Lost Visions: A Descriptive Metadata Crowdsourcing and Search Platform for Nineteenth-Century Book Illustrations

by Ian Harvey and Nicky Lloyd

Citation

Havery, Ian and Nicky Lloyd. 'Lost Visions: A Descriptive Metadata Crowdsourcing and Search Platform for Nineteenth-Century Book Illustrations'. In: Clare Mills, Michael Pidd and Jessica Williams. *Proceedings of the Digital Humanities Congress 2014*. Studies in the Digital Humanities. Sheffield: HRI Online Publications, 2014.Available online at: <https://www.dhi.ac.uk/openbook/chapter/dhc2014-harvey>

Abstract

Despite the mass digitization of books, illustrations have remained more or less invisible. As an aesthetic form, illustration is conventionally positioned at the bottom of a hierarchy that places painting and sculpture at the top. The hybridity or bi-mediality of illustration is also problematic, the genre having fallen between the cracks of literary studies and art history. In a digital context, illustration has fared no better: new technologies can aid the editing of a literary text far more successfully than they can deal with the images that accompany it.

This article considers the challenges and research implications of the AHRC-funded ‘Lost Visions’ project. The project focused on the development of a fully-searchable online database of over a million book illustrations from the British Library’s collections. The images span the late eighteenth to the early twentieth century, cover a variety of reproductive techniques (including etching, wood engraving, lithography and photography) and are taken from around 68,000 works of literature, history, geography and philosophy.

A ‘big data’ project of this nature inevitably raises a variety of challenges: this paper focuses on the methods undertaken to supplement the initial bibliographic metadata for the collection and to analyse and tag iconographic features of illustrations in order to aid searchability. In doing

so, it reveals a number of new research questions about the digital image archive and about how we might read images in the digital age, allowing for further development in the fields of both illustration studies and digital humanities.

Lost Visions: A Descriptive Metadata Crowdsourcing and Search Platform for Nineteenth-Century Book Illustrations

by Ian Harvey and Nicky Lloyd

1. Introduction

Developing a digital image archive is a process fraught with complexity. Certainly, as Patricia Harpring of the Getty Research Institute points out, “[i]mages are notoriously difficult to retrieve with accuracy” and “[r]etrieval of appropriate images depends on intelligent indexing” (Harpring 20). If the ultimate goal of the digital image archive is user retrieval, the archive must be attuned to the needs of the user and must anticipate how they might search: the format chosen will determine the questions that users can subsequently ask of the archive. In creating an archive designed for scholarly research, multiple methods of interrogating the dataset are essential: as Richard Pearson observes, “the more ways in which we can search an archive the better, and approaches that will open up ‘browsing’ and ‘searching’ – putting the search firmly into research – are going to be of the greatest assistance to future scholars” (Pearson 92-3).

In the digital age the relationship between word and image has become newly significant. On platforms that present literary texts in the form of facsimiles of their original editions or that use optical character recognition (OCR) technologies – such as Gale Cengage’s Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO) – the printed page occupies a newly hybrid form, at once both text and image. Conversely, in digital image repositories textual information is necessary in order to organise visual content. As James Mussell puts it, “the way in which images are described with metadata also determines how they behave and so this linguistic material plays both an interpretive and a structural role” (Mussell 72-3).

This reliance of the user on textual description to search for an image in a digital archive evokes a series of theoretical implications about the complex interaction between word and image in the digital age. These textual descriptions can be divided into two categories: bibliographic and iconographic. Some of this metadata details the image’s bibliographic

information: the title, the artist or author and so forth. However, while there may be correlation between bibliographic descriptions and the content of images – for example, in cases where the title of work might also describe its content – this is certainly not always the case, particularly when it comes to illustrations, which do not always have specific titles. Annotating images with keywords in order to create iconographic metadata that describe a picture's content is in itself an interpretative practice. There can be no objective method of labelling an image: the linguistic terms and what they describe are highly subjective and there are, as Julia Thomas observes, "aspects of a visual image […] that cannot be easily fitted into a linguistic structure: its surface, marks, lines, its very status as a visual object" (Thomas 2007 199). While Christine L. Sundt asserts that "controlled vocabularies and thesauri can be enormously powerful tools for bridging these kinds of verbal gaps" in searches for art images, book illustration brings with it an entirely different set of challenges (Sundt 70).

