IFLA WLIC 2013
SINGAPORE

Submitted on: June 12, 2013

Content-Based Image-Recognition on Printed Broadside Ballads: The Bodleian Libraries' ImageMatch Tool

Giles Bergel, Faculty of English Language and Literature, University of Oxford, Oxford, United Kingdom. giles.bergel@ell.ox.ac.uk

Alexandra Franklin, Bodleian Libraries, University of Oxford, Oxford, United Kingdom. alexandra.franklin@bodleian.ox.ac.uk

Michael Heaney, Bodleian Libraries, University of Oxford, Oxford, United Kingdom. michael.heaney@bodleian.ox.ac.uk

Relja Arandjelovic, Department of Engineering Science, University of Oxford, Oxford, United Kingdom.

Andrew Zisserman, Department of Engineering Science, University of Oxford, Oxford, United Kingdom.

Donata Funke, Bayerische Staatsbibliothek, Munich, Federal Republic of Germany

Abstract:

This paper introduces the Bodleian Ballads ImageMatch tool, developed by the Visual Geometry Group of the University of Oxford's Department of Software Engineering on behalf of the Bodleian Libraries. ImageMatch was designed to assist with the cataloguing and study of the pictorial content of early British printed broadside ballads, but has potential value for many other kinds of printed material. The paper outlines the nature of the materials to which ImageMatch has been applied; describes how the tool works and what it can do; and will offers some discussion on the benefits of ImageMatch's for image-cataloguing in Rare Books collections.

Keywords: Bodleian, ballads, broadsides, woodcuts, image-recognition

This paper introduces the Bodleian Ballads ImageMatch tool, developed by the Visual Geometry Group of the University of Oxford's Department of Software Engineering on behalf of the Bodleian Libraries. ImageMatch was designed to assist with the cataloguing and study of the pictorial content of early British printed broadside ballads, but has potential value for many other kinds of printed material. The paper outlines the nature of the materials to which ImageMatch has been applied; describes how the tool works and what it can do; and will offers some discussion on the benefits of ImageMatch's for image-cataloguing in rare books and special collections.[i]

Broadside ballads are cheap printed sheets carrying lyrics, illustrations and the names of popular tunes. They were sold, displayed, sung and read in the streets and alehouses of Britain from the 16th until the early 20th centuries.[ii]

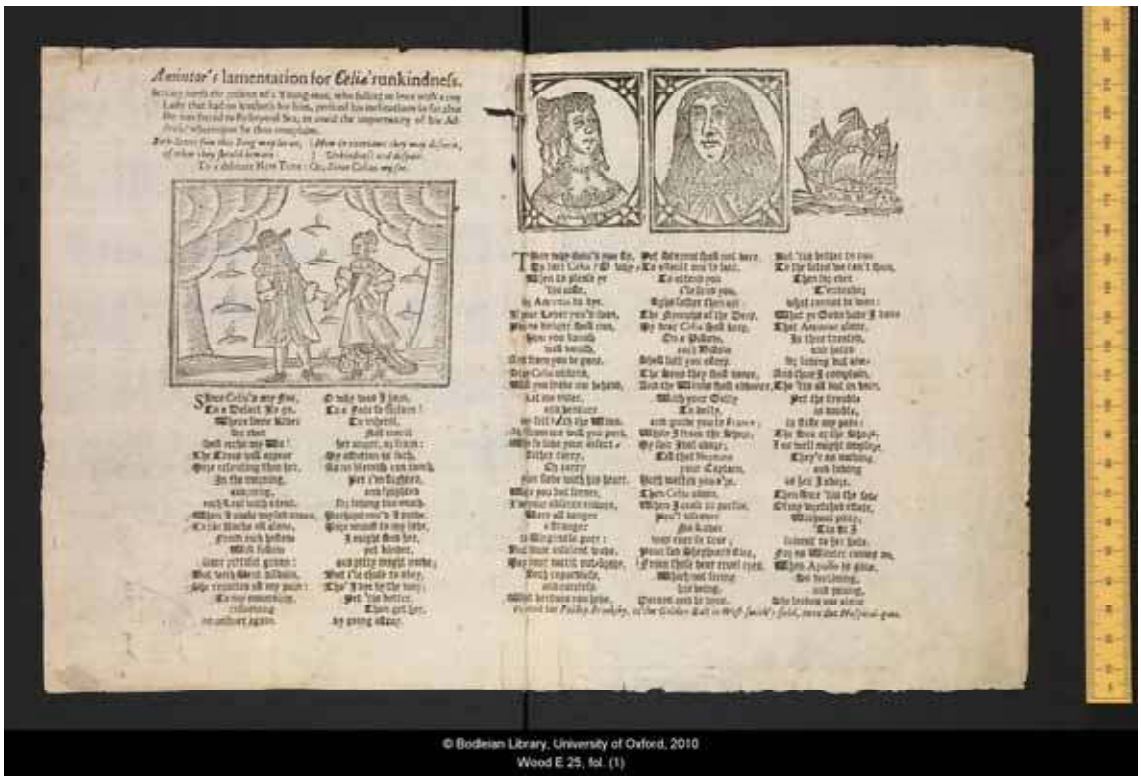


Fig.1: An example of broadside ballad from the 17th Century (Bodleian MS Wood E25 (1))

In addition to their literary and musical content, broadside ballads are specimens of visual culture. They were printed on one side of a sheet, for the purposes of public display, and are frequently illustrated, usually with woodcut prints or, in a later period, wood engravings. The illustrations may accompany a particular song-text, by illustrating the subject of a ballad-song, or an incident within the narrative. At other times, the illustrations relate to other illustrations on the same sheet or are independent works in their own right. Sometimes, there is no clear relationship between text and illustration, the illustration having been chosen to balance the mise-en-page of the sheet, to convey a generic quality or supply a mood.[iii]

Broadside ballads were printed in literally millions of copies, but very few of them survive. However,the Bodleian Library has over 34,000 specimens of broadside ballads, in collections that date from the 17th until the 20th centuries. Most of the ballads have been available online

for the last fifteen years, but over the last 18 months the Bodleian has been working to replace the database using a new relational cataloguing model that incorporates all the previous metadata, images and other content, alongside many enhancements, including ImageMatch.

