***Textes et contextes***
ISSN : 1961-991X
: Université Bourgogne Europe

**19-2 | 2024**
**Iconomorphoses : appropriation, éthique et partage - Représentations du monde hispanique actuel dans les séries télévisées**

# From the page to the web: a survey of digital illustration archives

*De la page au web : panorama des archives d'illustrations numériques*

Article publié le 15 décembre 2024.

**Ali Hatapçı**

http://preo.ube.fr/textesetcontextes/index.php?id=5025

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Ali Hatapçı, « From the page to the web: a survey of digital illustration archives », *Textes et contextes* [], 19-2 | 2024, publié le 15 décembre 2024 et consulté le 30 janvier 2026. Droits d'auteur : Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.. URL : http://preo.ube.fr/textesetcontextes/index.php?id=5025

La revue *Textes et contextes* autorise et encourage le dépôt de ce pdf dans des archives ouvertes.

PREO

PREO est une plateforme de diffusion voie diamant.

# From the page to the web: a survey of digital illustration archives

*De la page au web : panorama des archives d'illustrations numériques*


## *Textes et contextes*

Article publié le 15 décembre 2024.

**19-2 | 2024**
**Iconomorphoses : appropriation, éthique et partage - Représentations du monde hispanique actuel dans les séries télévisées**



**Ali Hatapçı**



http://preo.ube.fr/textesetcontextes/index.php?id=5025



Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.


---



---

1 Writing in 2012, Paul Goldman pointed out to the problem of "original material" in illustration studies. "In simplest terms," he writes, "one cannot study a subject if the raw materials are not available relatively easily and equally importantly internationally" (Goldman 2016, 20). This was only five years after one of the early scholarly digital illustration archives, the Database of Mid-Victorian Illustration (DMVI), was published online in 2007 (Thomas 2017, Ch 1, note 8). In a matter of years, an unprecedented number of illustrations began to pour onto the web. Nowviskie observes that "hand-crafted, boutique digit-


Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

ization by humanities scholars and archivists (in the intrepid, research-oriented, hypothesis-testing mode of the 1990s) was jarred and overwhelmed" (Nowviskie 2015, 386). As I would like to show in this article, with the increasing availability of large collections of illustrations from the 2010s, the concern over the availability of original material has been replaced by the need to make the existing material useable and useful. Since the mid-2010s, millions of illustrations have been extracted from the scanned pages of books, periodicals, newspapers from the fifteenth to early twentieth century. Computer scientists have not only proved the power of computers, but they have also confirmed the importance and quasi-ubiquity of visual culture in the previous centuries. It is now up to the (digital) humanists to tame this growing ocean of illustrations.

2 For a long time, the text monopolised attention in digital humanities. Text-mining, concordance analysis, voyant tools have become familiar terms in the field. This central place occupied by the text in digital humanities is largely due to the fact that it is the most widely available source thanks to the advance of the Optical Character Recognition (OCR) technology. OCR is a machine learning model trained to recognize the characters (letter, numbers and other symbols) on the photograph of a page. These characters are then added over the photograph as a layer, enabling us to see the same thing as the computer. It is no wonder that this technology which prioritises the text over any other element on the page promoted such an avid interest in the text. It would, however, be unfair to put the blame for the obliteration of illustrations entirely on the OCR technology. Julia Thomas contends that the modern editorial practices have imposed an imbalance between the text and the illustration by cutting out the visual content from the literary works (Thomas 2017: 17-19). This gave the illusion of the mid-Victorian period as an age dominated by the text whilst it was actually the golden age of illustration. [1]

3 Nevertheless, a visual turn has been underway in several fields including digital humanities. In periodical studies, [2] in the history of science, [3] the interest in visual elements has been reinvigorated. The availability and growing number of digital archives contribute to this renewed interest in the illustration and photography on the printed page. [4] The visual turn is most clearly observed in digital humanities as it is apparent from the Appendix listing about thirty projects. [5]

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

New projects tap into the experience of the existing ones to develop their own answers to the challenges previous projects faced. Every project, on the other hand, is different in its aims which inform the way the material is selected, treated, presented, and the functionalities offered to the "users".[6] To such an extent that the same material might be treated differently in different projects. In other words, as Mussell contends, every digital humanities project is an editorial project which involves a considerable amount of "editorial practice" towards their aims.[7] The aims, material treated, limitations and challenges of most of these projects are recounted on their websites and, albeit much fewer, in a monograph or journal articles. Julia Thomas's *Nineteenth-Century Illustration and the Digital: Studies in Word and Image* stands out in this literature not only as an overview of the important questions in the creation of digital visual archives, but also as a guide to the field. As the Principal Investigator of two projects (DMVI and the *Illustration Archive*), Thomas provides insights into the relationship between the text and the image while detailing the making of digital archives. In less than a decade since the publication of this important work, however, the advances in machine learning technology gave rise to another generation of digital archives, and thus a need for a new review of the field. Besides incorporating recent projects in its analysis, this article seeks to understand how digital archives of illustrations have been constructed in the last decade. In this article, as a first step towards the construction of a digital database of the nineteenth-century scientific illustration, I explore the landscape of visual projects. Inspired by Barman *et al.*'s (2021) paper which investigates the user interfaces of existing digital newspaper archives, this article examines a selection of digital humanities projects which take visual content from publications as source. By analysing the solutions these finalised or on-going projects offer regarding the extraction of visual content, its presentation on a digital platform, and its discoverability, this article seeks to contribute both to visual studies and to digital humanities.

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

# 1. Detecting and extracting images

4 Despite its initial development with the textual material in mind, OCR engine has been shown to be an effective tool for detecting visual content in scanned pages. Digital archives of illustrations extracted from published material make use of OCR data to detect visual content on the photos of the printed pages. The OCR process produces an XML file which, Leetaru explains, "contains an enormous wealth of information *normally invisible to users of the consumer version*".[8] Most importantly for the purpose of extracting visual content from scanned books and periodicals, it contains information indicating the boundaries of non-textual content in the scanned page. While denoting the parts of the page which they omit due to the absence of text, OCR engines demarcate the boundaries of the regions where there might be visual content. The irony was not lost on the scholars of illustration studies who lament the undue supremacy of the text over the image. "The identification and isolation of the illustration in the digitised and OCRed… pages were enabled not so much because of their visual difference but because of their negative relation to the text: they were recognised as images because they were not words" (Thomas 2017: 35).

5 Early digital archives of visual content from historical books and periodicals relied exclusively on OCR data to detect and extract images. In recent projects, however, more advanced techniques incorporating custom machine learning models complement it. A decade ago, two large projects, each taking a different corpus of scanned documents as its source, brought this aspect of the OCR to limelight. First of these was Ben O'Steen's *Mechanical Curator*.[9] Although unconventional in its logic of showing content "*almost* randomly" and not aiming at the discoverability of the items, this project required the extraction of images from scanned books.[10] O'Steen used the Microsoft Books/BL collection which included 25 million pages, 68,000 volumes, 65,000 titles scanned by Microsoft.[11] In his quest to develop a cost-effective and simple method for extracting visual material from this vast corpus, he discovered that the regions of the pages which the OCR engine predicted to contain visual elements were

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

marked with the annotation “GraphicalElement” in the XML file containing OCR data.[12] The result is a corpus of about a million illustrations and photographs extracted from the Microsoft/BL collection, available for viewing on Flickr.[13]

6 The second project came a few months after O’Steen’s large batch was uploaded to Flickr. Taking a larger corpus as its source, Kalev Leetaru’s *500 years of the images of the World’s Books*, sought to extract visual content from 600 million digitised books available on the Internet Archive.[14] Examining the OCR XML files which are provided for almost each scanned publication on the Internet Archive, Leetaru found that the regions of the page which the proprietary Abbyy OCR engine “thought” to include visual content were marked as “Picture” block.[15] Like O’Steen a few months earlier, Leetaru developed a software which detects the picture blocks in each page, if there is any, and then extracts the corresponding coordinates of the page from the photograph. The result is a corpus of visual material containing about five million images hosted on Flickr.[16]

7 Today, the process of using the XML file containing OCR data, available for the majority of items hosted on the Internet Archive has been simplified. Anyone with a limited knowledge of the command line and the popular programming language Python can conduct their own projects (Krewson 2019). It must, however, be recognized that OCR does not always produce the best results. This is especially the case with hybrid formats such as newspapers and periodicals in which the content is usually organised in several columns. To overcome this difficulty, more sophisticated solutions to visual content detection have recently been developed.

