Investigating User Experience and Bias Mitigation of the Multi-Modal Retrieval of Historical Data

by

Soham Hanuman Singh[1]

Dissertation presented for the degree of
MASTER OF SCIENCE
in the
Department of Computer Science,
University of Cape Town.

March 2021

Supervisor: Hussein Suleman[2]

[1] sngsoh004@myuct.ac.za
[2] hussein@.cs.uct.ac.za

Plagiarism Declaration

I know the meaning of plagiarism and declare that all the work in this thesis, save for that which is properly acknowledged, is my own.

- Soham Hanuman Singh, 3rd of March 2021.

Acknowledgements

I would like to thank my fiancé, Sonali, for her constant patience, love, and support throughout this journey. You were always there to keep me going and I cannot thank you enough. To my parents, Sharona and Suveer, you have taught me the value of education and have motivated me throughout my life to be the best I can be. Thank you for being the best parents I could have ever asked for. I would also like to thank my employer, Derivco, for their compassion, understanding and financial assistance while I spent time working on my postgraduate studies.

To the Archive & Public Culture Research Initiative at UCT, thank you for allowing me to use the data from the Five Hundred Year Archive in this research, as well as for your ongoing feedback and support. This was invaluable to this research. Lastly, I would like to thank my supervisor, Hussein Suleman. Your assistance, wisdom and patience was immeasurable, and I am a better student for it.

Abstract

Decolonisation has raised the discussion of technology having the responsibility of presenting multiple perspectives to users. This is specifically relevant to African precolonial heritage artefact data, where the data contains the bias of the curators of the artefacts and there are primary concerns surrounding the social responsibility of these systems. Historians have argued that common information retrieval algorithms may further bias results presented to users. While research for mitigating bias in information retrieval is steered in the direction of artificial intelligence and automation, an often-neglected approach is that of user-control. User-control has proven to be beneficial in other research areas and is strongly aligned with the core principles of decolonisation.

Thus, the effects on user experience, bias mitigation, and retrieval effectiveness from the addition of user-control and algorithmic variation to a multimodal information retrieval system containing precolonial African heritage data was investigated in this study. This was done by conducting two experiments: 1) an experiment to provide a baseline offline evaluation of various algorithms for text and image retrieval and 2) an experiment to investigate the user experience with a retrieval system that allowed them to compare algorithms. In the first experiment, the differences in retrieval effectiveness between colour-based pre-processing algorithms, shape-based pre-processing algorithms, and pre-processing algorithms based on a combination of colour- and shape-detection, was explored. The differences in retrieval effectiveness between stemming, stopword removal and synonym query expansion were also evaluated for text retrieval. In the second experiment, the manner in which users experience bias in the context of common information retrieval algorithms for both the textual and image data that are available in typical historical archives was explored. Users were presented with the results generated by multiple algorithmic variations, in a variety of different result formats, and using a variety of different search methods, affording them the opportunity to decide what they deem provides them with a more relevant set of results.

The results of the study show that algorithmic variation can lead to significantly improved retrieval performance with respect to image-based retrieval. The results also show that users potentially prefer shape-based image algorithms rather than colour-based image algorithms, and, that shape-based image algorithms can lead to significantly improved retrieval of historical data. The results also show that users have justifiable preferences for multimodal query and result formats to improve user experience and that users believe they can control bias using algorithmic variation.

Table of Contents

List of Figures

List of Tables

1. Introduction

In the digital age, there needs to be a greater emphasis on the preservation of cultural heritage. There is a number of important heritage collections containing physical artefacts that are subject to dilapidation and that are inaccessible to the general public (Williams, Manilal, Molwantoa, & Suleman, 2010). If these collections of physical heritage artefacts and oral data are digitised, institutions would be able to collaborate and reunite related materials and collections, improve the annotations of the data, and make it easily accessible to the general public (Benson, 2010; Shenton, 2009; ARL Working Group on Special Collections, 2009).

Archiving and preservation of heritage data is typically done using an *information retrieval system*. An information retrieval system is a system that is capable of the storage, retrieval, and maintenance of information (Kowalski, 1997). Popular examples of information retrieval systems are commercial search engines such as Google or Bing. While popular commercial search engines allow users to search for web pages or media, other search engines may allow users to browse knowledge collections, such as online museums, libraries, or wikis (such as Wikipedia). These systems typically contain documents with information about specific entities or topics, so that users can research and acquire knowledge about the given topic. This can be seen in Figure 1.1, which is an example of an online collection of Native American heritage artefacts that is explorable through a search engine.

