Rediscovering the lost photographs of the Exhibitions of the Royal Photographic Society

À la redécouverte des photographies perdues des expositions de la Royal Photographic Society

Stephen Brown
Translator: Sandrine Chene
Rediscovering the lost photographs of the Exhibitions of the Royal Photographic Society

À la redécouverte des photographies perdues des expositions de la Royal Photographic Society

Stephen Brown
Translation : Sandrine Chene

This research is funded by the UK Arts and Humanities Research Council grant number AH/J004367/1. Thanks are due to Birmingham City Library, the British Library, Musée du Louvre, the Metropolitan Museum, Musée d’Orsay, the National Archives, the National Media Museum, National Museums Scotland, St Andrews University, the V&A and Professor Roger Taylor for their generous support.

1 The early history of photography coincides with significant scientific, industrial, artistic, social, political and economic changes in leading countries such as France, Germany and the United Kingdom. It can inform understanding of contemporary issues such as the spread of scientific ideas, the relationship between science and art, the interplay between new technologies, popular culture and commerce, and the creation of personal and national identities. However, the Study of photographic history as cultural history is relatively new and texts in this field only form the beginnings of a scholarly corpus. Access to primary resources is essential therefore for continued progress but these resources are often ephemeral, fragile, widely dispersed, poorly catalogued and difficult to access.

2 De Montfort University (DMU) has created a corpus of digital research resources on 19th century and early 20th century photography, working from primary materials such as exhibition catalogues, diaries and letters (http://kmd.dmu.ac.uk/kmd_photohistory_page/). This resource includes two databases of photographic exhibition catalogues that together comprise the most comprehensive record of British photographic exhibitions at this time: Photographs Exhibited in Britain 1839-1865 (PEIB) (
http://peib.dmu.ac.uk) and Exhibitions of the Royal Photographic Society 1870-1915 (ERPS) (http://erps.dmu.ac.uk). They contain information about around 65,000 exhibits from the earliest days of photography, including the premier photographic society of the day, the Photographic Society of London, which later became the Royal Photographic Society. These exhibition records are a highly significant research resource, offering a unique insight into the evolution of aesthetic trends and photographic technologies, the response of a burgeoning group of photographic manufacturers to business opportunities and the activities and fortunes of individuals concerned with the technical, artistic and commercial development of photography. The Royal Photographic Society’s exhibitions in particular attracted a wide constituency of photographers, from Britain, Europe and America and many individuals launched their photographic career through them. While these resources have received critical acclaim from the international photohistory research community, they are singularly lacking in one important respect. They contain very few images of the exhibits referred to. PEIB is entirely devoid of pictures, only 3% of the ERPS records are illustrated and many of these illustrations are merely artists’ impressions of the originals because, at the time of the exhibitions, mechanical reproduction of photographic images was technically difficult and expensive and in any case unnecessary since the photographs themselves were displayed in the exhibitions the catalogues referred to (Brown, 2008). Despite their richness and scope, the value of the records is constrained therefore by the absence of images to illustrate them.

Fig. 1. Example of an illustration of a photograph from the RPS exhibition catalogue of 1901.

Fortunately, there are increasing opportunities to match these exhibition catalogue entries with surviving historical photographs as image collections are increasingly being made available online by museums, libraries and archives. In the USA an analysis of 18,142 museums and libraries by the Institute of Museums and Library Services (IMLS, 2006) revealed that the majority of museums and larger public libraries make digital
records available to the public via the Internet and the same was true for a smaller proportion of public and smaller academic libraries. Individual institutional postings can run to hundreds of thousands of objects. As of January 2013, the Museum of Modern Art had posted more than 540,000 images, the British Museum offers images of more than 730,000 works and the Victoria and Albert museum (V&A) website offers access to more than 300,000 images of works in its collection (Kelly, 2013). Collectively these resources amount to a vast quantity of data. The Europeana aggregator website offers access to 21.3 million objects from 33 countries (Europeana, 2012).