8 The recent availability of artificial intelligence chat tools for public use has drawn a great deal of attention to machine learning. As its name suggests, this technology consists in “teaching” the computer with a training dataset based on the analysis of which the computer creates an algorithm and can analyse and classify data which it had not “seen” before. Instead of OCR data, more recent projects rely on machine learning to detect and extract visual content from the scans of the printed pages. This is especially the case with projects which involve the treatment of visual content from newspapers and periodicals. Due to their multi-column organisation, these hybrid formats

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d’être soumis à des autorisations d’usage spécifiques.

tend to defy even the recent improvements on OCR engines. From many perspectives, the machine learning technology and its application to visual digital humanities makes enormous projects including millions of images more viable than before. It does not, however, diminish the importance of human contribution. The success of a machine learning model depends on the quality of annotations in the training dataset as well as the diversity of its content. The images in the training dataset need to be representative of the corpus on which the model is intended to run. If a training dataset is based largely on newspapers from the early twentieth century, it might not produce the best results when the machine learning model is fed newspapers from the nineteenth century. Also, a model which classifies images requires a training dataset which includes annotated images for each category proportionate to the scale of the corpus. Finally, the quality and precision of the image annotations play a vital role in the outcome. Therefore, while the creation of digital archives consisting of millions of images which are more discoverable than ever has become a possibility due the minimised need for human input, for annotating a small part of the corpus rather than its entirety, the need for expertise has become ever greater. Besides technological expertise, better guided annotators and a good knowledge of the corpus to select the documents or pages which represent it best have become very important in the creation of the training dataset. In this sense, machine learning, ever more than the previous decade's large-scale projects which relied mainly on the expertise of computer scientists, brings the two sides of the digital humanities together.

9 One of the recent digital archives of newspaper illustrations, the *Newspaper Navigator* based on the images from the *Chronicling America* newspaper archive adopts such an approach.[17] To detect the visual content in 16 million pages of American newspapers published between 1770 to 1963, instead of relying on the OCR data, Benjamin Lee developed a machine learning model trained on an augmented version of the *Beyond Words*, a crowdsourced project where volunteers were asked to identify a variety of content on scanned newspaper pages (Lee *et al.* 2020). Lee's model does not only detect visual content on the scanned pages, but it also classifies them. However, the *Beyond Words* dataset containing 3,437 images with 6,732 verified annotations did not, Lee found, proportionately represent the six cat-

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

egories into which he wanted to classify the images. In order to increase the number of these underrepresented categories in the dataset Lee added 32,424 annotations to pages including headlines, advertisements and maps. By using this augmented version of the *Beyond Words* data as training data, Lee, in his *Newspaper Navigator*, uses machine learning to impose the following categories on the images: photograph, illustration, map, comics/cartoon, editorial cartoon, headline, advertisement. Unlike Leetaru's project which does not impose any classification on the illustrations, albeit a great asset for a project involving newspapers which include a variety of images, *Newspaper Navigator* opens up new possibilities for the treatment of digitised pages. This project also provides important insights into the best practices in the creation of a training dataset. For instance, as the *Beyond Words* dataset only includes annotations on the newspapers from the WWI era, Lee's model does not perform so well on scanned pages of the newspapers from the nineteenth century as it does on those from the twentieth (Lee *et al.* 2020: 3061). In its final form, the *Newspaper Navigator* includes about 1.5 million items of visual content from newspapers published between 1900 and 1963. [18]

10 *Newspaper Navigator* exemplifies the current interest in machine learning models to detect and extract visual content from digital archives. [19] Despite the attempts to gamify and optimise citizen science platforms, the larger the datasets grow and the more tedious the tasks demanded from volunteers become. The difficulty of motivating volunteers for time consuming tasks such as transcription has grown the need to develop methods which minimise human intervention while at the same time optimising the results. Only with machine learning models can we imagine an archive of big data whose contents can be discovered. In *Newspaper Navigator*, for instance, the ratio of the training dataset to the corpus is about 3%. [20] Recent studies promise that the need for large training datasets, hence large-scale volunteer effort, might soon be something of the past with the development of new methods which minimise the size of the training dataset (Barman *et al.* 2021). This might further increase the importance of the selection of the items in the training datasets, hence the expertise of the literary scholars and historians.

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

# 2. Platforms

11 Extraction of the visual content from printed pages is now a relatively straightforward process, and it is rarely the ultimate aim of a visual archive project. The aspect of a project which determines its lifecycle, aims, and publics is related to the platform on which the images are made available (Tabak 2017). Some projects like *Science Gossip* (see below) have a predetermined lifespan during which they aim at collecting data about their corpora of images through crowdsourcing. The output is datasets. Others, usually with larger datasets, opt for an "in-progress" approach which allows both data collection and consultation for an undetermined period of time on a dedicated or generic platform. In this second case, the choice of platform, its interface determine the limitations and potentials of a project. Here I first treat the issues regarding the interface and search parameters and then focus on the questions related to permanence and re-usability.

## 2.1. Interfaces

12 One common feature of the large-scale projects which began around the mid-2010s was that they did not rely on custom interfaces for their repositories of images. Instead, both Kalev Leetaru and Ben O'Steen uploaded the output of their projects, five million and one million illustrations, respectively, to Flickr.[21] The illustrations extracted from the Biodiversity Heritage Library (BHL) as part of the "Art of Life" project were also uploaded to Flickr.[22] As a social platform optimised for sharing visual content, Flickr provides a venue where the output of visual digital humanities projects can be viewed, tagged, and shared. Although historical illustrations on Flickr theoretically have a better chance of being viewed by a larger public than if they had been deposited on a custom digital archive, from several perspectives Flickr poses problems.

13 The *Illustration Archive* launched in 2015 by a group of researchers from Cardiff University is one of the few and probably the best example of an "in-progress" digital archive.[23] The team led by Julia Thomas seeks to make one million illustrations extracted by O'Steen from the Microsoft/BL collection into a searchable illustration archive. As mentioned above, this corpus is already available on Flickr

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

and the existence of the *Illustration Archive* is a clear sign that the limited functionalities of Flickr do not meet the requirements of every research project, although the tags added to the items on Flickr are also fed into the *Illustration Archive* (Thomas 2017: Ch 1, note 43). On the *Illustration Archive*, visitors to the website can both contribute to the archive by annotating the illustrations through a guided questionnaire and consult the illustrations. In this part, I only focus on the interface and search functionality of the website by comparing it to Flickr, and below I treat its annotation interface.

14 The bespoke search functionality of the *Illustration Archive* includes the following search fields: keyword, illustrator, author, book title, publisher, place of publication, publication date by range or by decade.[24] This is a significant improvement on the limited number of search fields which Flickr offers. First, Flickr's search interface is not intuitive. In the British Library's (BL) Flickr photostream which largely includes the same items as the *Illustration Archive*, the user needs to click on the magnifying glass just above the images in the photostream, instead of the search bar at the top of the website, to run a query within the BL collection. Furthermore, to the frustration of illustration scholars, Flickr is biased against illustrations. Searching for "girl" within the BL photostream produces no results because the user first needs to click on "Advanced" search and tick "illustration/art" from the "content" types which is by default unticked, unlike "photos" and "videos" (Thomas 2017: 22). Secondly, Flickr's search function is very limited and certainly not suitable for historical research. It can be difficult on the search interface to find illustrations from the same book unless they were put together in a collection by the owner of the account. Similarly for the search between date ranges, Flickr offers only two options "date taken" and "date uploaded", there is no option to indicate such essential information as the date of publication. For these reasons, Flickr is far from meeting the basic requirements of scientific research on the illustrations unless the user is quite adept at programming and can use Flickr API.

