 10-13 Sep. 1996), 1693–96.

Mosley 2015 Mosley, James. “Dabbing, abklatschen, clichage...'” *Journal of the Printing Historical Society* 23 (2015).

Ross 1990 Ross, John C. *Charles Ackers' Ornament Usage*. Oxford: The Bibliographical Society, 1990.

Ryder 1972 Ryder, John. *Flowers and Flourishes*. London: The Bodley Head, 1972.

Toner 2015 Toner, Anne. *Ellipsis in English Literature: Signs of Omission*. Cambridge: Cambridge University Press, 2015.

Wilkinson 2013 Wilkinson, Hazel. “Printers’ Flowers as Evidence in the Identification of Unknown Printers: Two Examples from 1715.” *The Library*, 7th ser., 14 (2013): 70–9.