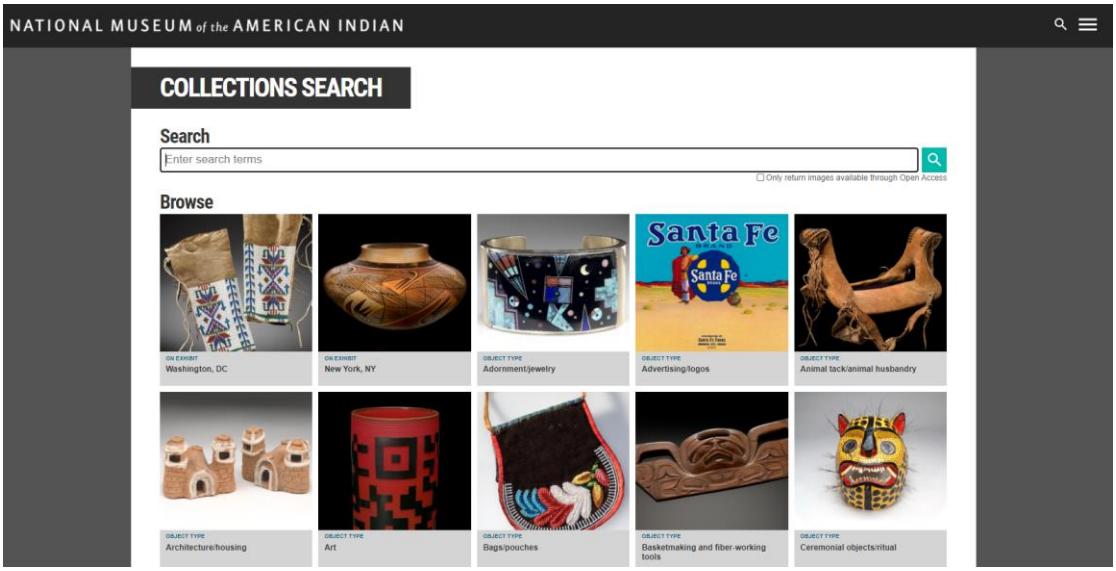

Figure 1.1: Search engine for a collection of Native American heritage data[3]

Most commonly, users would navigate these systems by submitting text phrases (known as a *query*) that the search engine would use to return to the user a corresponding list of matching text documents. Advancements in this field have led to what is known as *multimodal retrieval.* This refers to the retrieval of results, of more than one format, in response to a query of a single format (Long, Cao, Wang, & Yu, 2016). This allows for users to interact with and navigate these systems with far greater control, as they are able to use images to find other similar images, or

[3] https://americanindian.si.edu/

even texts, in the case where an image better describes the information they are looking for than a text query would. This can be seen in Figure 1.2, where an image of a flower was used as the query (A), instead of the name of the flower. This allowed for the search engine (Google in this instance) to return text documents related to the image of the flower (B), as well as images of similar or the same flowers (C). This example demonstrates the value of multimodal retrieval. If someone were to see this flower in reality, and want to learn more about it, it would be very difficult to find information about it if they did not know the name of it. However, with multimodal retrieval, they are able to instead take a picture of the flower and use that to find information about it.