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

## 2.2. Permanence of records and the re-usability of data

15 Flickr has been one of the popular platforms for researchers working with visual material (Spyrou and Mylonas 2016). Launched in 2004, like Facebook, and purchased by Yahoo the following year, Flickr soon became one of the prodigious children of the Web 2.0. The 2010s, nevertheless, saw the entry of new competitors such as Instagram in the market of visual content and Flickr again changed ownership in 2017. During this period, several features such as free storage space offered, and search parameters also went through changes. [25] When Flickr announced that it was going to delete the volume of photos exceeding its new freemium limit in 2018, it was met with criticism. [26] Its audience is estimated to have thinned to "an old community of photographers," and its prospects as a business are assessed to be bleak, unless it "come[s] up with a new and revolutionary feature,". [27] We have very recently witnessed the drastic changes Twitter went through after its sale, and it is always a possibility that any Web 2.0 platform might go through similar experiences. It can be argued that from the perspective of the permanence of records also, Flickr is probably not the best solution.

16 Nevertheless, the problem of longevity is not only limited to the privately-owned generic content-hosting platforms. Custom digital platforms, too, run the risk of becoming obsolete in a short period of time, requiring a partial and sometimes complete renovation, or abandonment before or after the completion (Solberg *et al.* 2021: 23). Visitors to the *Illustration Archive*'s tagging interface might notice that the image for "advertisement" option is missing in the initial questions of classification. [28] Furthermore, after spending some time on the website, the user is asked if they want to fill out a feedback form, nevertheless, this form turns out to be empty. More importantly, the server gives an error when the user tries to sign up. As a result, more advanced features of the platform such as creating user collections are currently not available. This lack of maintenance becomes all the more unfortunate as the *Illustration Archive* is an ongoing project still seeking to crowdsource annotations from volunteers.

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

17 These two examples show that platforms might falter, become outmoded, or remain unmaintained. To overcome such a situation and to save the efforts made throughout the lifecycle of the project, DH projects must make the data they collected as well as their source code available to the public. In this way, in the event that the interface is no longer available, the possibility of using the data remains intact. This also enables other researchers to use the data for purposes other than the original aim of a DH project, for instance, as a training dataset. Fortunately, the data from the majority of digital visual archives have been made available to public. Recent projects almost by default share their source codes as well as their outputs, and older projects follow their lead.[29] The data including tags and descriptions, as well as the images themselves hosted on Flickr can also be downloaded for more convenient manipulation by using its API (Application Programming Interface).[30]

18 Permanence and reusability have more recently become a bigger concern than interfaces. A new trend is afoot in the digital visual archives which aim at providing only the data and a script to the users in the form of a Jupyter Notebook.[31] Besides reusability, this move seems to favour environmental concerns. Lee, for instance, calculates the carbon emission as a result of the implementation of his *Newpaper Navigator*, about 380 $CO_2$, comparable to the carbon emission of one person flying from Washington, D.C. to Boston (Lee 2020: 29).

# 3. Making Images Discoverable

19 The difference between an archive and a pile is a catalogue or an index which helps finding what we look for. Once the visual content is detected and extracted from scanned files, how to make those image files discoverable? The problem posed by discoverability far supersedes that of the extraction of images. Here again there is no solution which fits all. The objectives of the project, the layout of the content on the pages, financial and human resources available enter into the equation. Most of the early projects depended solely on the OCR output to discover the visual content extracted from publications. Relatively smaller projects which outlined a clear historical research perspective employed labour-intensive, standardised and manual

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

method of annotating visual content. Some projects, on the other hand, opted for a continuous annotation by volunteers on their own bespoke platforms. In recent years, more sophisticated methods have been developed.

## 3.1. Metadata

20 Literally meaning data about data, metadata have enabled humans and computers to communicate through a standardized structure. In the libraries, museums, archives, and repositories, metadata impose an order on the items, making them accessible to the readers, visitors, and users (Gartner 2021: 1-6). According to Buckland, metadata has two main purposes. First, it describes technical (format, size, colours, etc) aspects, administrative (copyright information etc.) limitations, and the content (period, the subject, author, etc.) of an item. This information is usually recorded in standard descriptive formats such as Dublin Core to facilitate storing and promote interoperability with other formats.[32] The second purpose of metadata is to enable and facilitate the discovery of documents in the repository. In this case, the relationship between the data and metadata is reversed because the user has the option to "*start* with a query or with the description rather than the document—with the metadata rather than the data—when searching in an index" (Buckland 2017: 118). For instance, in a digital archive of illustrations from Buffon's *Histoire Naturelle*, a user should thus be able to search for "insects" or "birds" which should be included in the metadata attached to the individual files even if the captions of the illustrations might not include the common names of the species they describe. The metadata can also include colours used in the illustration as well as the technique employed such as wood engraving or photography. Inclusion of such information transforms the metadata from a secondary position to a primary source of information.

21 In a DH project, probably the most important step is the design of the metadata fields and schemes. Drucker contends that this is "where the intellectual and conceptual modelling of research project takes place" (Drucker, 2021, 53). What information to include, to what detail in the metadata are crucial and might be controversial in the project design. Although there are well-established metadata schemes and

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

classification systems, they might not be suitable to a particular project. For instance, the classification system based on the twenty-six letters of the Roman alphabet used by the Library of Congress (LoC) reserves the letters E and F for the history of Americas and only one letter, D, for the history of the rest of the world (Le Deuff 2018: Ch 8). It would probably not be very suitable for a library outside of the US to follow the LoC classification exactly. For the classification of illustrations however we need to look elsewhere. *Iconclass*, "the most widely accepted scientific tool for the description and retrieval of subjects represented in images (works of art, book illustrations, reproductions, photographs, etc.)," can be considered one of the options. [33] In spite of its detailed and wide-ranging definitions however, *Iconclass* is considered scarcely adaptable to describe illustrations. Invented in mid-twentieth century by the art historian Henri van de Waal, Iconclass is an opiniated classification of visual art into "28,000 hierarchically structured definitions within 10 main divisions" (Thomas 2017: 57). Although it has gone through frequent modifications to accommodate such visuals as novel artistic imagery, it prioritises fine art over popular visual forms. "A browse through the classification system begins with the category 'Religion and Magic', while 'The Bible' and 'Classical Mythology and Ancient History' feature as part of the ten main iconographic categories" (Thomas 2017: 57).

22 Due to these reasons, Thomas explains that her team opted not to use Iconclass in the classification of the illustrations in the DMVI. [34] Launched in 2004 and completed in 2007, making it the earliest digital archive of historical illustrations, the DMVI contains 868 "literary illustrations" extracted from books and periodicals published in 1862. [35] This limitation is explained both by the coverage of the illustration collections cut by Victorian collectors which DMVI uses as its source and by the desire to show the abundance and diversity of wood-engraving illustrations in their heyday in the 1860s. [36] The DMVI project divides its items into seven main categories (periods, geography, settings, people, activities, objects, themes) and then into 1,123 hierarchical categories. The category "holidays" under "travel and tourism" under the main category "themes," for instance, yields thirty-six illustrations, while "acrobatics" under "physical motions" under "actions and speech" under "activities" main category produces only two results. Thomas explains that "The relatively small and uni-

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

form corpus of material in DMVI enabled us to develop the classification for the iconographic search by a prior analysis of the images rather than fitting the illustrations into preconceived categories."[37] Indeed in a large dataset such an approach departing from the illustration to classification is unfeasible. The growing number of historical illustrations extracted in the last decade prompted the development of several methods with varying degrees of success.