However, with a total of circa 65,000 exhibition records to match, manual searching of museum, library and archive records is impractical because of the time and costs involved. Instead, the FuzzyPhoto project set out to develop and test a computationally based finding aid that can identify likely matches between the exhibition catalogues and online photographic collections from partner institutions and present the results to researchers as a set of suggestions for similar items to follow-up.

Heritage institutional records tend to be created using natural language because they are intended to be read by human beings rather than computers. Although some text mining tools can analyse natural language, they have mostly been developed to analyse large volumes of text such as found in books, reports, journal and newspaper articles (Tognini-Bonelli, 2001; McEnery and Hardie, 2012). Museum, library and archive records present a different challenge because while they may amount to thousands of words, each individual record typically contains relatively few words, often just a person name, an object title, a date and a short description. For this reason traditional corpus analysis and text mining tools were not appropriate for processing the FuzzyPhoto data. Instead, the plan was to ingest partner records into a common data warehouse where the records could be mined for similarities. Identified matches would be exported as links to a second separate database from which they could be returned to the partner’ Web pages. Figure 2 outlines the overall process.

Fig. 2. Overview of the FuzzyPhoto data work flow.
The intention was to display the results in partner Web sites rather than create yet another Web site that researchers have to remember. To this end it was decided to create a “widget” to be embedded in a partner’s Web page so that whenever a visitor lands on an object page within a partner’s Web site, links to potential matches are displayed in context. Figure 3a shows the widget embedded in a page from the Web site of the Musée d’Orsay and 3b shows the same widget expanded to reveal matching items ranked in order of similarity. Notice how visitors can choose between similarity by person name, by title or by overall combined measure. FuzzyPhoto seeks for matches using four key fields: person name, title, date and photographic process. The combined metric takes into account all four individual metrics.

Fig. 3a. A prototype widget embedded in a Musée d’Orsay Web page
Data acquisition

Six institutional collections were initially selected for comparison with PEIB and ERPS: Birmingham City Library, the British Library, the Metropolitan Museum of Art, the Musée d’Orsay, the National Media Museum, the Victoria and Albert Museum and St. Andrews University library. Subsequently, data from other collections were added (Brooklyn museum, Culture Grid, the National Archives, and National Museums Scotland). The initial stage of importing partner data into a common data warehouse was complicated by the variety, and in some cases the quality, of the data received. While no two data sets were alike, overall there were significant issues:

- Different data structures – because institutions use different records management systems that cluster and describe the data differently.
- Confused structure – some contributor records combined together a variety of different data types, for example date, person name, title, description etc. in a single field or cell. While such information is easy enough for a person to read, computers need it separated out into separate, labeled, categories.
- Inconsistent metadata – different labels were used for essentially the same descriptor by different institutions, eg. “creator”, “artist/maker”, “photographer”, “exhibitor”, “auteurs”.
- Incomplete records – empty fields.
- Inconsistent data entry – eg. name order such as Henry Tomas Malby; H.T. Malby; Henry; T. Malby; Malby, Henry; Thomas, and so on.
- Junk data – eg. mixed data fields such as person name and date of birth/death or date and place of birth in the same field, or date information expressed in text form: “circa pre Great War”. Again, while such information is intelligible to human beings, computers find it challenging when numerical and text data are combined or different types of text such as person and place names are bracketed.
The messiness of the data meant that it had first to be cleaned before being mapped to a common metadata schema. The ICOM Lightweight Information Describing Objects schema (LIDO) was developed specifically for exposing, connecting and aggregating information about museum objects on the Web and was therefore ideally suited to the task of standardizing the metadata provided by each of the contributors. Data cleaning and standardization took around 70 person days of effort.