As an aesthetic form, the hybridity or bimediality of illustration is problematic. Conventionally positioned at the bottom of a hierarchy that places painting and sculpture at the top, illustration occupies an unstable position within the field of art history. Neither has it been privileged within literary studies; indeed, despite the mass digitization of books, illustrations have remained more or less invisible. Illustrations are – in their original form – images that signify through their relationship with text. This signification becomes even more pronounced in the digital image archive. Just as the meaning of illustration and text in the printed book are mutually dependent, so too must illustration be accessed and 'read' via linguistic markers in the digital archive. The semiotic challenge posed in generating these linguistic markers is problematised further by the lack of an established terminology to describe even the most basic aspects of an illustration.

The AHRC-funded 'Lost Visions' project focused on the development of a fully-searchable online database of over a million book illustrations from the British Library's collections. Scanned by Microsoft, the images span the late eighteenth to the early twentieth century, cover a variety of reproductive techniques (including etching, wood engraving, lithography and photography) and are taken from around 68,000 works of literature, history, geography and philosophy. As described above, within the already challenging context of online image retrieval, a dataset of over a million illustrations poses still more problems, exacerbated by the fact that the images were accompanied by bibliographic metadata relating to the books from which they derived but lacked any descriptive or iconographic metadata.

2. Objectives

There have been numerous attempts to create online repositories of images alongside associated data and metadata. Flickr was chosen by the British Library as a location for the upload of over a million images from books in their collection in 2013. However, as a platform for enabling scholarly research Flickr has a number of limitations. The user is presented with minimal information relating to the provenance of an uploaded image, and the bibliographic metadata stored during the digitisation process in the case of the British Library images was either lost or stored as an image tag. Flickr does not provide any information regarding the book in which an illustration was originally published, as every image is essentially either stand-alone or contained within user-curated collections and neither of these options has a metadata structure suitable for storing or searching for connected illustrations via the author, illustrator or publisher of a book. While the uploading, sharing, tagging and curation of images is successfully performed on a very large scale on Flickr, the 'Lost Visions' project sought to develop mechanisms for more comprehensive organisation and connectivity of the image metadata using a combination of computational methods and crowdsourced tagging.

In practical terms, the objective of the 'Lost Visions' project was to create a fully-searchable archive and to improve the existing metadata in order to aid image retrieval. The project provided an opportunity to investigate the partnership between traditional computer science conceptssuch as production of algorithms and data management, the expectations of users in humanities research disciplines and feedback from members of the public. As well as producing a framework for describing illustrations using metadata from multiple sources, it was a high priority to produce a stable and functional service which would be open to use by both researchers and the general public at the end of the project.

Given that manual markup and keywording of individual illustrations is impractical in such a large dataset, we also sought to develop a system for crowdsourced iconographic tagging of illustrations. Computationally, efforts were made to analyse collected metadata to make predictions about a illustration's content based on assigned tagsand so create features and functions within the Illustration Archive which allow users to navigate a large set of images in new and interesting ways. Specifically, comparing the semantic content of tags and their groupings between illustrations formed the basis of a 'More Like This' function, which shows similar images when viewing a previously tagged illustration.

In a wider context, the 'Lost Visions' project operated as a critical practice in its own right, enabling illustrations to be viewed in new and different ways. The development of crowdsourced tagging brings to light the various theoretical implications of the relationship between the textual and the visual in the digitisation of images. While this was necessarily grounded in linguistic descriptions of images, we also wanted to provide users with a variety of ways to use the database that would go some way towards challenging the persistent subordination of the visual to the textual in the digital archive. The gallery view and the browse function of the Illustration Archive allow users to view multiple images simultaneously with no textual content displayed, offering the opportunity to discover hitherto undiscovered connections and correlations between illustrations from disparate periods and genres.

3. Improving Bibliographic Metadata

On release of the one million images, the British Library also made a full set of bibliographic metadata available in Tab Separated Variable (.tsv) format on their public GitHub account. This data contained, amongst other elements, details of the author, title and publisher for the book each illustration was retrieved from, as well as the page number and an index in the case where multiple illustrations appear on a single page. This data became the initial data object which future data collection techniques would refer to and build upon. However, as is frequently the case, the bibliographic metadata accompanying the illustrations is uncertain, incomplete and often ambiguous.