## 3.2. Crowdsourcing

23 With the availability of large amounts of illustrations from the mid-2010s and the popularity of social media outlets such as Flickr as well as dedicated citizen science platforms such as Zooniverse, a new method of making images discoverable emerged: crowdsourcing.[38] Two main practices of crowdsourcing can be discerned in the existing projects based on their demands from the volunteers: Folksonomy and citizen science. Defined as "A user-generated system of classifying and organizing online content into different categories by the use of metadata such as electronic tags," folksonomy has quickly become one of the ways for dealing with large corpora of visual material.[39] Although projects seeking to understand publics' engagement with art have produced interesting results, folksonomy is far from providing a standard description of images (Thomas 2017: 71-4). When applied to the "million images" of the Microsoft/BL collection on Flickr, free tagging produced hardly useable results.

24 Citizen science projects, on the other hand, demand more specific tasks from volunteers, and so the volunteers are provided with a guide or training. Apart from the convenience of providing a ready-made platform for visual content, Flickr is also a social media platform. Projects have been developed to tap into this social aspect of Flickr. The "Art of Life" project (2012-5) which uses the illustrations from the holdings of the Biodiversity Heritage Library (BHL) was one of the early projects to crowdsource the annotation of about 300,000 illustrations in its Flickr photostream.[40] A nineteen-page guide to tagging BHL illustrations on Flickr was produced and is available on the BHL blog. In this guide, the volunteer is given detailed instructions how to "read" a historical scientific illustration. Step three dwells on how to identify the name of the species; step four explains

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

how to “machine tag” species in the form of “taxonomy:binomial="Genus species"” or with the name of the illustrator, for instance, “engraver:name="[Givenname Middlename Familyname]"”, so that these tags can be differentiated from folksonomic tags.[41] At level three of the fourth step, volunteers are asked to “box tag” the illustrations if there are multiple species in the illustration. Further advice is provided in the document to find the current taxonomic names of the species as well as their common names and to enter this information as tags.[42] As of September 2023, almost a decade after the completion of the Art of Life project, out of about 300,000 images, around 52,000 have been tagged with at least one species name, and 30,000 with the name of the artist.

25 In contrast with other collections treated in this article, some of the content in the BHL Flickr photostream is organised under collections. This is especially the case with periodicals and multi-volume works which are organised in 227 collections.[43] Volumes of periodicals like the well-known *Curstis's Botanical Magazine* are organised further into sub-collections, one for each year.[44] Because the search function of Flickr is not geared towards researchers' needs, organisation of content in collections is the only way to enable the user to see entirety of the visual content in a volume.

26 Flickr is lacking from many perspectives to provide a platform for scholarly presentation and crowdsourcing of visual content. As a result, DH projects with defined research objectives either use more adaptive platforms such as Zooniverse or create their own bespoke platform. One of the first Zooniverse projects to go viral, “Science Gossip” as part of “Constructing Scientific Communities” (2014-9) project, asked the members of the public to annotate the scanned pages from nineteenth-century science periodicals on Zooniverse.[45] The “Science Gossip” project produced 34,108 annotations for 10,535 pages including at least one illustration from sixteen periodicals.[46] Although the project did not include the creation of an archival platform where users could consult this output, it constitutes an important source of training data for the future projects on the visual content of nineteenth-century periodicals.

27 *Illustration Archive*, on the other hand, has its own bespoke platform. The visitors to the website are invited to classify and tag illustrations

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

shown to them in a guided manner. In the first step, the volunteer is asked to choose one of the ten pre-defined categories (advertisement, portrait, decoration, title page, location, map, scientific, literature, photograph or none of these) which describe the illustration best. A second question follows to put the illustration in a subcategory. For instance, location is followed by a two-choice question: "by name" or "on a map"; while scientific is followed by eight subcategories: geological, medical, engineering, botanical, zoological, archaeological, architectural, and none of these. After this two-step classification of the illustration, the user is asked to enter tags and it is especially here that the *Illustration Archive* makes an improvement on the existing crowdsourcing platforms. Its tagging interface is connected to WordNet, a large lexical database of English where "Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept".[47] When the user enters a tag, a list of choices comes up and the user is asked to choose the most suitable one. According to Thomas, the benefits of this link with WordNet are twofold: First, it improves the collected data by recording at the backend of the Archive, not only the term entered but also its broader categories up to three inherited hypernyms. In this way, when the user enters "ship" it is recorded with the broader categories of "vessel, watercraft", "vehicle", and "transport". This does not only ensure a more precise description of the illustrations and link them semantically, but it also facilitates the task of the tagger and reduces the amount of labour required for the description of images. The second benefit of this approach is to curtail the vocabulary of the user-generated tags and standardise them (Thomas 2017: 75-6). On the other hand, this approach has significant limitations, especially for projects which use more specialist illustrations as source. First, the user is limited to one-word tags or phrases which exist in the WordNet database. Furthermore, WordNet is only suitable for more general descriptions. For a digital archive of scientific illustrations, for instance, a combination of Encyclopaedia of Life (EoL) and WordNet would provide a better source.[48] Finally, although the tagging process concludes with the typing of the caption of the illustration, it omits the artist.

28 The final aspect to consider in crowdsourcing of illustration tags regards the validation of tags by several users. Thomas observes that

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

"In Flickr, the inconsistency and instability of the 'language of retrieval', which comes about because the images are tagged differently by different contributors, means that relevant images will not always be retrieved" (Thomas 2017: 22). While platforms like Flickr which are not specially geared towards scientific projects do not provide such functionalities, it is one of the defining features of bespoke crowdsourcing platforms and dedicated citizen science platforms. The final question in the *Illustration Archive* asks the user to confirm the existing tags by other users on the illustration and remove the incorrect ones. For instance, on the most popular citizen science platform Zooniverse, validation by agreement is called "subject retirement" and functions as a project parameter to "Decide how many people you want to complete each task". Zooniverse advises validation by five to ten volunteers for tagging, and by three to five volunteers for transcription tasks before an illustration is retired, hence accepted as completed.[49] On the other hand, although essential, validation by agreement increases the need for volunteer engagement and delays the completion of a project.

## 3.3. Text around the image

29 Assuming that the text around the illustration should describe it, some scholars have developed tools and methods to make the text sandwiching the illustration speak for it. One of the early big data projects on historical illustrations, Leetaru's database of historical illustrations offered a solution. Leetaru's aim was to 'make them all browseable and searchable (via both the metadata of the original book and the text surrounding each image), "reimagining" the world's books'.[50] To enable users to search the millions of images extracted from the scans of printed pages on Flickr, alongside such information as the unique Internet Archive identifier for the item, page number, URL, Leetaru added 1000-character text chunks preceding and following the illustration to the description of each illustration on Flickr.[51] This method, however, raises a number of questions. Even if it is assumed that the 1,000 characters preceding and following the illustrations satisfactorily describe them, it is also possible that these chunks of text sandwiching the illustration will also include other "keywords," which give rise to false positives. Furthermore, without other search parameters such as limitations on dates and publication

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

titles, the user is only bombarded with illustrations without any anchor points.