**Data mining**

The endpoint of the data ingestion stage was a MySQL database of records in LIDO format. Although the records contained a great variety of different types of information the project focused in just four fields common to all the sets:

**Date**

The key factor in relation to date is the amount of time between the dates described in the fields. The greater the amount of time between them, the less similar they are. Some of the dates given describe a span of time rather than a specific year (i.e. "1890s", "the 19th century"). Greater differences between time spans indicate less similarity between fields. For example "19th century" and "1900" are less similar than "1900s" and "1900". Extraction of the date information from the various formats used was achieved with a combination of the Python dateutil library, regexes and a rule based system.

**Person name**

Name comparison is a well-established problem in many application areas and a large number of comparison algorithms exist which can handle typographical errors, alternative spellings etc. We used established edit distance techniques that measure similarity in terms of the number of changes (edits) that are required in order to convert one string into another (Winkler, 1990) combined with a heuristic best fit approach to match up individual name elements across fields.

**Process**

Process is a more complicated issue because many of the historical photographs in GLAMs collections entailed several separate processes in order to produce the final image that would today be called a photograph. We matched the stated process with a list of preset keywords describing various known processes using established techniques (Winkler 1990) to deal with typographical errors and minor spelling variations. The various photographic processes were organised into a dendrogram in which processes sharing specific traits appear on the same branch. Once the field had been matched to a specific locus it was compared to other fields using a graph transversal algorithm to find the shortest path between the processes. The shorter the distance between the approaches, the more similar they are considered to be.
The title similarity metric was particularly challenging because titles do not follow normal grammatical rules for sentences and are generally brief. (The average title length across all the exhibition records is just 8.1 words, of which only 5.4 are useful.) This ruled out standard corpus analysis tools. Short text semantic similarity tools such as Latent Semantic Analysis (LSA) and Sentence Similarity (STASIS), were developed to handle short texts but cannot easily process large numbers of records (O’Shea et al, 2008). As we had to deal with 1.4 million records we developed a simplified semantic similarity measure “Lightweight Semantic Similarity” (LSS) that can handle large numbers of very short records without significantly reducing the accuracy of results (Croft et al, 2013). LSS is based on standard statistical cosine similarity metrics (Manning and Schütze, 2003) but additionally takes into account the semantic similarity between words, using WordNet, in combination with Inverse Document Frequency (IDF) values (Papineni, 2001).

The final step was to combine the individual field metrics into an overall record similarity metric. A series of rules describe how to combine and weight the individual metrics, based on the programmed knowledge of domain experts. Rule based systems function poorly when faced with imprecise information (eg. “circa 1890”). The solution to this difficulty was to use a fuzzy rule-based approach that allows an object to simultaneously belong to more than one category to various degrees. The full set of fuzzy rules used draws on the individual field similarity metrics to ascertain if the match between any pair of fields is good and then combines that comparison with other field comparisons as follows:

If bad_title AND bad_person THEN terrible_match
If bad_title OR bad_person THEN bad_match
If good_title OR good_person OR (good_process AND good_date) THEN good_match
If good_title AND good_person THEN excellent_match

In order to reduce the time taken by this final stage of the comparison process, a novel defuzzification algorithm was developed (Coupland et al, 2014).

Good matches have been identified for over half (57%) of which around half are excellent matches. Table 1 shows an example of an excellent match between an image from the Library of Congress that FuzzyPhoto matched to a record in the 1893 RPS exhibition catalogue. Since the RPS catalogue does not contain an image of the photograph, we cannot be absolutely certain that these records co-reference the same photograph, nevertheless the high degree of correspondence between them suggests that they are very probably the same and therefore figure 4 shows us for the first time in 121 years what this exhibit looks like.
In addition FuzzyPhoto has identified matches that are clearly not the same and yet are strikingly similar. Figure 5 shows one such pair. Notice how none of the lines of description match exactly and yet daffodils and narcissi are the same genus, Henry
Thomas Malby and H.T. Malby are undoubtedly the same person despite differences in the way the names are presented, and 1895 is very close to 1896.

Fig. 5a. Fair Daffodils, Henry Thomas Malby, Bromide (Print), ERPS 1895, exhibit no. 410. Source: Photograms of the Year.