Fig. 2: Bodleian Ballads Online

Because of their complex multi-media nature, broadside ballads are difficult to describe. Many typical features of ballads are challenges to traditional Anglo-American cataloguing practice: these include the importance of the first line of the ballad in establishing the identity of the text,; the absence of an overall title for a ballad sheet that may contain more than one song; the common lack of printing details (place and date of publication, the printer's name); the variable number of ballads on a single sheet; the central importance of their illustrations, and the re-use of components of one sheet on other sheets. The Bodleian has employed a series of cataloguing solutions to accommodate the complexity of the materials and the needs of researchers, who may be specialists in social history, the history of folksong, of popular culture, of illustration, music, printing or the British book trade. A new relational cataloguing model has been published as an RDF schema and is available for others to use and to adapt.[iv] The first online database of broadside ballads in Bodleian collections, created between 1995 and 1999, implemented a subject index of illustrations, based on the iconographic indexing system, ICONCLASS.[v]

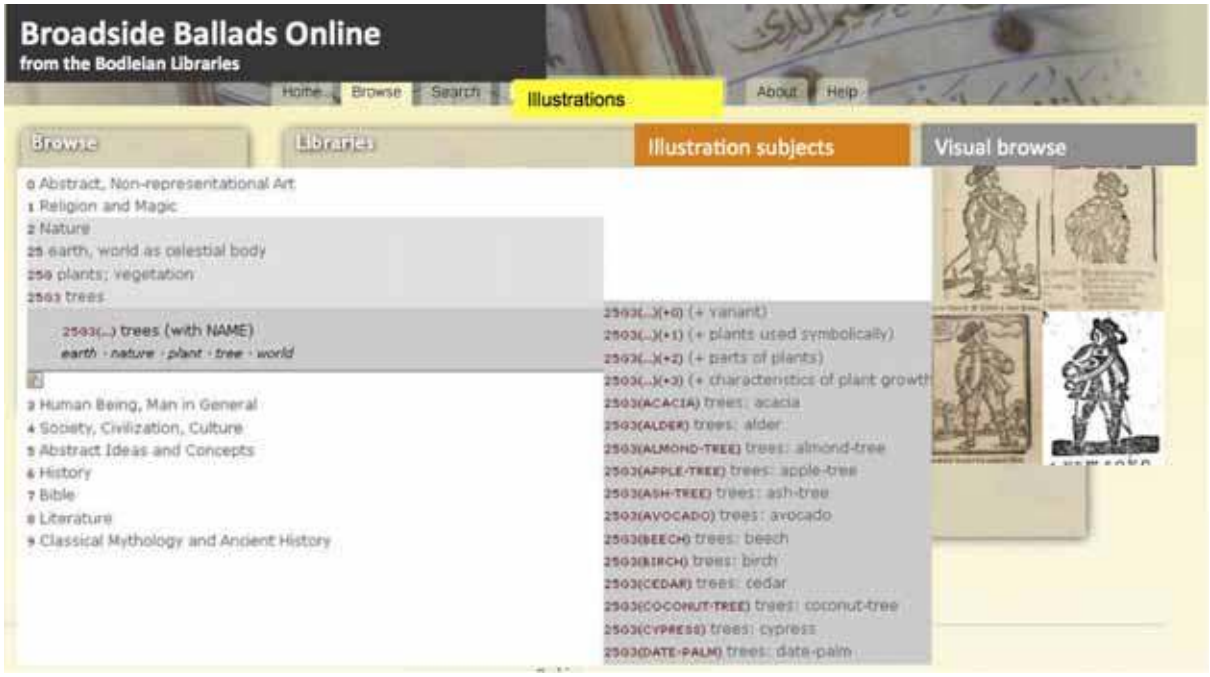


Fig. 3: Mockup of Bodleian Ballads ICONCLASS browser

ICONCLASS is more detailed than other tools such as the Library of Congress Thesaurus of Graphic Materials, but classification schemes require a key or code to understand and rely on the indexer assigning the same codes in the same situations. This can be very difficult when classifying a large collection.

Contrasting with ICONCLASS's iconographic analysis and description, ImageMatch does not offer a subject categorization of images, but is designed to assist the cataloguer and user in finding illustrations derived from the same woodblock or other printing-surface.

Fig. 5: A ballad woodblock (British Museum 2000,0723.9) and a corresponding impression

The ability to match woodblocks is interesting for cataloguers, bibliographers and art historians for a number of reasons. The use of an identifiable woodblock may enable us to assign a date to the printing of a ballad-sheet. The condition of a woodblock, as revealed by features within the printed impression such as wormholes or other damage, may allow the cataloguer to assign a date or date-range to the printing of a ballad-sheet within a sequence of other impressions.[vi] The use of an identifiable woodblock may reveal the printer of an anonymous ballad-sheet; its place of printing; the sale or loan of a woodblock; or a relationship between printers.

Identifying common woodblocks is a challenging task that is complicated by several factors. Variations in inking, damage to blocks and to sheets, the resolution, scaling and skewing of the surrogate digital image; and the presence of many copies of popular subjects, which in some cases are difficult to distinguish. Here we have an example of many copies of the same subject, printed from different blocks:

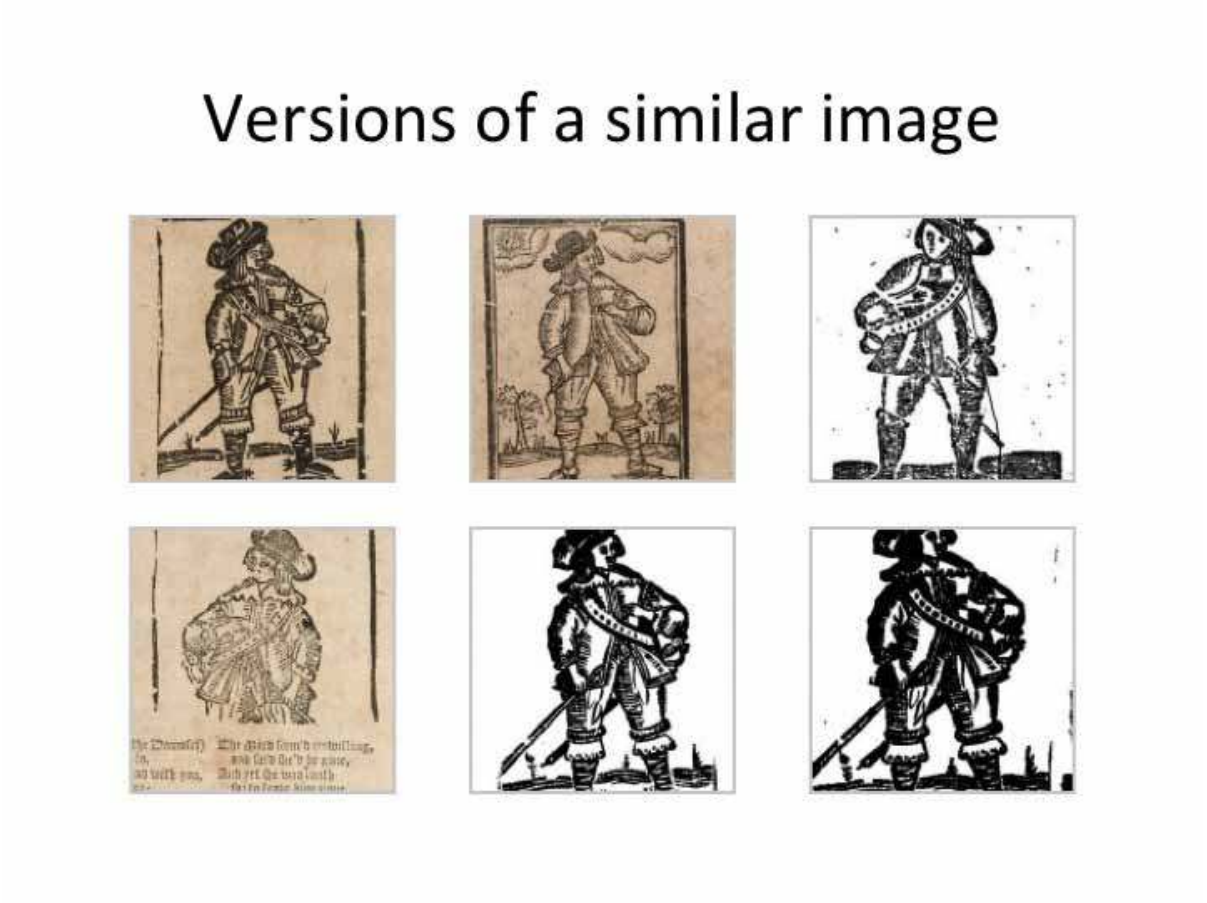


Fig. 6: Variant illustrations

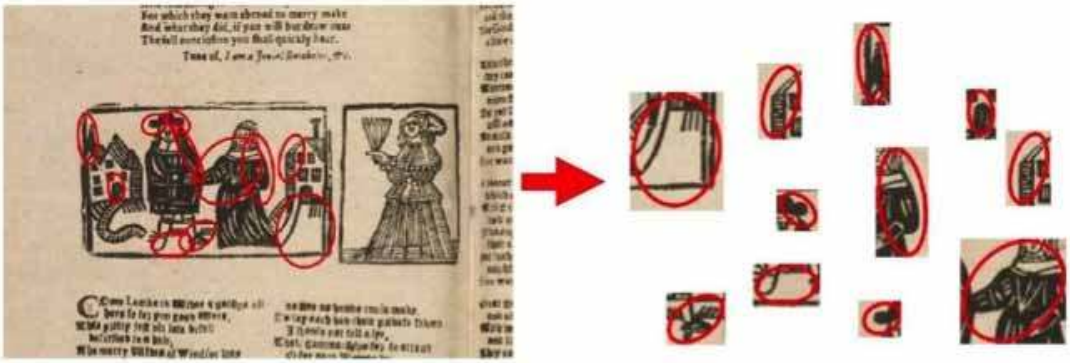
ImageMatch provides two ways of helping the cataloguer in matching woodblocks. Both of them are also useful in distinguishing copies of separate blocks. The two modes of operation are first, an image retrieval mode, and second, an image comparison mode.

ImageMatch is available for image retrieval within the Bodleian Ballads Online resource, integrated with metadata about the collections. It is also available as a standalone tool. It has been implemented on a subset of some 800 high-resolution images of ballads from the 17th Century, when broadsides were most frequently illustrated with a common set of woodcuts. It is invoked by selecting a region of the image of a ballad sheet within the collection, or by uploading an image file, or by providing the URL of an image stored elsewhere.

Matching and retrieval works by treating the selected image as a collection of 'visual words' - iconic patches, or fragments, of the image as a whole, which are bounded by ellipses of various sizes. The software queries the visual words contained within the selected image against a pre-built index of images within the database, considered as a collection of identically-constituted visual words, using an efficient search algorithm similar to keyword searching.[vii]

How it works

- Representation: bag of (visual) words
- Visual words are 'iconic' image patches or fragments
  - represent the frequency of word occurrence
  - but not their position