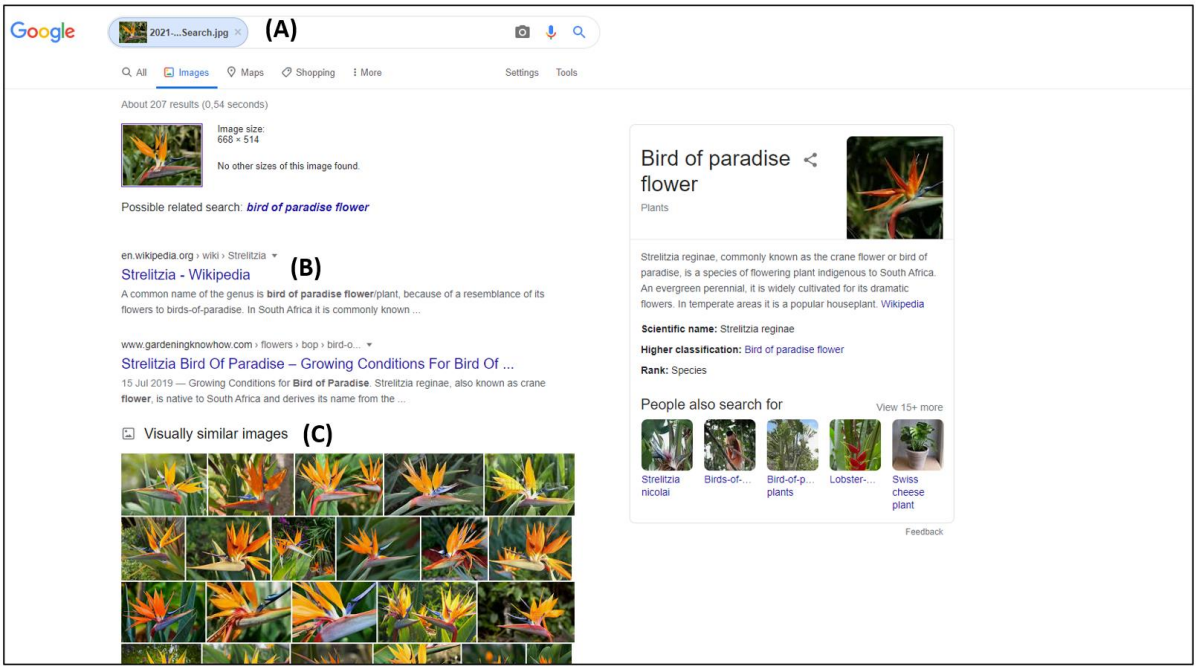

Figure 1.2: Multimodal retrieval using Google to search by image

While it is clear to see the value in preserving cultural heritage data in information retrieval systems, African heritage data preservation introduces a problem set that is different to the problems faced in the global North-West (Suleman, 2011). Heritage data preservation initiatives in Africa have the task of addressing the unique permutations of problems such as the deterioration of the artefacts and, importantly, rewriting the history of these cultures (Suleman, 2011). Given the history of colonisation in South Africa, there is a social responsibility in bringing to the fore the true representation of Africa's precolonial cultural heritage. This aligns with the discussions in academia and the mainstream media surrounding decolonisation.

Decolonisation is the process of unlearning unjust practices and ideals, and dismantling institutions with the intention of facilitating other perspectives of knowledge (Kessi, Marks, & Ramugondo, 2020). The decolonisation discussions motivate for digital preservation initiatives to consider the origins and creation of the data they allow users to explore, and how they might influence the interpretation and navigation of the cultural data. In many instances, precolonial heritage data would have been curated in the colonial era. This data would then have been annotated by the curators of the artefacts. The Archive and Public Culture research initiative at the University of Cape Town has expressed concerns regarding the findability of such heritage artefacts in retrieval systems. This

is due to the potential of the biases of the curators of the artefacts to misattribute the data and negatively impact the *retrievability* of that data. Retrievability is the potential for a document to be retrieved by a given query. This can be seen in Figure 1.3, where the curator's description is used as the title of the artefact. This description is largely responsible for the textual retrieval of this artefact, and any misattribution of it by the curator will prevent the retrieval of the data. In addition, the vocabulary used by the curator may not match that of the researchers, and specificities known to the original creators of the artefacts are lost in translation when an archivist thousands of kilometres away curates the object without full knowledge of its context.

Headrest
Tsonga artist
South Africa

Figure 1.3: Captioned African heritage data from an online archive collection[4]

While the metadata may contain the bias of the curators of the artefacts, these information retrieval systems may potentially also have biased underlying algorithms. These are algorithms that favour the matching of certain documents for reasons unrelated to their relevance to the query. Bias in retrieval systems has been the recent subject of many research initiatives and has led to the trend towards transparent, explainable, and accountable systems (Long, Cao, Wang, & Yu, 2016; Burrell, Kahn, & Jonas, 2019). It is understood that the biases present in these systems can impact users' interactions with search engines (White, 2013), and can also have serious consequences for the users, such as with search engines that recommend medical treatment courses (Caruana, et al., 2015; Pogacar, Ghenai, Smucker, & Clarke, 2017). The combination of algorithmic bias, as well as biased metadata, makes for the problematic retrieval of precolonial African heritage data.

Alternative strategies need to be considered to retrieve such data in the case where textual metadata is not accurate or complete, such as multimodal retrieval, the presentation of results in different formats that do not rely on the textual metadata alone, or the use of other algorithms that are able to find relevant documents by focusing on different features in the images or metadata. Thus, a potential solution to these problems of algorithmic and metadata bias is the granting of control over the retrieval process to the user. Historically, user-control in information retrieval systems yielded mixed results with respect to retrieval performance, yet was found to make the retrieval process easier by users and was requested to be implemented in more search engines by users (Nemeth, Shapira, & Taeib-Maimon, 2004; Beaulieu, 1997). It has since been found to be successful in several other research areas such as in recommender systems, adaptive systems, user interface design, personalization systems and educational systems (Orji, Oyibo, & Tondello, 2017; Tsandilas & Shcraefel, 2003; Rahdari & Brusilovsky, 2019). Given that user-control aligns strongly with the core concepts of decolonisation, such as interrogation, autonomy, and freedom (Orji, Oyibo, & Tondello, 2017), it was investigated in this study whether

[4] https://africa.si.edu/collections/collections

affording users greater control over the retrieval process can potentially lead to improved retrieval effectiveness, user experience and the mitigation of bias in heritage data retrieval systems.