Due to the way in which the metadata was collected and recorded over the years, many of the book titles in the British Library catalogue have been abridged. For example, compare the title in the British Library metadata in the 'Lost Visions' dataset for the 1806 work *Indian Antiquities* with the title of the same edition from a facsimile edition on Google Books:

From the BL metadata:

> *Indian Antiquities: or, Dissertations relative to the ancient geographical divisions, the ... primeval theology, the grand code of civil laws, the original form of government, and the ... literature of Hindostan, compared ... with the religion, laws, government and literature of Persia, Egypt and Greece. The whole intended as*

> *introductory to ... the History of Hindostan.*

From Google Books:

Figure 1

> *Indian antiquities: Or, dissertations, relative to the ancient geographical divisions, the pure system of primeval theology, the grand code of civil laws, the original form of government, the widely-extended commerce, and the various and profound literature, of Hindostan. Compared, throughout, with the religion, laws, government, and literature, of Persia, Egypt, and Greece. The whole intended as introductory to the history of Hindostan, upon a comprehensive scale.*

In the example from the BL dataset, words from the title have been removed and some additional punctuation added. This could be for any number of reasons, such as different sources of catalogue records, the processes used to digitise handwritten records, or to technological limitations such as maximum length text entry fields in a database. In addition to incomplete or abridged titles, the publication dates associated with individual illustrations refer to the publication of the specific volume from which an illustration derives, which can differ considerably from the date of creation or the original printing of the illustration.

While it was not an objective of the 'Lost Visions' project to provide the data cleansing required to perfect these bibliographic records, the challenges of using inherited metadata were acknowledged and solutions considered. Consultation workshops that ran throughout the project frequently raised questions about the possibility of data cleanup as part of the crowdsourcing process. The issue of updating data is commonplace, as is the concept of retaining information relating to changes for future reference. The problem faced here is that of the crowdsourcing of expert knowledge, and the uncertainty of accuracy in the new information as well as that it would potentially replace. Such a process would require a large amount of moderation in order to validate and verify any changes in records.

The main challenge presented by the bibliographic metadata provided with the images was that there was no specific field for the name of the illustrator or engraver of an illustration, in a gesture which is consistent with the subordinate status of illustration within the hierarchies of literary studies. In order to create a fully-searchable archive of illustrations, it was essential that the archive could be searchable using these terms and that new descriptive metadata was produced. With the above limitations in mind, efforts were made to create an algorithm able to extract information from the title 'strings' stored in the database. The aim was to find sub-strings such as 'with drawings by the author' and so create new searchable metadata relating specifically to the illustrator or creator of images.

This produced mixed results due to the lack of any standardisation across titles, the wide variety of languages used throughout this set of books, and the repetition of multiple names in some titles. In the initial run of the algorithm of approximately 30,000 images around 200 'image creator' sub-strings were discovered, of which around 30-40% were false positives. Work is continuing to increase the quality of these results. As a result of the difficulties of extracting the names of the creators of illustrations, the current solution to providing a search feature is essentially a standard text search within the title field. This is useful only when the searched name is already known, but has already been promising in terms of new illustrations discovered.

4. Collection of Crowdsourced Descriptive Metadata

There has been considerable focus on the phenomenon of crowdsourcing within the field of digital humanities in recent years. Operating as both a means of improving online archives and as a valuable form of public engagement with cultural heritage, there have been numerous successful crowdsourcing endeavours over the past decade. Probably the most well-known academic crowdsourcing organisation is Zooniverse, which began with a 'citizen science' project in astrophysics in 2007 and has gone on to develop more than 30 projects across the sciences and the humanities, attempting to provide accessible user interfaces for crowdsourcing very specific information for specialised datasets. Zooniverse humanities projects include *Ancient Lives* (http://www.ancientlives.org/) – which involves the transcription of Egyptian papyri – and the 2014 project *Operation War Diary* (http://www.operationwardiary.org/), which combines transcription and tagging of archival material from World War I. These projects have been

hugely successful: *Ancient Lives* has benefitted from the work of over 250,000 unique online participants and to date *Operation War Diary* has seen over 10,000 unique volunteers add nearly 500,000 tags and transcriptions. Another successful venture is the AHRC-funded *Transcribe Bentham* project at UCL (http://blogs.ucl.ac.uk/transcribe-bentham/), launched in 2010. It has – according to their latest weekly update at the time of writing – transcribed 13,696 manuscripts of Bentham's handwritten documents through integration of user communities such as school children and amateur historians.