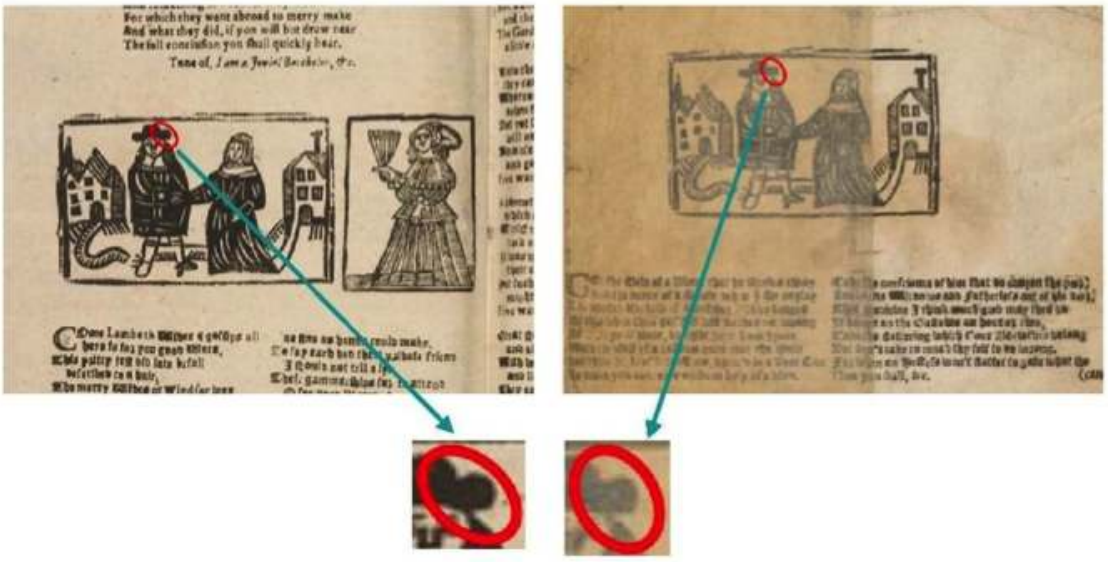


Image

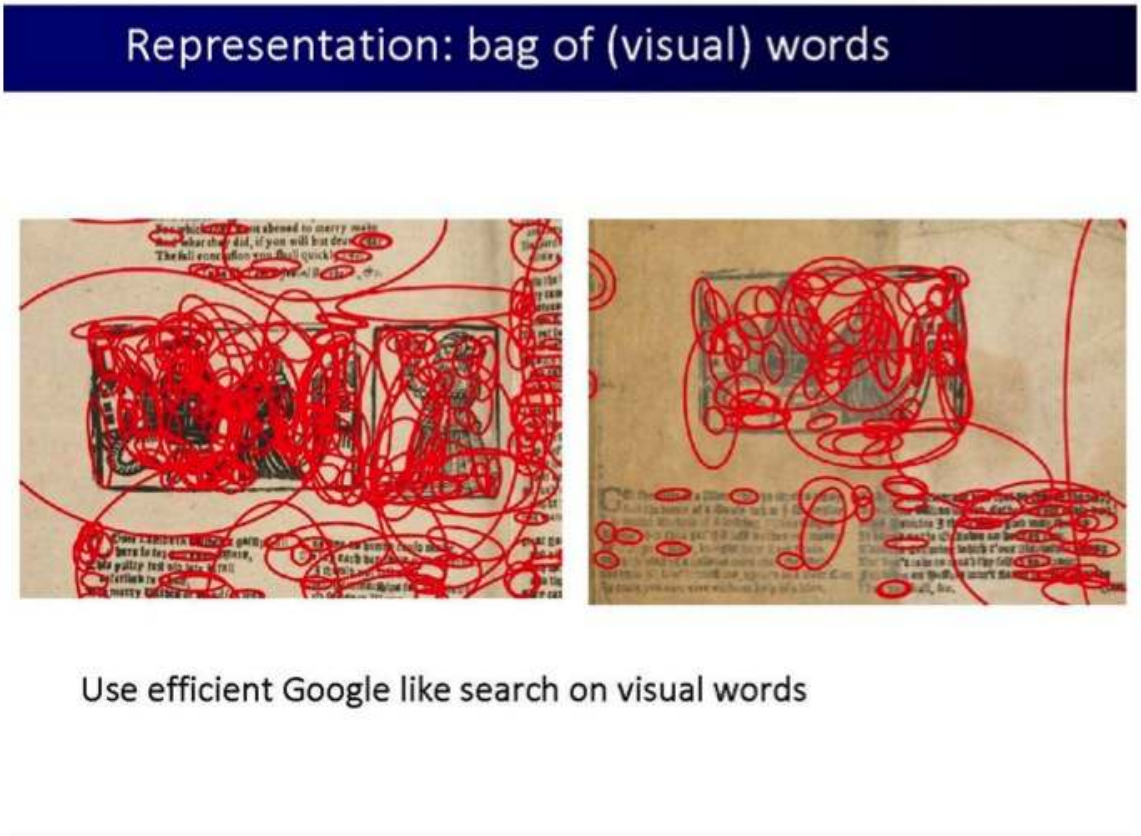
Collection of visual words

Visual vocabulary unaffected by printing artifacts

- Differences in printing strength, spatial deformations



The same visual word



Figs, 7, 8, 9: Searching on 'Visual Words'.

The system returns matching images ranked in order of similarity, expressed by a numerical score, the highest score being assigned to the starting image, that is, an identical match.

Fig. 10: Numerical scores indicate closeness of matching

We have tested the accuracy by comparing the results of a sample of images, in both bitonal and high-resolution formats and with both uploaded images and images held within the database as queries, against ground-truth established by manual identification of blocks and similar illustrations.

The following slide shows the software's performance in matching four blocks, two of which are close copies.