1.1 Research Questions

This research aims to explore the effects on user experience and bias mitigation from the addition of user-control and algorithmic variation to a multimodal information retrieval system containing heritage data. The addition of user-control and algorithmic variation will be explored by allowing the user to conduct queries via text or image, select the retrieval pre-processing algorithm and alternate among various result views. These facilities will ultimately assist with discovering answers to the following research questions:

1) What is the impact of text vs. image querying on user experience and retrieval effectiveness of unbiased results?
2) What is the impact of variation of retrieval pre-processing algorithms on retrieval effectiveness of unbiased results?
3) What is the impact of result display format on user experience?

1.2 Methodology

This study was designed in order to answer the research questions listed in section 1.1 by following an approach consisting of two experiments: 1) an experiment to provide a baseline offline evaluation of various pre-processing algorithms for text and image retrieval (Experiment 1) and 2) an experiment to investigate the user experience with a retrieval system that allowed them to compare pre-processing algorithms (Experiment 2).

In Experiment 1, the retrieval performance across 6 different image retrieval pre-processing algorithms was calculated, with an emphasis on determining whether shape-based pre-processing algorithms, colour-based pre-processing algorithms, or pre-processing algorithms that are both colour- and shape-based, perform the best for the African heritage data used in this study. For text retrieval, the experiment focused on the difference in retrieval performance from stemming processing, stopword removal and synonym query expansion. A multimodal information retrieval system was built that contained indexed documents from the Five Hundred Year Archive[5], an archive that consists of precolonial African heritage text and image data. In the experiment, a collection of heritage image- and text-based data was used as sample queries that were submitted to the information retrieval system. The results of these queries across all pre-processing algorithms were recorded, and participants graded the relevance of these results. These graded relevance measures were then used to calculate the average retrieval effectiveness of each pre-processing algorithm across a number of metrics (i.e. recall, precision, F-measure, NDCG etc.). These results were then analysed to determine the similarities between the pre-processing algorithms that were recorded to have better retrieval performance.

An interface for the information retrieval system was developed and was used by participants during Experiment 2. This experiment aimed to explore the effects on user experience and bias mitigation from the addition of user-

[5] http://www.apc.uct.ac.za/apc/research/projects/five-hundred-year-archive

control and algorithmic variation to the multimodal information retrieval system. Specifically, this experiment investigated the effects of user-control of query format, pre-processing algorithms, and result format. In this experiment, users were asked to perform several tasks using a retrieval system in a controlled environment and then provide feedback, via a survey, regarding the impact on retrieval effectiveness and bias from the granting of control over the retrieval process and algorithmic variation. Their responses to the survey were then analysed and compared to prior research and the results from Experiment 1.

1.3 Thesis Overview

Chapter 2 provides the necessary context for understanding the principles of information retrieval and multimodal retrieval as well as the issues of bias in retrieval. In this chapter we also present a summary of prior research of using user-control to improve retrieval and mitigate bias in retrieval. The principles and background for decolonisation and heritage data preservation and retrieval is also discussed.

Chapter 3 and **Chapter 4** provide detailed descriptions of the research design, methodology and implementation followed for Experiments 1 and 2 respectively.

Chapter 5 presents answers to the research questions, as well as the conclusions reached from this study. This chapter also acknowledges the limitations of this study and provides recommendations for future research.

Appendix A lists the 70 text-based queries used for Experiment 1. **Appendix B** presents screenshots of the pre-processing algorithm explanation pages available to users in the retrieval system user interface in Experiment 2. **Appendix C** is the template instructions sent to participants for Experiment 2. **Appendix D** is the approved Informed Voluntary Consent to Participate in Research Study form that was issued to participants of this study. Lastly, **Appendix E** is the survey completed by participants during Experiment 2.

2. Literature Review

This chapter provides context for the issue of heritage data retrieval, understanding multimodal information retrieval and the biases present in retrieval, the historic effects and potential of user-control on user experience and bias mitigation, and the discourse surrounding the decolonisation of retrieval. This chapter shows that there is an inherent concern of user experience and bias mitigation when adopting decolonisation principles and this forms the foundation of the study.