30 Leetaru's method of identifying illustrations by extracting the text around them is used in other projects. SherlockNet, which became finalist of the BL Labs Competition in 2016, combined Leetaru's approach with machine learning to make the Microsoft/BL collection on Flickr discoverable. However, instead of the 2,000 characters sandwiching the illustration, SherlockNet utilises the text on the whole of the preceding page, on the page of the illustration, and the text on the following page to describe the illustration in a limited number of tags.[52] The noun phrases extracted from these three pages through Natural Language Toolkit (NLTK) are then compared with those extracted from the pages sandwiching twenty other illustrations in the collection which the algorithms have determined to be the most similar.[53] Currently about 836,000 illustrations carry at least one "sherlocknet:tag=" on BL Flickr photo stream.[54] A search for "cow" in the BL Flickr stream produces some cows but also a considerable number of false positives, for instance an illustration of zebras. On the inspection of SherlockNet tags, it can be observed that the automated tagging process also included such unrelated tags as port, bird, antelope, stone, and none for zebra. The query for cow yields the illustration because the folksonomic tags include cow but also zebra.[55]

31 More advanced projects employing the state-of-the-art machine learning methods and adapted to the source material have emerged more recently. The Library of Congress (LoC) *News Navigator* developed by Benjamin Charles Germain Lee in 2020 includes 1.56 million newspaper photos sourced from LoC's *Chronicling America* project, which includes 16.3 million pages from newspapers published between 1770 and 1963.[56] To improve the searchability of the images, Lee leverages both the XML files resulting from OCR process and *Beyond Words* training dataset. In *Beyond Words* dataset, volunteers were asked to annotate scanned pages to indicate the textual content in the boxes on the newspaper pages where images occur. Assuming that the textual content sharing the same box with the visual content describes the visual, as captions or titles, *Newspaper Navigator*, like Leetaru's project but somewhat more finely, relies on the text around the visual content to describe and retrieve it. However, as the visual

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

content in newspapers are likely to be described more precisely by the captions around them than illustrations in books in diverse subjects published through a long period and in varying formats, the outcome is much more satisfactory.

32 There are two points to underline from these experiments. First, the tools and methods at our disposal to describe images and make them discoverable based on the text around them is not yet adequate. This is especially the case for collections like that of Microsoft/BL, eclectic both in terms of the formats of publications from which the illustrations were culled, and their genres. Indeed, the relationship between the text and the image seems to vary with genre, period, and format of publication. In literary illustration, where captions are rare, Thomas contends, relying on the text around the illustration does not always produce the expected results. The text referring to the illustration might appear several pages before or after the illustration or the illustration might describe the entirety of the work rather than a specific part of the text (Thomas 2017: 47-53). In scientific illustration, especially in botany, the illustration is also dependent on the text for description (Chansigaud 2009: 7, 10). Saunders, however, observes a change through time in the relationship between the text and the image in botanical illustration. With the establishment of a standard, universal language of description in the eighteenth century by Linnaeus "many works of botanical theory dispensed with illustrations altogether, though they remained essential for books of a practical, descriptive, decorative or reference nature" (Saunders 1995: 8). Where botanical illustrations appear, however, they are likely to be accompanied by an adjacent caption or description.[57] It can therefore be argued that in a digital archive, botanical illustrations, more than literary illustrations, might benefit from a combination of the methods mentioned above for the description of its contents.

33 Recent projects redefine the relationship between the text and image by providing an option to search for images based on their similarity. Lee's *Newspaper Navigator* enables user to compare images based solely on image embeddings, that is, a numerical description of the semantic content of the image in computer-readable format.[58] The users can create their own collections by picking images and using this collection as a training dataset to search for similar images in the database. The bigger is the size of the training dataset in the user's

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

collection, the better are the results. Image similarity detection is a developing field which might open up new ways to study large collections of digitised visual content and might decrease to a limited extent our dependence on the text to describe the images, for in this case the images speak for themselves.

# Conclusion

34 Digital archives challenge the way we have been studying illustrations, and even only for that reason they are invaluable sources. In what context could the relationship between text and image be better analysed than the problem of describing digitised images? They encourage the researchers to engage with the public in the framework of citizen science projects. But they are not only challenging. They also contribute to our understanding of the publishing history by showing that visual culture was not the reserve of the twentieth and twenty-first centuries. A greater portion of the public beyond the specialists have a chance to see, be inspired by this heritage of visual culture. They bring illustrations to the comfort of our homes, which encourages research. They also encourage collaboration between experts from different disciplines. Looking back at the past decade of the digital archives, it can be argued that the future is promising for larger and more useable archives.

35 Digital illustration archives have gone through four stages in the last three decades. The small-scale, hypothesis-oriented projects of the late twentieth and early twenty-first centuries were succeeded by large-scale technology-dominated projects, only to invite human intervention again in the form of voluntary classifications in the 2010s. Today, the scale of human intervention is not any less important, but its scale has been limited to the making of training datasets. Illustration studies is once again going through a drastic change, this time combining the scale and usability, as well as the specialities of the computer scientists and humanists.

---

'19th Century Digitised Books Collection'. OCR text dataset. British Library Research Repository, 2014. https://bl.iro.bl.uk/collections/b7fd2482-debd-44

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

5-9494-72aa2ead00bb?locale=en.

Barman, Raphaël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan (2021). 'Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers'. *Journal of Data Mining & Digital Humanities*: *HistoInformatics*, 6107, 1-26. https://doi.org/10.46298/jdmdh.6107.

Belknap, Geoffrey (2016). *From a Photograph: Authenticity, Science and the Periodical Press, 1870-1890*, London: Bloomsbury Academic.

Brake, Laurel, and Marysa Demoor (2009). 'Introduction: The Lure of Illustration,' in Brake, Laurel / Demoor, Marysa, Eds., *The Lure of Illustration in the Nineteenth Century: Picture and Press*, Basingstoke, England; New York: Palgrave Macmillan, 1–13.

Brian Maidment (2008). '*The Database of Mid-Victorian Wood-Engraved Illustration* (Review),' in *Journal of Victorian Culture* 13, no. 1: 108–13. https://doi.org/10.1353/jvc.0.0015.

Buckland, Michael Keeble (2017). *Information and Society* (= The MIT Press Essential Knowledge Series), Cambridge, Massachusetts London, England: The MIT Press.

Chansigaud, Valérie (2009). *Histoire de l'illustration Naturaliste*, Paris: Delachaux et Niestlé.

Chansigaud, Valérie (2016). 'Scientific Illustrators,' Lightman, Berman, Ed., *A Companion to the History of Science*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 111–25. https://doi.org/10.1002/9781118620762.ch8.

Drucker, Johanna (2021). *The Digital Humanities Coursebook*. Oxon: Routledge.

'History of the Database: Database of Mid-Victorian Illustration'. Available via: https://www.dmvi.org.uk/guide.php?page=13. Consulted 23 August 2023.

Fyfe, Paul, and Qian Ge (2018). 'Image Analytics and the Nineteenth-Century Illustrated Newspaper,' in *Journal of Cultural Analytics* 3/1, 1-25.

Gartner, Richard (2021). *Metadata in the Digital Library: Building an Integrated Strategy with* XML. London: fp, facet publishing.

Goldman, Paul (2016). 'Defining Illustration Studies,' in Goldman, Paul / Cooke, Simon, Eds. *Reading Victorian Illustration, 1855-1875: Spoils of the Lumber Room*, London: Ashgate, 13–32.

Goldman, Paul (1994). *Victorian Illustrated Books 1850 - 1870: The Heyday of Wood-Engraving; the Robin de Beaumont Collection*. London: British Museum Press, 1994.

Hughes, Damian (2022). *Picturing Ecology: Photography and the Birth of a New Science*. Singapore: Palgrave Macmillan.

Le Deuff, Olivier (2018). *Digital Humanities: History and Development*. London Hoboken: ISTE Wiley.

Lee, Benjamin Charles Germain (2020). 'Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset'. http://dx.doi.org/10.17613/k9gt-6685.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld (2020). 'The Newspaper Navigator

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America,' in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, New York, NY, USA: Association for Computing Machinery, 3055–62. https://doi.org/10.1145/3340531.3412767.