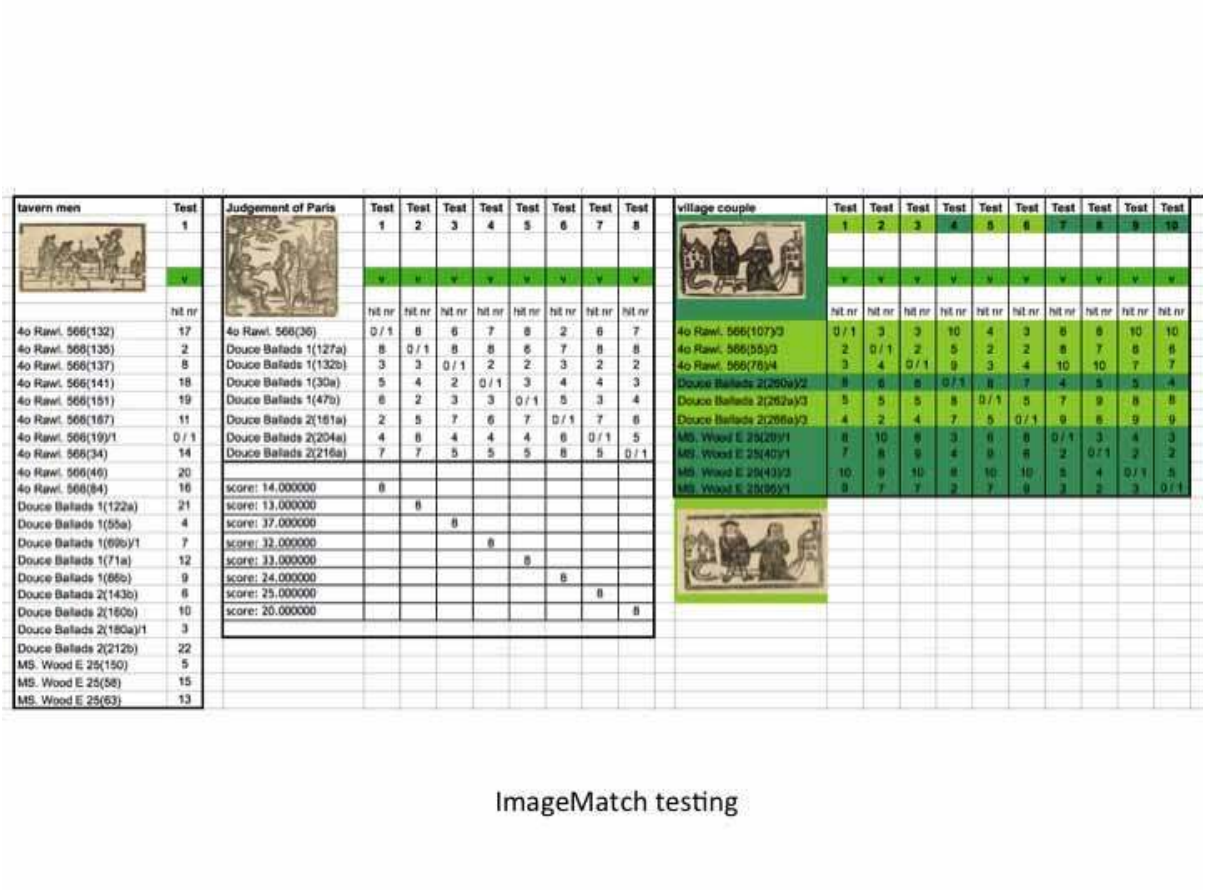


Fig. 11: Testing ImageMatch accuracy using ground-truth

Summarised, we can see that ImageMatch finds over 90% of related images in a test of seventeen blocks divided into four groups of varying degrees of similarity.

Test-group	Number of tests	Number of blocks	Avg. precision	Avg. recall
1	14	2	0.66	1
2	3	2	0.69	0.86
3	12	5	0.96	1
4	7	8	0.72	0.85
TOTAL	**36**	**17**	**0.76**	**0.93**

Fig. 12: Table of ImageMatch precision and recall scores

While ImageMatch is designed to match impressions taken from whole blocks in order of confidence, the ranking may also include close copies of blocks, even when only part of a block has been copied. Here, you can see two different blocks, but which include copies of the surrounding frame and some similarities between the subject in the middle.

Fig. 13 Copies of blocks with similar elements

It can also trace fragmented images, where a block or an impression has been damaged

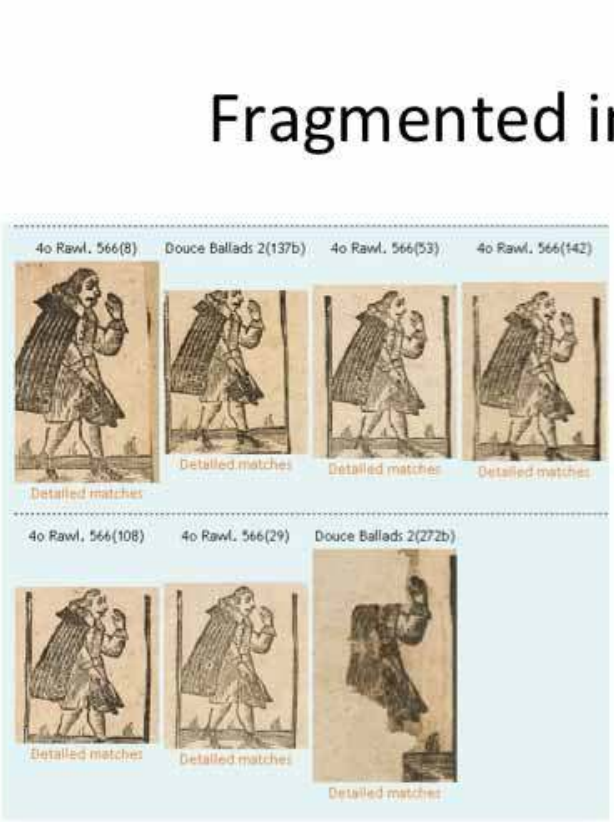


Fig. 14: Retrieval of details of incomplete (damaged) illustrations

And it can match composite images made up of fragments of blocks that appear in different combinations elsewhere.