2.1 Heritage data archiving and preservation

Efforts by the Open Content Alliance and Google's book search partnership have assisted with universal accessibility to the world's knowledge (Guill, 2009). While this has largely assisted with the availability of digitised print material, there is now a concern of digitising material that is unable to be similarly easily scanned and shared, such as special collections in libraries and archives (Gracy & Kahn, 2012). Whitelaw notes that the process of digitising these heritage collections and materials will assist greatly towards new discrimination of the heritage data, that was not previously possible, given the potential for enhancement, technology and tools that are now available to us (Whitelaw, 2009). Other benefits of digital preservation include being able to overcome the hurdles of obsolete formats and poor physical conditions (among others) to make previously inaccessible archived materials accessible (Benson, 2010), and facilitating collaboration between institutions to bring together related materials that are held in physically separate institutions (Shenton, 2009). This collaboration can potentially lead to increased contextualization and annotation of archived materials (ARL Working Group on Special Collections, 2009). This process of digitisation, however, faces many obstacles. Digitising heritage data is prone to recurrent issues surrounding file formats, antivirus checks, metadata generation and preservation-ready file conversion (among others) (Oliver & Harvey, 2016).

Sweetnam, et al. (2012) identify 4 different types of viewers for digital collections of historical data: professional researchers, apprentice investigators, informed users, and the general public. Sweetnam et al. also explain that despite the fact that these 4 groups all have separate interests in the historical data, there is an overlap in their requirements for these digital collections, such as the ability to conduct searches accurately, visual interactions with the content in the collections and more (Sweetnam, et al., 2012). Following UNESCO's motivation for the preservation of cultural heritage data for accessibility to that data, a number of projects have emerged in Europe and the USA to support historical data preservation. PREMIS (Preservation Metadata: Implementation Strategies), an initiative from the University of Illinois, set out to define metadata elements that support the preservation of data (Caplan & Guenther, 2005). This was followed by work such as CASPAR (Giaretta, 2006), which adopts the Open Archival Information System model for European preservation infrastructure, and PLANETS (Becker, Kolar, Küng, & Rauber, 2007), which has developed tools and services that assist with evaluating the effectiveness of preservation approaches. New Zealand and Australia have also participated in the development of tools to assist with data preservation, such as the National Archives of Australia's Digital Preservation Software Platform (National Archives of Australia, 2010) and the National Library of New Zealand's Metadata Extraction Tool (National Library of New Zealand, 2007). The Library of Congress Transfer Tools, for example, assists with the transfer of digitised heritage data from curators to the heritage data repository and is able to authenticate that the

files have been unaltered during the digitisation and transfer process (Ashenfelder, 2009). An example of a large-scale digital heritage repository is Europeana, which aggregates heritage data from archives, museums and libraries across Europe and Eurasia (Petras, Hill, Stiller, & Gäde, 2017). Europeana contains more than 53 million objects for which it is able to provide detailed descriptions, thumbnails and references to the institution that owns the object (Petras, et al., 2017). The UNESCO World Heritage Centre is another example of a repository that is consistently updated, has a target audience of public users and researchers, and currently references 1121 sites from its World Heritage List[6].

While much of the focus of research for preservation in Europe and the USA is centered around the importance of trustworthiness and auditing, developing countries are faced with very different, practical, concerns with regards to preservation of historical data (Suleman, 2011). Africa presents unique instances of the problem of preserving historical data given the sheer number of important collections needing to be preserved and a need for the skills and funding required for this (Suleman, 2011). Suleman explains that while there is a need to address the problems of trustworthiness and auditability in the global North-West, Africa has the need for strategies of digital preservation that solve the problems typical of developing countries, where the heritage data is prone to dilapidation and inaccessibility (Williams, et al., 2010). An example of this can be seen in the case of the Khoisan people from Southern Africa. Williams, et al. (2010) state that the entire generation of Khoisan people with a specific cultural background will have passed on in the coming years, meaning that there is a great need to preserve whatever physical artefacts or oral history and knowledge remain so that this information can be accessed in the future. Suleman (2011) identifies a number of problems related to preservation that must be considered within the African context: artefact deterioration, rewriting history, skills and education, funding, and Internet bandwidth. Despite these issues, there is a number of digital preservation initiatives that have been undertaken in the African continent. These include the Mapungubwe Collection (Brand South Africa, 2017; Huffman, 2000), the Timbuktu Manuscripts (Minicka, 2006), the Kirby Collection (University of Cape Town, 2007), the Digital Imaging South Africa at the University of KwaZulu-Natal (Pickover & Peters, 2002) and the Bleek and Lloyd Collection (Skotnes & Bleek, 2007). The Bleek and Lloyd Collection, for instance, contains artefacts pertaining to the Khoisan people of Southern Africa. A visual dictionary developed by Williams, et al. (2010) for understanding the |xam Khoisan language (named BOLD) showed promising results where 78% of subjects believed the system helped them acquire knowledge in African cultural heritage and 100% of subjects believed the system is useful for learning about the |xam language. This is promising, given that engagement with historical data was previously found to be a significant problem (Olojede & Suleman, 2015).