Leetaru, Kalev (2014). 'Using the Imagery of Books Extraction Tool v1.0. Technical Documentation'. Available via: https://ia904508.us.archive.org/21/items/vrr-texts-imageryofbooks-fullres-imageextractortool/USERGUIDE-IMAGEEXTRACTOR.pdf. Consulted 12 August 2023.

Maidment, Brian (2009). 'The *Illuminated Magazine* and the Triumph of Wood Engraving,' in Brake, Laurel / Demoor, Marysa, Eds., *The Lure of Illustration in the Nineteenth Century: Picture and Press*, Basingstoke, England; New York: Palgrave Macmillan, 17-39.

Mika, Katherine A., Joseph De Veer, and Constance Rinaldo (2017). 'Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data,' in *Biodiversity Informatics*, 12, 58-75.

Münster, Sander, and Melissa Terras (2020). 'The Visual Side of Digital Humanities: A Survey on Topics, Researchers, and Epistemic Cultures,' in *Digital Scholarship in the Humanities* 35/2, 366–89. https://doi.org/10.1093/llc/fqz022.

Mussell, James (2009). 'Science and the Timeliness of Reproduced Photographs in the Late Nineteenth-Century Periodical Press,' in Brake, Laurel / Demoor, Marysa, Eds., *The Lure of Illustration in the Nineteenth Century: Picture and Press*, Basingstoke, England; New York: Palgrave Macmillan, 203-19.

Mussell, James (2012). *The Nineteenth-Century Press in the Digital Age*. London: Palgrave Macmillan UK, 2012.

Mussell, James, and Suzanne Paylor (2012). 'Editions and Archives,' in *The Nineteenth-Century Press in the Digital Age*, London: Palgrave Macmillan UK, 114-48.

National Endowment for the Humanities. Interim Performance Report May 2014-Oct 2014, 29 November 2014. Available via: https://drive.google.com/file/d/1hSBA2xBq3RNg0owJjKmEGKeC-QGQeyraM/view?usp=sharing&usp=embed_facebook. Consulted 16 August 2023.

Nowviskie, Bethany (2015). 'Resistance in the Materials,' in Svensson, Patrik / Goldberg, David Theo, Eds., *Between Humanities and the Digital*, Cambridge, Mass. London: The MIT press, 386-9.

O'Steen, Ben. 'A Million First Steps'. British Library Digital Scholarship Blog, 12 December 2013. Available via: https://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html. Consulted 14 August 2023.

O'Steen, Ben. 'Peeking behind the Curtain of the Mechanical Curator'. British Library Digital Scholarship Blog, 7 October 2013. Available via: https://blogs.bl.uk/digital-scholarship/2013/10/peeking-behind-the-curtain-of-the-mechanical-curator.html. Consulted 14 August 2023.

Roued-Cunliffe, Henriette (2018). 'Developing Sustainable Open Heritage Datasets,' in Levenberg, Lewis /Neilson Tai / Rheams Davis, Eds., *Research Methods for the Digital Humanities*, Cham:

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Springer International Publishing, 287–307.

Saunders, Gill (1995). *Picturing Plants: An Analytical History of Botanical Illustration*, London: I.B. Tauris.

Solberg, Kaitlyn, Lisa Tweten, and Chelsea A. M. Gardner (2021). 'From Stone to Screen: The Built-in Obsolescence of Digitization'. in Ed. Shane Hawkins, *Access and Control in Digital Humanities*, London: Routledge, 22–40.

Spyrou, Evaggelos, and Phivos Mylonas (2016). 'A Survey on Flickr Multimedia Research Challenges,' in *Engineering Applications of Artificial Intelligence* (= Mining the Humanities: Technologies and Applications), 51, 71–91. https://doi.org/10.1016/j.engappai.2016.01.006.

Tabak, Edin (2017). 'A Hybrid Model for Managing DH Projects'. *Digital Humanities Quarterly* 11/1. Available via: http://www.digitalhumanities.org/dhq/vol/11/1/000284/000284.html. Consulted 22 August 2023.

Thomas, Julia (2017). *Nineteenth-Century Illustration and the Digital: Studies in Word and Image*, (= The Digital Nineteenth Century), London: Palgrave Macmillan.

Thomas, Julia (2007). 'Reflections on Illustrations: The Database of Mid-Victorian Wood-Engraved Illustration (DMVI)'. *Journal of Illustration Studies*, December 2007. Available via: https://web.archive.org/web/20150325063852/http://jois.uia.no/articles.php?article=37. Consulted 30 July 2013.

Topham, Jonathan R (2020). 'Redrawing the Image of Science: Technologies of Illustration and the Audiences for Scientific Periodicals in Britain, 1790–1840,' in Dawson, Gowan / Lightman, Berman / Shuttleworth, Sally / Topham, Jonathan R., Eds., *Science Periodicals in Nineteenth-Century Britain: Constructing Scientific Communities*, London: University of Chicago Press, 65–103.

Wevers, Melvin, and Thomas Smits (2019). 'The Visual Digital Turn: Using Neural Networks to Study Historical Images,' in *Digital Scholarship in the Humanities*, 35/1, 194-207. https://doi.org/10.1093/llc/fqy085.

# Digital Archive and URL

Anatomia: https://anatomia.library.utoronto.ca/

British Printed Images to 1700: www.bpi1700.org.uk

Broadside Ballads Online project: http://ballads.bodleian.ox.ac.uk

Cervantes Project: https://cervantes.tamu.edu

Database of Mid-Victorian Illustration (DMVI): https://www.dmvi.org.uk

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Database of Scientific Illustrators: https://dsi.hi.uni-stuttgart.de/index.php?function=show_static_page&id_static_page=1

Fleuron: https://fleuron.lib.cam.ac.uk/

Illustrated Classics of Engineering (NYPL): https://digitalcollections.nypl.org/collections/illustrated-classics-of-engineering-from-the-william-barclay-parsons-collection#/

Illustrating Scott: https://illustratingscott.lib.ed.ac.uk

IMPRESSO: https://impresso-project.ch/

Marie Duval Archive: https://www.marieduval.org

NCNA Image Analytics: https://ncna.dh.chass.ncsu.edu/imageanalytics/analyses.php

Newspaper Navigator: https://news-navigator.labs.loc.gov/

Nineteenth-Century Serials Edition: www.ncse.ac.uk

Photogrammar: https://photogrammar.org/maps

Project by Kalev Leetaru (Internet Archive illustrations): https://blog.archive.org/2014/08/29/millions-of-historic-images-posted-to-flickr/

Rossetti Archive: http://www.rossettiarchive.org

Science Gossip: https://www.sciencegossip.org

The Illustration Archive: https://illustrationarchive.cardiff.ac.uk

Ukiyo-e: https://ukiyo-e.org

Victorian Illustrated Shakespeare Archive: www.shakespeareillustration.org

Victorian400: Colorizing Victorian Illustrations: https://github.com/eliboo klover/Victorian400

Visual Geometry Group (Exploring British Library Images): http://www.robots.ox.ac.uk/~vgg/research/BL

Visual Haggard: http://www.visualhaggard.org

William Blake Archive: https://www.blakearchive.org

Yellow Nineties Online: http://www.1890s.ca

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

1 Paul Goldman, *Victorian Illustrated Books 1850 - 1870: The Heyday of Wood-Engraving* (London: British Museum Press, 1994).