Fig. 5b. Pheasant-eye Narcissus, H.T. Malby, Platinum (Print), ERPS 1896, exhibit no. 142. Source: Photograms of the Year.
Thus far we have tested a sample of the outcomes on a panel of subject experts, to see how their speed and accuracy of co-reference identification compares with FuzzyPhoto. The results indicate that FuzzyPhoto is at least as accurate as experts and considerably faster. However these trials also revealed that “similarity” is a rather nebulous concept. Some researchers interpreted the notion of something similar to a given seed record as meaning something rather similar to the same photographer. Others understood it to mean something with a similar title or subject matter, or process. The interface shown in figure 5 allows users to choose between similarity by person name, title or by overall similarity based on the combined metric. However, better ways need to be found to allow users to nuance the results of the matching algorithms to more accurately reflect their expectations and needs.

Conclusions

In the last decade there has been tremendous effort on the part of heritage institutions to digitize their records and make them available to users online. The increased availability of online collections and searchable metadata creates enormous potential for using information about objects from one data set to enrich records in others, thus enhancing their value and revealing relationships not previously apparent. There is also growing recognition among cultural heritage institutions that there are significant benefits to be gained from sharing and connecting data and that users are increasingly likely to expect to be able to navigate seamlessly across separate collections. However, drawing together information from diverse collections is time-consuming and increasingly complex as the volume of online information increases and comparisons between collections are difficult because of incompatibilities between different institutional record management system, idiosyncrasies in the ways in which they are implemented and the sheer messiness of legacy data in terms of gaps, duplications and typographical errors.

“Without tools and methods for filtering, categorizing, clustering, weighting, and disambiguating our expanding datasets, museums will face increasing difficulty in managing and organizing their online resources, while their visitors will struggle to locate and make use of them” (Klavans et al., 2011).

The FuzzyPhoto project has developed an approach that has successfully identified matches for over half the records in the catalogues of the Royal Photographic Society annual exhibitions, allowing images of the exhibits to be seen for the first time in around 120 years and allowing a sense of the overall feel of the exhibitions to be partially recreated. In addition FuzzyPhoto has identified matches between partner records, even where these do not match exhibits in the RPS exhibitions. So, for example, photographs in the Musée d’Orsay have been matched with similar items in the British Library, and items from the National Media Museum collection have been linked to photographs in the Library of Congress. This shows that FuzzyPhoto is a powerful tool not only for rediscovering the lost images from the RPS exhibitions but more generally for identifying potential connections between important photographic collections in major international institutions, enriching our understanding through cross-referral. Further expansion of the FuzzyPhoto data warehouse has the potential to create a rich web of interconnections between different collections world-wide that will be invaluable for photographic historians, curators, researchers, teachers, students and dealers.
However, notwithstanding these encouraging results, the FuzzyPhoto approach is not without limitations. Scalability is an issue, firstly because of the messiness of the data. It took 70 person days to get the 1.4 million records in the FuzzyPhoto data warehouse cleaned up and mapped to a common metadata schema. While this equates to less than 0.1 second per record, nevertheless it remains a potential barrier to further expansion of FuzzyPhoto. Further work is being undertaken to create a customisable batch loader that will automate the data ingestion process but this will be able to process only rigorously and consistently structured records. A second scalability issue is processing capacity. At present 8 gigabytes of memory and around one month are required to process all 1.4 million records. These figures will increase as more records are added. Fortunately record matching is performed off line and the results are cached ready to be retrieved virtually instantly by end users. It does mean however that continuous updating of the records is impractical. Instead a regime of 6 monthly updates is planned.

Finally, although developed to address the specific problem of finding pictures to match the exhibition records in the ERPS and PEIB databases, the algorithmic approach described here has generated an unexpected outcome in the form of the “similar but different” matches identified, such as those shown in figure 4. These are quite different kinds of results compared with normal search engines, which strive to deliver items that match the search query as closely as possible, and they suggest a further use for FuzzyPhoto as a free-standing search engine.