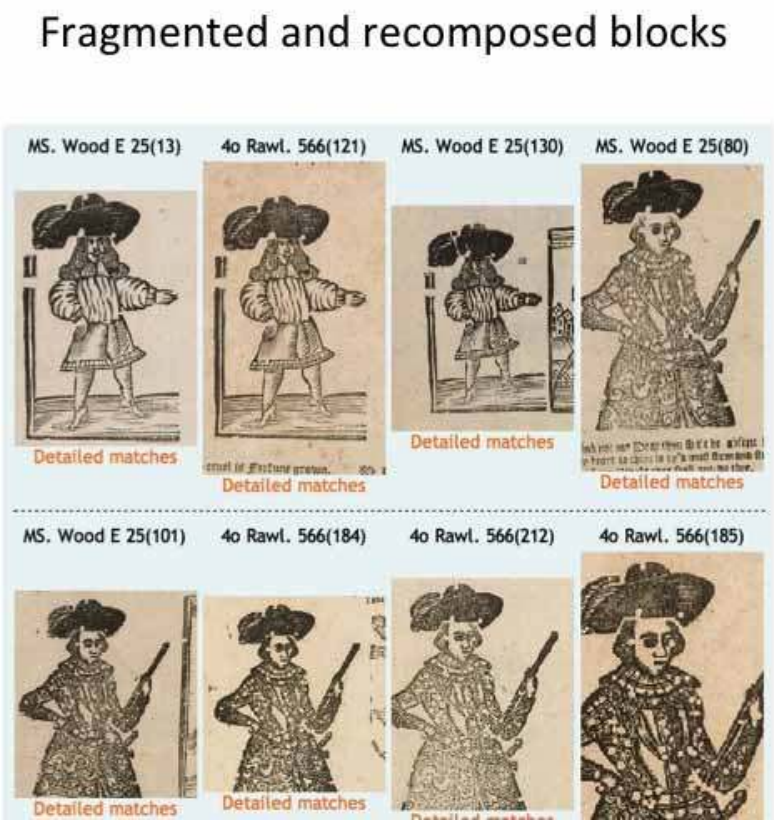


Fig. 15: Fragments of blocks

This robustness gives useful results for iconographic cataloguing, tracing visual traditions, while also showing the complexity of a print culture in which blocks or their elements were swapped, divided, reformed and copied.

Turning to the image comparison mode, this is used for closer examination of potential block-matches. Like optical bibliographical collators, which are typically based on overlaying or stereoscopy, the image comparison mode provides an overlay, or registration, between two images, with the effect of highlighting any differences.[viii]

Registration is provided by a simple two-dimensional transformation. This can adjust for most differences in scale introduced by photography. Identifying common blocks or plates is not a well-documented procedure. Most researchers rely on experience to adjudicate matches between impressions: a distinctive area of damage (such as a wormhole) is the most common source of positive evidence for a match, while a significant discrepancy between the dimensions of an impression, or between sets of points within impressions, is the most powerful source of negative evidence against a match. Strikingly, the human eye is often overwhelmed by a profusion of data-points, complicated by variations in inking, wear and tear, and the possibility of a traced impression or cast block.[ix] All these obstacles make it difficult to come to a conclusion within a short space of time. The image comparison mode can, first, reduce the population size of potential common blocks by ruling out obvious mismatches. It can then assist the cataloguer in judging whether more similar images derive from the same block

The original Bodleian ballads database included an iconographic index of the illustrations. The rather laborious workflow fifteen years ago involved examining each broadside at the edition level and describing the illustrations, doing the best that could be done to repeat the descriptions when illustrations were known to be similar. It took 2 years to create index entries for 22,800 illustrations. Image matching offers a speeding up of this process. If, as soon as we describe one image, we can immediately find and identify identical and similar images, then we can confidently assign the same codes to all of them. For example, the 20 occurrences of this scene of men talking around a table could have been coded simultaneously.