2.2 Information retrieval

Typically, heritage data is preserved in an information retrieval system, that is, a system that is capable of the storage, retrieval, and maintenance of information (Kowalski, 1997). Common examples of information retrieval systems are commercial search engines such as Google or Bing. These commercial search engines typically allow users to search for web pages or media using a query, that is, a set of terms that are a representation of a user's needs. The search engine finds documents (e.g.: web pages or media) that are relevant to the given query.

[6] https://whc.unesco.org/

Relevance refers to the usefulness of a document result to a user for a given query. When the search engine retrieves all the relevant documents for the query, these documents are ranked by relevance and then presented to the user.

Text-based retrieval is conducted through the use of a number of potential models. A popular model is the Standard Boolean Model. Frants, Shapiro, Taksa, and Vladimir (1999) describe the origins of this model, which dates back to the late 1950's where it was used as the basis for document selection in almost every information retrieval system, even going into the 21st century where it was the standard. This is a simple method of text retrieval that determines the relevance of a document to a user's query by conducting an exact matching to the terms in the query and the terms in the documents (Lashkari, Mahdavi, & Ghomi, 2009). This model uses Boolean algebra and set theory to conduct retrieval and all terms in the query are given equal importance (Lashkari, Mahdavi, & Ghomi, 2009). Despite its simplicity, Frants, et al. (1999) notes that there has always been a large discussion around the drawbacks of the model. Frants, et al. (1999) note early criticisms of this model stemming from the 1980's were predominantly around the difficulty users experienced in understanding the Boolean logic, and this led to an increased difficulty in users being able to formulate their queries. Frants, et al. (1999) also acknowledged that a major criticism of Boolean information retrieval systems is the poor ranking performance. The poor ranking performance of Boolean systems was mitigated to some extent by using weights to give importance to query descriptors (Lancaster, 1968) but Frants, et al. (1999) note that this did not prevent further criticism of the ranking performance of the Boolean model compared with other approaches. Despite this, the Standard Boolean Model is still largely popular and adopted in many modern day information retrieval systems (Lashkari, Mahdavi, & Ghomi, 2009). Another popular model of text retrieval that addresses some of the drawbacks of the Standard Boolean Model is the Vector Space Model. The need for a vector retrieval model to address the shortcomings of the Standard Boolean Model was identified as early as the 1980's by Salton, Buckley, and Fox (1983). This relies on a vector that represents each document in the collection, and each vector component is a representation of a concept or term in the document (Wong & Raghavan, 1984). Retrieval of documents is conducted by determining the similarity between the query vector and document vectors, and ranking documents by similarity (Wong & Raghavan, 1984). As a result of this, the vector retrieval model was able to improve on the relatively poor ranking performance of the Boolean model (Salton, Buckley, & Fox, 1983).

Text-retrieval also has a variety of pre- and post-processing strategies that can assist with improving retrieval effectiveness. One such popular technique is that of stopword removal. Stopwords are words that, while meaningless with respect to information retrieval, are frequently occurring (Lo, He, & Ounis, 2005). As a result of their ability to be frequently matched, while offering very little semantic meaning, it has been argued that there is value in removing stopwords from documents during indexing-time (pre-processing) as well as query-time (post-processing) (Lo, He, & Ounis, 2005). Another strategy for improving the retrieval of text data is the use of thesauri in retrieval. Thesauri are used in information retrieval systems to solve the problems of word mismatch (Xu, 1997), that is, being able to match different expressions that are semantically similar (Imran & Sharan, 2009). Thesauri are implemented through a technique known as query expansion, where additional terms are appended to the original query to potentially improve retrieval effectiveness (Imran & Sharan, 2009). In the case of thesauri, synonyms of the tokens in the original query are appended to the query with the expectation that they may solve

the issue of word mismatch (Imran & Sharan, 2009). Stemming is another popular technique used for the improval of text retrieval (Tala, 1999). To stem a word is to provide its base word through mapping different morphological variants of it (Tala, 1999). Stemming in information retrieval follows the assumption that words that share the same base word (i.e. stem) are semantically similar (Tala, 1999). Based on this, stemming can be used for pre-processing and post-processing to potentially improve text retrieval by matching words that are morphological variants of the same stem (Tala, 1999).