The majority of content searches begin with a general-purpose search engine (Brown and Greengrass, 2010; Housewright et al., 2013). In the hands of skilled users, powerful search engines such as Google and Wolfram Alpha are highly effective tools for discovering resources using keyword searches. However, keyword searches do not make it easy to browse for interesting ideas and relevant information when one is not sure exactly what one is looking for, even though it is often easy enough to recognise the potential of such information when one sees it. The ability of FuzzyPhoto algorithms to identify similar but different items has the potential therefore to create a new kind of search engine that supports browsing and accidental discovery. Although the Web is already well-suited to accidental discovery of information not sought for (Campos and de Figueiredo, 2002), computational support for browsing is much less well-developed than for keyword searching. Yet arguably browsing behaviour is just as important for developing research ideas, particularly in the early stages of research before specific questions have been formulated (Budd, 1989). A “similarity engine” that can be used to conduct targeted searches but which extends the imagination of the researcher by returning results that are different from the search query, yet fundamentally related, would be a valuable tool for expanding the scope of their research beyond the boundaries of what they already know.
BIBLIOGRAPHY


NOTES
2. Based on a sample of 35 records.
3. Reproduced with permission.

ABSTRACTS

This paper describes a successful attempt to link together hitherto disconnected historical photographic records owned by different institutions using just textual record information, in order to identify images that have been missing from important photographic exhibition catalogues for over 100 years. The context for this research is the UK Arts and Humanities Research Council funded FuzzyPhoto project that has developed a finding aid for researchers interested in identifying pictures in museum online collections even when the records are incomplete, imprecise and incorrect. It describes the challenges of working with historical museum records and the approach developed to deal with them. This new approach is based on semantic matching techniques and fuzzy logic based algorithms that reflect the kinds of choices made by human decision makers when carrying out such tasks. Results are presented and the significance of this research is discussed for cultural heritage institutions and more widely.

Cet article présente une tentative réussie de rassembler des archives photographiques jusque-là éparpillées, conservées dans différentes institutions, par l’utilisation seule d’informations textuelles de ces archives, afin d’identifier des œuvres qui ont fait défaut à d’importants catalogues d’exposition pendant plus de cent ans. Conduite dans le cadre du projet FuzzyPhoto financé par le Arts and Humanities Research Council (AHRC) au Royaume-Uni, cette étude a
permis de développer un outil de recherche pour les chercheurs intéressés par l'identification des œuvres de collections de musées en ligne en dépit des lacunes, imprécisions et erreurs que ces documents peuvent comporter. Ce texte décrit les défis rencontrés dans le travail avec des archives de musées, ainsi que la méthode développée pour les traiter. Cette nouvelle approche repose sur des techniques d'appariement sémantique et sur des algorithmes basés sur une logique floue qui illustrent les types de choix que tout humain preneur de décision opère en poursuivant ce genre de tâches. Les résultats sont présentés et l'enjeu de cette recherche est débattu pour servir les institutions de préservation du patrimoine culturel et plus largement encore.

AUTHORS

STEPHEN BROWN

Stephen Brown is Professor of Learning Technologies and Director of Knowledge Media Design at De Montfort University and a Visiting Fellow at the Centre for Distance Education, International Programmes, University of London. Prior to this he was Head of the School of Media and Communication and Director of the International Institute for Electronic Library Research at De Montfort University; His career includes course design, research and tutoring for the Open University; He has managed 24 research projects. He has produced over 150 publications and he has led numerous workshops on technology enhanced learning at many different universities. He is a member of the editorial boards for Research in Learning Technology and Interactive Environments journals and has extensive experience reviewing for a range of other journals and academic conferences. Since 2005 he has been a registered European Commission expert in the fields of Technology Enhanced Learning, Digital Libraries, and Cultural Heritage and he was a member of the Arts and Humanities Research Council Peer Review College for ten years.