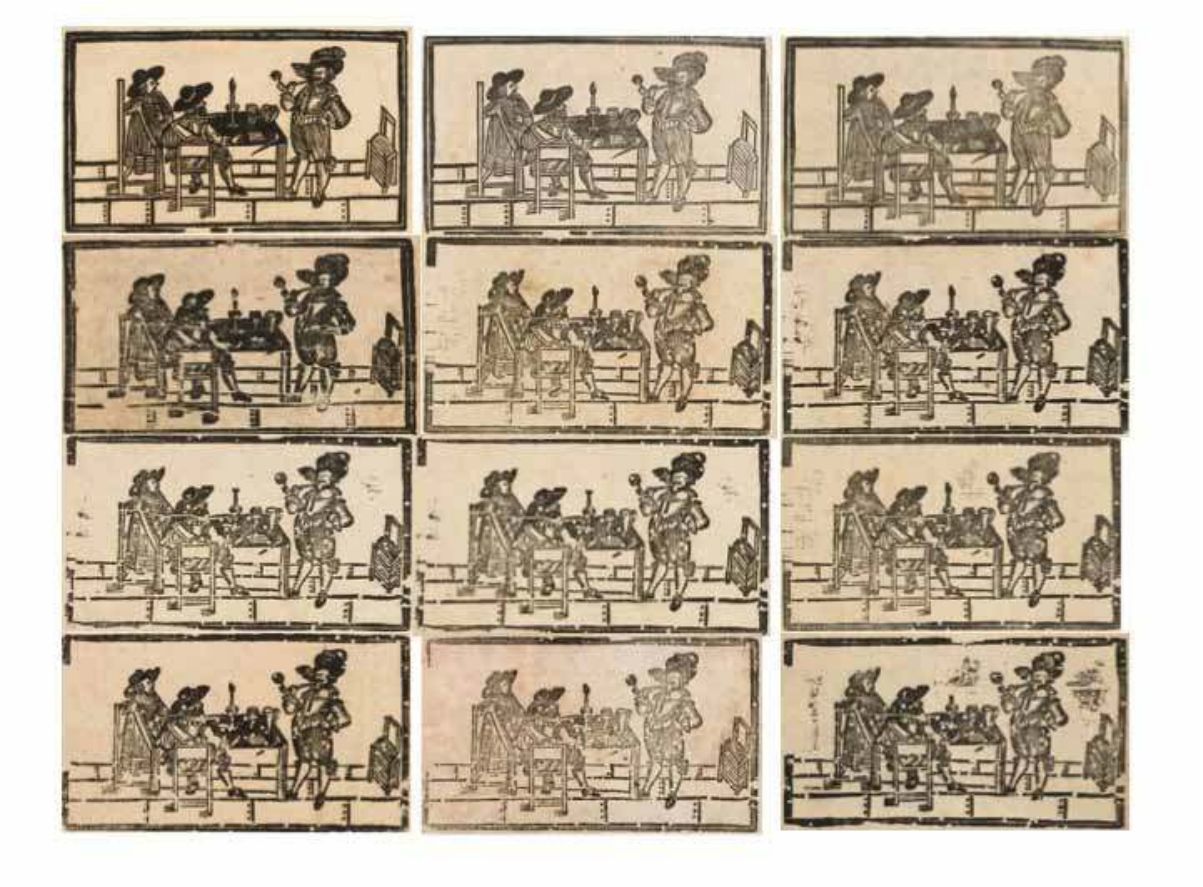
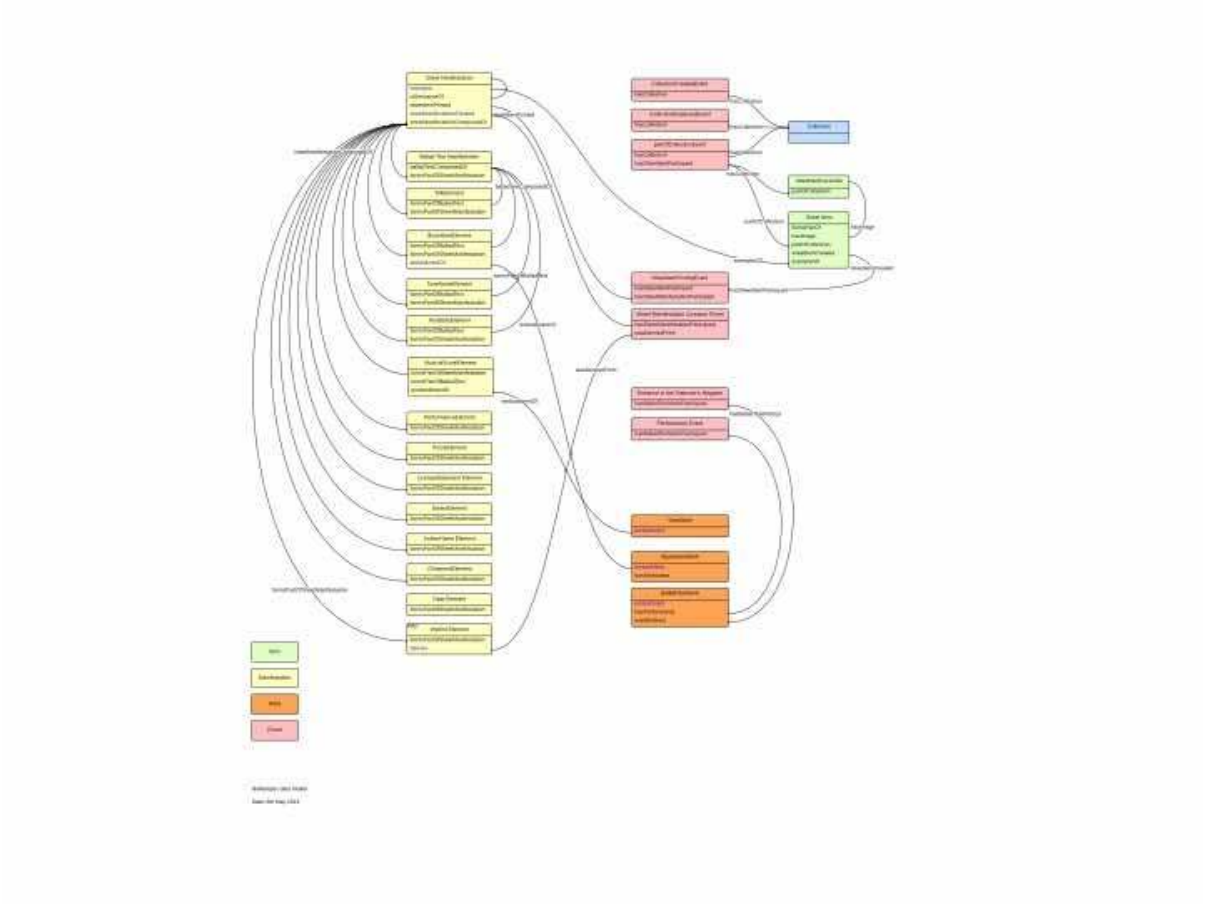


Fig. 16: Multiple impressions from the same block

What was not attempted in 1999 was a catalogue that gave an identity to each woodblock.[x] Now that it is possible to use image matching to group impressions together, it's possible to create an identity for the image matrix (in this case, the woodblock used). This will greatly simplify the workflow for making an index of illustrations in any collection of printed works. Matching images taken from the same woodblock can now be grouped together at the start, and assigned identities. Once an image has been identified, its coordinates on the sheet can also be added to the catalogue record, and associated with the correct ICONCLASS codes and other descriptors already present in the catalogue record.[xi] Then the first layer of description, that simply explains what this is a picture of, can be done for all the impressions of that woodcut just once. Image matching that finds similar, but not identical, images is also helpful and would save time for the cataloguer: descriptions can be replicated for several versions of an image, or modified when slightly different images are discovered. Finally, an explanation of what the illustrations might mean in connection with the text of a ballad is yet another layer of description, and once the images themselves are properly identified,

cataloguers could approach this step confident that the inherent description of the pictures remains linked to their identities. All of these aspects of illustration-cataloguing – at the level of the item; the manifestation, corresponding to the block, and at the level of the iconographic work, are comprehended in the new ballads cataloguing model.



Slide 17: Balladspec: an RDF schema for broadside ballads

ImageMatch shows strong potential for reuse on a broad range of other printed images. Certainly, it can readily be applied to other woodblock-printed images of a similar provenance: experimentally, images within the corpus of indexed ballads have been successfully matched against images taken from other digital collections, such as Early English Books Online (EEBO). Graphical elements of letterpress printing other than woodblocks, such as printers' ornaments, rules, initials and even small-scale features such as distinctive type sorts, may also be retrievable. It is likely that ImageMatch and similar image-recognition engines will rapidly become a common feature of printed image databases, whether in a federated model or in union indexes for specific classes of material.[xii] Users' expectations of a degree of accuracy similar to that which can be obtained through text search, however, may be challenging to meet, given the wide range of graphical elements present in the corpus of print, in varying states of preservation and image-quality.[xiii] Other printing processes, such as wood-engraving, intaglio or lithographic methods that produce a wider tonal range than high-contrast letterpress printing, pose different challenges, which image-retrieval systems may have to be modified to address.