In terms of crowdsourcing for online image archives, one of the most successful models for tagging of digital images is the *BBC Your Paintings Tagger* (http://tagger.thepcf.org.uk/), which asks users to enter a combination of generic and iconographic descriptors. While this offers a useful model for the Illustration Archive in many respects, it is specifically related to art images; for example, the user is prompted to refer to the title of paintings to inform the tags for three of these five sections, whereas the illustrations in the Illustration Archive dataset did not usually have specific titles. This highlights the difficulties faced in designing a method for iconographic tagging on illustrations: there is no defined vocabulary with which to describe it.

Iconclass (http://www.iconclass.nl/home) is the best-known controlled vocabulary for describing the content of images, with a system of 28,000 hierarchically-ordered definitions divided into ten main divisions, each containing an alphanumeric classification code (notation) and the description of the iconographic subject (textual correlate). However, as James Mussell and Julia Thomas point out, that fact that it was originally designed for medieval and Renaissance works of art and that its categories and hierarchies tend to be biased in favour of the religious subject matter of these images makes it unsuitable for the iconographic tagging of nineteenth-century illustrations (Mussell 108, Thomas 2007 203). Indeed, there are also issues of range and scale to take into account: the 'Lost Visions' dataset is so disparate (containing not only the illustrations of works of fiction but diagrams, handwriting, title pages and musical scores to name but a few broad types) that the use of tightly-controlled tagging hierarchies was not viable. Asking users to navigate a hierarchy of 28,000 categories in order to accurately describe an illustration would be impractical, while a smaller subset of categories would lose the desired accuracy required to properly define an illustration's attributes.

The Database of Mid Victorian Illustration (DMVI) (http://www.dmvi.org.uk/)

– an earlier AHRC-funded project at Cardiff University – catalogued 868 wood-engraved illustrations using accurate bibliographic classifications and a customised iconographic taxonomy, which was subsequently adapted for the 5000 images in the *Ninteenth-Century Serials Edition* (NCSE) (http://www.ncse.ac.uk/index.html). However, both archives were able to manually assign iconographic keywords to their relatively small datasets, which was not practical for the 'Lost Visions' project. Our considerations for designing a process for crowdsourced iconographic tagging, then, needed to take into account a number of factors. While the detailed hierarchies and categories offered by the *DMVI* were hugely appealing with the end goal of user image retrieval in mind, the level of complexity and specialist knowledge they represent was not consistent with large scale crowdsourced tagging.

Reducing task complexity was important – any potential uncertainty about how to fill in text box, for example, disengages the tagger and necessitates awkward and complex pre-task tutorials, or guides, which we wanted to avoid. We also wanted to avoid asking for specialist or technical knowledge (for example distinguishing between etchings, engravings, lithographs etc.) which we hope we will eventually be able to determine using computational methods. The issue we faced was that, if the retrieval of images takes place at the level of language, how, within the context of crowdsourcing, can we control the labelling of images? If the creator of the archive would usually be responsible for annotating images with words so that the user can find them by searching these same words, what are the implications of public crowdsourcing for big data, where the archive creator can have minimal control over the interpretative process behind assigned tags?

As discussed previously, a lack of established linguistic hierarchies for illustrations necessitated the production of a set of features with which illustrations could be categorised. While it was considered important for the user to have the tools and the freedom to add whichever information about an illustration they deemed pertinent, there was also an interest in guiding the user to place the illustration into one or more generic or topical categories. In order to facilitate a iconographic search, it was necessary to consider both what illustration is 'of' and what it is 'about'. However, while this could be addressed through user-generated keywording, neither of these aspects of an illustration was entirely adequate for helping to classify the wide range of images in the dataset. As a result, we decided to implement a closed system of tagging the 'type' of images that would encourage users to select a broad generic category for each illustration as the first stage in the tagging process (as depicted below).

These categories needed to fit a series of requirements: they should be readily understood by a non-expert tagger, they should be easily demonstrated in a visual example on a button on the website and the information obtained regarding the category should be relevant in terms of searchability and image retrieval. This type of 'scaffolded' crowdsourcing system not only makes the process simpler and quicker for the tagger: it facilitates the convergence of professional and amateur knowledge by allowing users to participate in the development of scholarly research without themselves requiring the skills and background of a researcher.

Figure 2

The selection of a relevant category from a fixed list as the first step in the tagging workflow performed six main functions:

1. To guide the user in thinking about the type of illustration in front of them.
2. To allow the user a simple 'one click' interface to begin adding information to the system.
3. To provide a very simple set of illustrated buttons as the first step to a more complicated data entry process.
4. To provide a closed vocabulary from which to build further detail.
5. To provide a controlled entry point to more complex data entry tools, i.e.
    a. The option to place the location illustrated on a map.
    b. The option to select a named area from a gazetteer.
6. In the event that the required category is not available as a default option, the user is already prompted to think about what to manually add instead. For this reason, a 'None of These' option was added at this step, to more easily guide the user forward through the tagging processes.