2 Paul Fyfe and Qian Ge, 'Image Analytics and the Nineteenth-Century Illustrated Newspaper', *Journal of Cultural Analytics* 3, no. 1 (2018); James Mussell, 'Science and the Timeliness of Reproduced Photographs in the Late Nineteenth-Century Periodical Press', in *The Lure of Illustration in the Nineteenth Century: Picture and Press*, ed. Laurel Brake and Marysa Demoor (Basingstoke, England; New York: Palgrave Macmillan, 2009), 203–19; Brian Maidment, 'The *Illuminated Magazine* and the Triumph of Wood Engraving', in *The Lure of Illustration in the Nineteenth Century: Picture and Press*, ed. Laurel Brake and Marysa Demoor (Basingstoke, England; New York: Palgrave Macmillan, 2009), 17–39; Laurel Brake and Marysa Demoor, 'Introduction: The Lure of Illustration', in *The Lure of Illustration in the Nineteenth Century: Picture and Press*, ed. Laurel Brake and Marysa Demoor (Basingstoke, England; New York: Palgrave Macmillan, 2009), 1–13.

3 Damian Hughes, *Picturing Ecology: Photography and the Birth of a New Science* (Singapore: Palgrave Macmillan UK, 2022), https://doi.org/10.1007/978-981-19-2515-3; Jonathan R. Topham, 'Redrawing the Image of Science: Technologies of Illustration and the Audiences for Scientific Periodicals in Britain, 1790–1840', in *Science Periodicals in Nineteenth-Century Britain: Constructing Scientific Communities*, ed. Gowan Dawson et al. (University of Chicago Press, 2020), 65–103, https://doi.org/10.7208/chicago/9780226683461.003.0003; Geoffrey Belknap, *From a Photograph: Authenticity, Science and the Periodical Press, 1870-1890*, 2020, https://www.taylorfrancis.com/books/9781003103530; Valérie Chansigaud, 'Scientific Illustrators', in *A Companion to the History of Science*, ed. Bernard Lightman (Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016), 111–25, https://doi.org/10.1002/9781118620762.ch8; Gill Saunders, *Picturing Plants: An Analytical History of Botanical Illustration*, 2. ed (Chicago London: KWS, 2009).

4 For a discussion on 'edition' and 'archive' see James Mussell and Suzanne Paylor, 'Editions and Archives', in *The Nineteenth-Century Press in the Digital Age* (London: Palgrave Macmillan UK, 2012), 114–48, https://doi.org/10.1057/9780230365469.

5 Sander Münster and Melissa Terras, 'The Visual Side of Digital Humanities: A Survey on Topics, Researchers, and Epistemic Cultures', *Digital Schol-*

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

*arship in the Humanities* 35, no. 2 (1 June 2020): 366–89, https://doi.org/10.1093/llc/fqz022; Melvin Wevers and Thomas Smits, 'The Visual Digital Turn: Using Neural Networks to Study Historical Images', *Digital Scholarship in the Humanities*, 18 January 2019, https://doi.org/10.1093/llc/fqy085.

6 Mussell differentiates between the reader and the user: 'Whereas 'reading implies engagement with text, the term "user" suggests that those interacting with digital sources must also "do" something. The shift from "reader" to "user"', he argues, 'recognizes the transformation that print must undergo to provide access to text in digital form.' James Mussell, *The Nineteenth-Century Press in the Digital Age* (London: Palgrave Macmillan UK, 2012), 6, https://doi.org/10.1057/9780230365469.

7 James Mussell and Suzanne Paylor, 'Editions and Archives', in *The Nineteenth-Century Press in the Digital Age* (London: Palgrave Macmillan UK, 2012), 114–48, https://doi.org/10.1057/9780230365469.

8 Kalev Leetaru, 'Using the Imagery of Books Extraction Tool v1.0. Technical Documentation', 30 August 2014, https://ia904508.us.archive.org/21/items/vrr-texts-imageryofbooks-fullres-imageextractortool/USERGUIDE-IMAGEEXTRACTOR.pdf. (emphasis added)

9 "Ben O'Steen: Experiments at British Library Labs." National Digital Forum YouTube Channel, https://www.youtube.com/watch?app=desktop&v=bIXB0ROyxcY

10 Ben O'Steen, 'Peeking behind the Curtain of the Mechanical Curator', British Library Digital Scholarship Blog, 7 October 2013, https://blogs.bl.uk/digital-scholarship/2013/10/peeking-behind-the-curtain-of-the-mechanical-curator.html.

11 '19th Century Digitised Books Collection', OCR text dataset (British Library Research Repository, 2014), https://bl.iro.bl.uk/collections/b7fd2482-debd-4495-9494-72aa2ead00bb?locale=en https://www.bl.uk/collection-guides/digitised-printed-books http://news.bbc.co.uk/2/hi/technology/6213260.stm https://en.wikipedia.org/wiki/Live_Search_Books.

12 O'Steen, 'Peeking behind the Curtain of the Mechanical Curator'.

13 Ben O'Steen, 'A Million First Steps', British Library Digital Scholarship Blog, 12 December 2013, https://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html.

14 https://blog.gdeltproject.org/500-years-of-the-images-of-the-worlds-books-now-on-flickr/

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

15 Leetaru, 'Using the Imagery of Books Extraction Tool v1.0. Technical Documentation'.

16 https://blog.gdeltproject.org/500-years-of-the-images-of-the-worlds-books-now-on-flickr/

17 https://chroniclingamerica.loc.gov/; https://news-navigator.labs.loc.gov/search.

18 https://news-navigator.labs.loc.gov/search/about

19 For instance, see the digital archive of Swiss and Luxembourgois newspapers *Impresso* (https://impresso-project.ch).

20 https://github.com/LibraryOfCongress/newspaper-navigator

21 https://www.flickr.com/photos/internetarchivebookimages/; https://www.flickr.com/people/britishlibrary/.

22 There was a convergence between Leetaru's project and 'Art of Life' amounting to no less than 20% of the output of the latter. 'National Endowment for the Humanities. Interim Performance Report May 2014-Oct 2014', 29 November 2014, https://drive.google.com/file/d/1hSBA2xBq3RNg0owJjKmEGKeCQGQeyraM/view?usp=sharing&usp=embed_facebook.

23 https://illustrationarchive.cardiff.ac.uk/

24 https://illustrationarchive.cardiff.ac.uk/search_advanced

25 https://www.britannica.com/topic/Flickrcom; https://www.techspot.com/article/2384-flickr/.

26 https://www.nytimes.com/2018/11/29/style/digital-photo-storage-purge.html; https://www.vox.com/the-goods/2019/2/6/18214046/flickr-free-storage-ends-digital-photo-archive-history.

27 https://www.techspot.com/article/2384-flickr/

28 https://illustrationarchive.cardiff.ac.uk/image/11267641693

29 For the source code of Leetaru's project see https://ia904508.us.archive.org/21/items/vrr-texts-imageryofbooks-fullres-imageextractortool/USERGUIDE-IMAGEEXTRACTOR.pdf; for his output data see https://blog.gdeltproject.org/500-years-of-images-from-the-worlds-books-as-an-ai-training-dataset/. For O'Steen's source code see https://github.com/BL-Labs/BL-Flickr-1-Million-Images_Update-April-2020 and https://github.com/BL-Labs/embellishments ; for his output data see https://github.com/BL-Labs/imagedirectory. For the source code of the Newspaper Navigator see http-

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

s://github.com/LibraryOfCongress/newspaper-navigator, and for its output of prepackaged dataset with confidence scores higher than 90% see https://news-navigator.labs.loc.gov/ .

30 For a guide to Flickr API see Henriette Roued-Cunliffe, 'Developing Sustainable Open Heritage Datasets', in *Research Methods for the Digital Humanities*, ed. lewis levenberg, Tai Neilson, and David Rheams (Cham: Springer International Publishing, 2018), 287–307, https://doi.org/10.1007/978-3-319-96713-4_16.

31 All of the three projects presented in the panel "Machine Reading Victorians" during the Annual Conference of the Research Society for the Victorian Periodicals (University of Stirling, 13-15 June 2024) are planned to provide the users with data accompanied by a script in Jupyter Notebooks.