With respect to image searching, historically this was done through the annotation of the images – reducing the task to essentially text-based searching (Carvalho, et al., 2018). Guangxin, Cai, Li, Yu, and Tian (2014) explain that features are extracted from the images themselves and similarity is determined by the cardinality of the features between images (i.e. the proximity of similar feature data points between the images). These features can be a variety of elements that exist within images, such as shapes, or colour composition. An example of a shape-based image processing algorithm is the Edge Histogram algorithm (Sun Won, Kwon Park, & Park, 2002), which captures information regarding the spatial distribution of edges that exist in images. An example of a colour-based image processing algorithm is the Colour Layout algorithm (Kasutani & Yamada, 2001), which is able to extract dominant colour data by dividing an image into sub-images. There are also algorithms such as the Colour and Edge Directivity Descriptor (Chatzichristofis & Boutalis, 2008) that incorporate both data regarding edges, as well as fuzzy colour, in their processing. Given that these algorithms all extract features from images through different protocols and by focusing on different elements within images, there is the potential for certain algorithms to perform better on a given dataset than other algorithms (due to an inherent similarity between images for a given dataset). There is now also emphasis on the concepts within the images and the content they depict. Zhao, Wei, Sui, Zhu, and Lo (2013) explain that by checking for similarity of these concepts, over and above textual and visual features, one would be able to see far better retrieval performance. This has resulted in the proposition that it may in fact be more beneficial to the user if they are presented with a diverse subset of results, rather than results conforming to a limited concept range. Hare and Lewis (Hare & Lewis, 2013) argue that this would increase the probability of the user finding images that are relevant to their query. Seah, Bhowmick, and Sun (2014) suggested a protocol similar to this, where results of user queries are grouped into similar concepts, enabling the user to view a spectrum of relevant images and to determine themselves by which axiom they were seeking relevance to the query at hand. An example of the use of these image-retrieval algorithms for cultural image data retrieval can be seen in the work done by Olojede (Olojede & Suleman, 2015), who sought to investigate the best performing image processing algorithms (that analyse features extracted from images) using African rock art images as the training data. Olojede developed a Content Based Image Retrieval (CBIR) system, which took as input an image from the user's smartphone camera and returned a ranked list of results containing rock art from the Cederberg rock art site in the Western Cape, South Africa.

2.3 Multimodal retrieval

While information retrieval was traditionally restricted to text-based or image-based searching, where retrieval of unimodal documents (i.e. only text- or image-based documents) was done through the use of a unimodal query (i.e. a query in either text or image format), this has now been extended to the possibilities of multimodal retrieval.

This is described by Long, Cao, Wang, and Yu (2016) as the retrieval of results, of more than one format, that are semantically relevant as a response to a unimodal query (such as the simultaneous retrieval of images and text data through a single text- or image-based query).

Text-based searching, image-based searching, and multimodal retrieval all present their own unique subsets of challenges. Multimodal retrieval poses an amalgamation of the challenges presented by the former pair. The ultimate goal in multimodal retrieval is to be able to retrieve images and text similar to a text or image query (Carvalho, et al., 2018). While multimodal retrieval has shown to be a great benefit in the product search space, Laenen et al. claim that multimodal retrieval's gain in popularity is also largely a result of the copiousness of available multimodal data and the strides in the deep learning visual processing field (Laenen, Zoghbi, & Moens, 2018). Multimodal information was also shown to be beneficial to the indexing and retrieval of images (Meng, et al., 2015). Roy, Ghosh, Basu, Gupta, and Ghosh (2018) note that multimodal retrieval must consider the challenges and propositions of image- and text-based searching, whilst also tackling the main issue of non-document-aligned data. This issue can potentially be handled in multimodal retrieval by embedding data of all the different available formats into a shared space wherein they can be compared to one another for similarity (Carvalho, et al., 2018). Roy, et al. (2018) use this approach as well, not only to solve the issue of non-document aligned data of the same format, but also for data across multiple sources. This approach was shown to be successful by Carvalho, et al. (2018), where it was used for a multimodal retrieval system for cooking recipes. In this way, data that was similar in their datasets were closer in the shared space, than data that were dissimilar. They achieved this through the optimisation of a minimising loss function that was learned by calculating the cost of violating distances between similar and dissimilar documents in a shared latent space that allowed the different representations to be compared. By keeping the size of the feature vectors to a minimum, thereby narrowing the search space, they were also able to maintain satisfactory performance (Carvalho, et al., 2018). Laenen et al. approach multimodal retrieval and ranking through the use of an artificial neural network that is able to determine the semantic relationship that exists between textual and visual content, and ranks images based on their visual similarities to other images (typically the query image) and their semantic similarity to other text (typically a text query) (Laenen, Zoghbi, & Moens, 2018).