For each of the pre-selected categories, a single tag is associated with the illustration. When a user presses the category button a linguistic tag is entered for the illustration in the same way it would be if the tag had been entered manually, except a synset from WordNet has been pre-selected.

The next stage in the tagging sequence is a free text section in which users are asked whether there are 'any other things or ideas' in the illustration. The examples given in the tagging command – 'bird', 'table', 'winter' and 'love' – are designed to cover a range of linguistic terms from objects to

more abstracts concepts to include both what the illustration is ‘of’ and what it is ‘about’. In the next two stages of the tagging sequence, users are then asked to enter the caption (a title below the image) for the illustration, if present, and given the opportunity to add any additional information in a free text box.

As mentioned above, an issue arises where two people use different words, and therefore different tags, to describe a single object or concept while tagging illustrations. For example, ‘church’, ‘religious building’ or ‘place of worship’ could all be used to describe the same illustrative depiction. This becomes a problem when a user attempts to search the collection, as unless an exact match for a term is found, the tagged illustrations would not be returned in the results. To combat this issue, the tagging process was extended to automatically generate new tags based on a single tag added by the user. WordNet, a lexical database of connected ‘synsets’ made available by Princeton University, allows discovery of synonyms and a hypernyms in a programmatically accessible way. With this database incorporated into the tagging process, the word ‘church’ can automatically trigger the addition of a multitude of other relevant tags as if the user had used a thesaurus themselves. Additionally, the option to add hypernyms as tags means that ‘church’ can also automatically be tagged as a ‘building’, ‘structure’, ‘construction’ and upwards to more general tags such as ‘object’ and ‘physical entity’. While the recording of such tags could be considered extraneous or redundant, in reality they proved highly useful in the production of a ‘More Like This’ browsing feature on the archive which allows users to discover related images without needing to perform a text search.

The fact that WordNet connects words based on their meaning, rather than mere spelling, is of high importance, and allowed extremely accurate production of automated tags. With the ‘church’ example, the user originally enters the term by typing the word into a text box. This process strips the word from the intended meaning, and so raises ambiguity in the tag production algorithms - for example ‘church’ the building as opposed to ‘church’ the institution, or collection of people. To tackle this problem, the input process was extended to provide a selection of homographs and a description of their usage to the user. Once the user selects the intended definition of the word, further tags could be generated based on the sense of the word, and so increase accuracy of future tagging and searching features. As an example of the importance of this feature, consider the word ‘field’, which returns 16 nouns and 4 verbs with this spelling. This problem was discovered early on in the implementation of this function as fields of various

types are common within the dataset, and automated tagging without the word sense was almost guaranteed to produce inappropriate tags for an illustration.

5. Search Functionality

Consideration was made into the different ways in which users perform searches with such a tool. In terms of audience, we aimed to create a scholarly research tool that would also open up the content to users who did not have specialist knowledge that would enable them search by bibliographic details. In addition, feedback from workshops guided the production of search tools away from simply performing text searches in bibliographic metadata, towards more vague queries or entirely random selections of illustrations from which to discover new content with a high degree of serendipity.

To this end, the option to have search results tiled on the screen was added, essentially hiding all metadata and allowing a visual focus which foregrounds the illustration itself. Aesthetically, this allows more results to be visible on a single screen, but it also creates a separation of illustration from its original volume. The ability to remove individual images from this tiled view of search results meant that a user can customise a view of results which is tailored to their interests. Of particular note when receiving feedback from alpha users and during workshops was the comparison between the search functions we were providing, compared to search engines such as Google. There is an expectation from users to be able to perform, for example, searches which use quotes to connect terms (‘blue boat’ being a different search to ‘blue’ + ‘boat’), or using mathematical symbols, as in the same example. During these developments, questions about how results should be displayed frequently arose, relating specifically to the number of results which should be generated and returned for a specific search, the order in which they should be returned and how a search could be amended or refined after the original search was performed.