32 Andy Powell and Pete Johnston, 'Guidelines for Implementing Dublin CoreTM in XML', DCMI, 2 April 2003, https://www.dublincore.org/specifications/dublin-core/dc-xml-guidelines/.

33 https://vocabs.dariah.eu/iconclass/en/; https://rkd.nl/nl/collecties/services-tools/iconclass; https://iconclass.org/. Other possible classifications systems to consider include the *Library of Congress Subject Headings* (https://id.loc.gov/authorities/subjects.html) and the *Art and Architecture Thesaurus Online* (https://www.getty.edu/research/tools/vocabularies/aat) developed by the Getty Research Institute.

34 After the DMVI website was launched, its classifications were "mapped onto Iconclass codes," but as Thomas explains, due to "the differences between the Catholic fine art privileged by Iconclass and the largely Protestant illustrations in DMVI," the compatibility was not complete (Thomas, 2017, 57).

35 The project used two collections of illustrations cut by Victorian collectors from the 1860s and 1870s held in the School of Art Museum and Gallery, Aberystwyth University and the Forrest Reid collection in the Ashmolean Museum, Oxford. Julia Thomas, 'Reflections on Illustrations: The Database of Mid-Victorian Wood-Engraved Illustration (DMVI)', *Journal of Illustration Studies*, December 2007, https://web.archive.org/web/20150325063852/http://jois.uia.no/articles.php?article=37.

36 Brian Maidment, '*The Database of Mid-Victorian Wood-Engraved Illustration* (Review)', *Journal of Victorian Culture* 13, no. 1 (2008): 113–14, https://doi.org/10.1353/jvc.0.0015; 'History of the Database', Database of Mid-

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Victorian Illustration, accessed 23 September 2023, https://www.dmvi.org.uk/guide.php?page=13.

37 Thomas, 'Reflections on Illustrations'.

38 For an overview of crowdsourcing platforms with an emphasis on textual material see Katherine A. Mika, Joseph De Veer, and Constance Rinaldo, 'Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data', *Biodiversity Informatics*; Vol 12 (2017)DO - 10.17161/Bi.V12i0.6646, 14 November 2017, https://journals.ku.edu/jbi/article/view/6646/6090.

39 "folksonomy". Oxford Dictionaries. Oxford University Press. https://premium.oxforddictionaries.com/definition/english/folksonomy (accessed via Oxford Dictionaries Online on October 12, 2023).

40 https://about.biodiversitylibrary.org/projects/past-projects/art-of-life/

41 It is indicated that "Machine-tagged images are auto-ingested into the Encyclopaedia of Life & added to species pages". https://blog.biodiversitylibrary.org/2011/08/bhl-on-flickr.html. For a definition and the use of machine tags on Flickr see https://www.flickr.com/groups/api/discuss/72157594497877875/.

42 "Machine Tagging Tutorial," Biodiversity Heritage Library. http://s.si.edu/BHLTaggingGuide;

43 https://www.flickr.com/photos/biodivlibrary/collections/72157713307356438/

44 https://www.flickr.com/photos/biodivlibrary/collections/72157681766674633/

45 https://app.dimensions.ai/details/grant/grant.3496117; https://gtr.ukri.org/projects?ref=AH%2FL007010%2F1.

46 https://portal.sds.ox.ac.uk/search?groups=46353&contentTypes=project; https://ora.ox.ac.uk/objects/uuid:57958a9e-b690-49cd-86f3-186193e23604.

47 https://wordnet.princeton.edu/

48 https://eol.org/

49 https://help.zooniverse.org/getting-started/

50 https://blog.gdeltproject.org/500-years-of-the-images-of-the-worlds-books-now-on-flickr/

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

51 https://ia904508.us.archive.org/21/items/vrr-texts-imageryofbooks-fullres-imageextractortool/USERGUIDE-IMAGEEXTRACTOR.pdf

52 https://blogs.bl.uk/digital-scholarship/2016/08/sherlocknet-tagging-and-captioning-the-british-librarys-flickr-images.html

53 https://blogs.bl.uk/digital-scholarship/2016/11/sherlocknet-update-millions-of-tags-and-thousands-of-captions-added-to-the-bl-flickr-images.html

54 https://www.flickr.com/search/?user_id=12403504%40N02&view_all=1&text=%27sherlocknet%3Atag%3D%27&content_types=0%2C2&video_content_types=0%2C1%2C2%2C3

55 https://www.flickr.com/photos/britishlibrary/11306464394/in/photolist-ie7zXh-i4Vyjo-hLQszv-hSyCH2-i5i4G9-i4cozC-i4DPAt-hYYhPb-i6oETe-hPxQxq-hPubVA-hLHf4w-i85KQq-i9pkmi-i4RBF3-ie63pi-hR2Lcr-i55r2Q-hLJVEU-2jzP2di-hNb76Q-i6zqjn-hNbXcx-hSw51H-2jzAZ9Z-icc1tY-ie8Fes-i7cjwE-2jzNdkR-hUM2Gd-hLK5fi-i4EYSd-2jzHUJx-i6o5HT-icWZWX-i4YvVZ-2jzHUFw-i6o56u-ibGdFw-hPDLEH-i4H8EM-i79sVR-hPpu71-i79DtJ-idUvbj-i4HPqm-hLFWB1-i49hif-ieapJM-2jzJHAD

56 https://github.com/LibraryOfCongress/newspaper-navigator;

https://news-navigator.labs.loc.gov/search; https://chroniclingamerica.loc.gov/about/.

57 For a discussion on the captions in literary illustrations see Thomas, *Nineteenth-Century Illustration and the Digital*, 43–47.

58 Lee et al., 'The Newspaper Navigator Dataset', 3057–58; Benjamin Lee, 'Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset', 1 September 2020, 24–28, https://hcommons.org/deposits/item/hc:32415/; Impresso project also provides this feature: https://impresso-project.ch/.

---

**English**

In the past decade, an unprecedented number of illustrations were extracted from the scanned pages of books, periodicals, and newspapers and made available online. Proving the importance of visual culture in the past, these millions of images also posed the challenge of making them searchable and useful. From labour-intensive crowdsourcing to technology-intensive machine learning, several methods have been developed to this end with varying degrees of success. The problems faced and solutions

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

offered in these projects provide insights into the making of visual digital archives. This article aims at analysing the landscape of digital visual archives by focusing on a selection of projects since the beginning of the 2000s. It is a critical review as well as a history of the digital visual archives.

**Français**
Au cours des dix dernières années, un nombre sans précédent d'illustrations a été trouvé au fil des pages numérisées d'ouvrages, de périodiques et de journaux, et rendu disponible en ligne. Témoins de l'importance que tenait la culture visuelle jadis, ces millions d'images posaient également la question de leur mise à disposition en ligne pour être cherchées et utilisées. Plusieurs techniques (allant de la très laborieuse production participative, jusqu'à l'apprentissage automatique, onéreux en technologie) ont vu le jour avec plus ou moins de succès pour tenter d'atteindre ce but. Les problèmes rencontrés et les solutions trouvées au cours de ces projets sont autant de traces de la création d'archives visuelles numériques. Le present article souhaite proposer une cartographie de ces archives, en se concentrant sur une série de projets entrepris depuis le début des années 2000, permettant ainsi un panorama critique et une histoire de ces archives visuelles numériques.

---

**Mots-clés**
archives visuelles numériques, humanités numériques, interfaces, métadonnées, illustrations historiques

**Keywords**
digital visual archives, digital humanities, interfaces, metadata, historical illustrations

---

**Ali Hatapçı**
Maître de conférences, Centre Interlangues TIL (UR 4182), Université de Bourgogne, 4 Boulevard Gabriel, 21000 Dijon

ali.hatapci@u-bourgogne.fr
ali.hatapci@u-bourgogne.fr

Le texte seul, hors citations, est utilisable sous Licence CC BY 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.