ImageMatch is perhaps, at this time, of most value as a cataloguing tool for a closed corpus of materials. As more corpora become indexed, it will become invaluable, via the upload

function, for addressing materials that have not yet been described. For these materials, it can assist both in cataloguing their graphical content, but also in copy-cataloguing a complete item. Future developments will explore the limits of the current technology both through experiment with large-scale indexing of already-digitised materials and through more detailed studies of particular corpora. Library curators and cataloguers, technical developers and end-users will all have a role to play in developing its potential.

[i] ImageMatch is currently running on Bodleian servers as part of the Bodleian Ballads Online resource at http://ballads.bodleian.ox.ac.uk, developed as part of the Integrating Broadside Ballad Archives project, funded by the United Kingdom's Joint Information Systems Council (http://www.jisc.ac.uk/whatwedo/programmes/digitisation/content2011_2013/broadsideballads.aspx) It is also hosted by the Visual Geometry Group alongside technical specifications, tutorials and source code (http://www.robots.ox.ac.uk/~vgg), A history of the project, including video demonstrations, can be seen at http://balladsblog.bodleian.ox.ac.uk/blog/570

[ii] The bibliography on broadside ballads is extensive: for overviews of the form, see Leslie Shepard, *The Broadside Ballad*, (Herbert Jenkins, 1962) and Claude Simpson, *The British Broadside Ballad and Its Music* (Rutgers University Press, 1966). New Brunswick: Rutgers University Press, 1966). In addition to Bodleian Ballads Online (http://ballads.bodleian.ox.ac.uk) other digital collections include the English Broadside Ballad Archive (http://ebba.english.ucsb.edu) at the University of California at Santa Barbara and the National Library of Scotland's Word on the Street (http://digital.nls.uk/broadsides/index.html). The English Folk Dance And Song Society's Roud Indexes (http://library.efdss.org/cgi-bin/textpage.cgi?file=aboutRoud)link broadside ballads content with ballads in other formats and media.

[iii] See Alexandra Franklin, 'The Art of Illustration in Bodleian Broadside Ballads before 1820', *Bodleian Library Record* 17 (2002): 327-352

[iv] The RDF schema will be published on vocab.ox.ac.uk.

[v] The ICONCLASS browser [http://www.iconclass.nl/iconclass-2100-browser] arranges iconographic subjects into hierarchies and assigns code to these in a format that increases the length of the code in relation to the specificity of the subject. Therefore a short two-digit code indicates a broad field, such as "Nature" or "Society", while added letters and numbers specify, "palm tree" or "musical concert". Using the ICONCLASS browser, hierarchical browsing from general categories to specific examples can be cut across by the threads of metaphor and reference which join items in different categories. This enables cross-searching of keywords attached to the codes, so that an "arrow" may appear either as an attribute of the god Cupid, or as a weapon carried by a soldier, in different parts of the hierarchy.

[vi] For wormholes, see SB Hedges, 'Wormholes record species history in space and time', *Biological Letters*, 9 (2012)

[vii] See R. Arandjelovic and A. Zisserman, 'Three things everyone should know to improve object retrieval', *IEEE Conference on Computer Vision and Pattern Recognition*, 2012; J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman , 'Object Retrieval with Large Vocabularies and Fast Spatial Matching', *IEEE Conference on Computer Vision and Pattern Recognition*, 2007

[viii] Registration (overlaying) to assist with the detection of differences in printed impressions can be accomplished with standard image-processing software; the use of photographic transparencies; or with specialised optical collation devices (for the latter, see Steven Escar Smith, 'The Eternal Verities Verified: Charlton Hinman and the Roots of Mechanical Collation', *Studies in Bibliography*, 53, (2000). 129-161 and the documentation for the McLeod collator referenced in http://blogs.bodleian.ox.ac.uk/theconveyor/2010/09/03/library-machines-the-mcleod-collator)

[ix] Blocks may be copied, either directly of from their impressions, with a high degree of accuracy. For copying of impressions, see W. M. Ivins, *How Prints Look: Photographs with a Commentary*, (New York Metropolitan Museum of Art, 1943), 117-41. For copying and casting of blocks in metal, see James Mosley, 'Dabbing, abklatschen, clichage...' at http://typefoundry.blogspot.co.uk/2006_01_01_archive.html

[x] Cataloguers may however have noted distinct blocks, in particular for the purpose of identifying variant states, issues or editions to which the item as a whole belongs. One resource which does aim to distinguish woodblocks within the record of printed impressions is Ruth Samson Luborsky and Elizabeth Morley Ingram, *A Guide to English Illustrated Books 1536-1603,* (Medieval and Renaissance Texts and Studies, 1998). For the copying, sale and reuse of woodblocks, see Barry McKay 'Cumbrian Chapbook Cuts: Some Sources and Other

Versions', in Peter Isaac and Barry Mckay (eds.), *The Reach of Print: Making, Selling and Using Books*, (St. Paul's Bibliographies, 1998) 65-83

[xi] Out of the total of 24,405 editions in the Bodleian collections, 15,919, or 65%, have illustrations. The total number of illustrations, remembering that some editions have more than one picture, is 22,800.

[xii] Content-based image-recognition for printed images has also been implemented on the Passe-partout International Bank of Printers' Ornaments (http://www3.unil.ch/BCUTodai/); Japanese Woodblock Print Search http://ukiyo-e.org; and the Bavarian State Library's similarity-based image search http://bildsuche.digitale-sammlungen.de

[xiii] The majority of Bodleian Ballads images are bitonal scans from microfilm: this set of images has proven to be significantly less amenable to ImageMatch retrieval, in particular when a high-quality image is selected to query against a lower-quality index. Further research and development in domain adaptation is needed to improve matching on heterogenous image collections and in defining optimal imaging standards for the future.