The number of search results discovered and returned to the user presents a complex challenge. It was discovered that while some users will refine their search if the illustrations they are seeking do not appear in the first few results, other users will happily scroll through many hundreds or thousands of results in search of items of interest. While the popular internet meme suggests that ‘the best place to hide a dead body is page two of Google’, and

a study showed that 91% of users do not go beyond the first page in a Google search, this statement caused consternation during the round-table feedback stage of multiple workshops. Indeed, it was the opinion of most humanities researchers present that they would be willing to peruse many pages of search results in the expectation of discovering things only tangentially related to the original search terms. This was highly at odds with the usage scenarios envisaged by the developer working on the search functionality of the Illustration Archive, and resulted in the addition of many new features to enable a less structured and more serendipitous search process, including a 'random search' function, which simply returns a few hundred random illustrations from the set without the need for a text search. From this point onwards, it was important to review expectations of all elements of the design of the archive in order to account for any preconceived opinions about how researchers from different backgrounds perform searches.

6. Conclusion: Findings and New Research Questions

The primary output of the 'Lost Visions' project is the Illustration Archive (http://illustrationarchive.cf.ac.uk/), a web accessible front-end to a database of bibliographic metadata and descriptive tags, which serves the functions of both crowdsourcing descriptions of the one million illustrations from the British Library dataset, and provides a variety of search functions for these illustrations. The website is set up to be hosted at Cardiff University for 10 years. The codebase is open source, and hosted on GitHub (https://github.com/CSCSI/Lost-Visions). Future projects will have the option of building from this code and/or the underlying database(s) to add further information about the existing illustrations, or to add new illustrations to the dataset.

Making a large number of illustrations available in this way allows researchers to analyse these images in new and varied ways. Bibliographic searches can reveal, for example, which authors and genres were most frequently illustrated in a given time frame, offering insights into the literary marketplace. Iconographic searches can reveal similarities between illustrations of texts from widely different periods and genres, suggesting, as Julia Thomas points out, "the ways in which these images signify by the repetition of stylistic features, characteristics and devices and, as such, form part of an aesthetic and artistic tradition that is distinct from its textual

analogues" (Thomas 2010 104). The practical and computational challenges associated with crowdsourcing descriptive metadata in the 'Lost Visions' project also find numerous parallels in the broader field of illustration studies.

The complexity of the how the visual might be accounted for and described in linguistic terms in the tagging process reflects the way that the hybridity or bimediality of illustration is also problematic, the genre having fallen between the cracks of literary studies and art history. The range of the collection offers an unprecedented opportunity for illustration studies research, in both a qualitative and quantitative sense, or – to use Franco Moretti's terms – as it relates to both 'close' and 'distant' reading. The crowdsourced data collected has the potential to tell us about how images are interpreted in the digital age and to understand how communities engage with online archives. Furthermore, the processes involved in the creation of the digital image archive raise a series of critical questions about the interaction between word and image, drawing attention to the process of remediation that takes place in the digital archive and to the theoretical implications of incorporating images into a linguistic structure. The Illustration Archive retrieves the 'lost vision' of historic illustration, restoring the visual element of the printed book and challenging both the hierarchies of textual production in which the author or creator of a work are privileged and the persistent subordination of the visual to the textual in the digital archive.

References

Harpring, Patricia. (2002) The Language of Images: Enhancing Access to Images by Applying Metadata Schemas and Structured Vocabularies. in Murtha Baca (ed.) Introduction to Art Image Access: Issues, Tools, Standards, Strategies. Los Angeles: Getty, pp. 20-39.

Moretti, Franco. (2013) Distant Reading. London: Verso.

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

Pearson, Richard. (2008) 'Etexts and Archives'. Journal of Victorian Culture 13.1:88-93.

Princeton University (2010) About WordNet. WordNet: Princeton University

<http://wordnet.princeton.edu>. [Online]

Sundt, Christine L. (2002) The Image User and the Search for Images. in Murtha Baca (ed.) Introduction to Art Image Access: Issues, Tools, Standards, Strategies. Los Angeles: Getty. pp. 67-85.

Thomas, Julia. (2008) Digital Transformations. Journal of Victorian Culture 13.1:101-107.

Thomas, Julia. (2007) Getting the Picture: Word and Image in the Digital Archive. European Journal of English Studies 11:2:193-206

Trant, J. with the Participants in the steve.museum Project. (2006) Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept. New Review of Hypermedia and Multimedia 12:1:83-105.

Van Deursen, A.J.A.M. & Van Dijk, J.A.G.M. (2009) Using the Internet: Skill Related Problems in Users' Online Behavior. Interacting with Computers 21:393-402