2.4 Bias

Given that information retrieval systems potentially make use of several different algorithms, they are susceptible to containing algorithmic bias. Algorithmic bias is often a result of the algorithm being an implicit reflection of the values of the individuals responsible for the development of the algorithms (Beer, 2019). Wilkie & Azzopardi define algorithmic bias in retrieval as the favouring of specific documents by algorithms, for reasons unrelated to the relevance of the document (Wilkie & Azzopardi, 2017). Hamilton, Karahalios, Sandvig, and Eslami (2014) explain that this often goes unnoticed by users, as even when they are accustomed to a system, they are largely unfamiliar with the effects algorithms have in the filtration of information that is shown to them. To this effect, algorithms within systems can be seen as "gatekeepers", given their responsibility in the determining of the information that is omitted and presented to end-users (Bozdag, 2013).

An emphasis in research towards bias in algorithms has led to an inspection of the elements of bias within algorithms, often referred to as the FAT (fairness, accountability, and transparency) of algorithms (Shin & Park, 2019). This is particularly of concern to opaque algorithms, given their potential to turn into risks, and this has largely motivated for the discussion surrounding FAT (Diakopoulos, 2016). Algorithmic fairness is referring to the ideal that algorithms, in their judgements, should not be unfair or discriminatory (Yang & Stoyanovich, 2017). This can have harmful effects on the livelihoods of individuals considering the role algorithms have in the deciding of financial loans, job recruitment and real estate allocations, and the potential for algorithms to discriminate applicants on the basis of their gender, race, qualifications and capabilities (Shin & Park, 2019). Algorithmic transparency can refer to various things depending on the algorithm in question. It can either refer to disclosing to the user the data about them that the system is implicitly gaining, or, the algorithm being open about its reasoning and data management (Ananny & Crawford, 2018). Once again this is also pertinent to opaque algorithms, and algorithmic transparency can be vital towards understanding the workings of algorithms that remain hidden for proprietary reasons or due to their complexity (Shin & Park, 2019). An instance where algorithmic transparency would have been beneficial would have been in the instance where Facebook's role in the alleged Russian meddling in the 2016 USA Presidential Election's was questioned (Shin & Park, 2019). Lastly, algorithmic accountability refers to the ideal that those responsible for the development (or management) of an algorithm are ultimately responsible for the effects that the system has on the public (Diakopoulos, 2016; Shin & Park, 2019). Shin et al. describe an "accountability loophole" that exists within the current Software Development Lifecycle, where developers cannot make important business decisions, and management are uninformed of the risks surrounding decisions pertaining to the design of algorithms (Shin & Park, 2019). This accountability loophole exposes entities to risks stemming from the consequences of their algorithms, risks that they should be held accountable for (Shin & Park, 2019). Examples of this can be seen in the propagation of "fake news" by news recommender systems, where the responsibility of the propagation of the "fake news" is blurred between the creator of the content and the creators of the algorithms that propagate the misinformation (Shin & Park, 2019).

An example of algorithmic bias can be seen with PageRank, which has a bias for older and more linked pages compared to newer pages (Wilkie & Azzopardi, 2017). An example of this in retrieval can be seen in the study done by Otterbacher et al. that found that when searching for images of a "person", they were able to retrieve significantly more images of men than women (Otterbacher, Bates, & Clough, 2017). Otterbacher et al. also found that retrieval of images seemed to reflect, and sometimes exaggerate, the stereotypes of modern society. This was observed in their study in the case when searching for images of a "doctor", where they retrieved more images of men once again than their female counterparts (Otterbacher, Bates, & Clough, 2017). Otterbacher et al. noted that this is a harmful example of representation bias as it influences searchers' view of the actual gender distribution of the given topic (Otterbacher, Bates, & Clough, 2017). During their study, Wilkie & Azzopardi found that they were able to quantify algorithmic bias using a measure known as retrievability (Wilkie & Azzopardi, 2017). They define retrievability as the probability of a document to be retrieved by any given query irrespective of the document's relevance (Wilkie & Azzopardi, 2017). They then found bias to be the distribution of the retrievability scores across the collection, such that bias is an inequality among the retrievability scores of the documents in the collection (Wilkie & Azzopardi, 2017). Wilkie & Azzopardi (2017) ultimately found that the contributing factors of bias in an information retrieval system were the documents and their representation, how the system is used by the user, the retrieval model, the indexing process, and the system configurations (Wilkie & Azzopardi, 2017).

The machine learning community have also been doing much work in determining ways to mitigate bias in retrieval systems. Hube & Fetahu (2019) made the case that bias is the presence of one or both of: inflammatory terms and non-factual statements that defy consensus. They adopted a neural approach to mitigate bias in the system, with an emphasis on capturing dependencies between words to determine bias in a phrase. Hube & Fetahu (2019) state that hand-crafted features are unable to capture the dependencies between terms in a phrase that are far apart within the phrase, and that their neural approach was able to solve this issue.

Work must also be done to manage the potential biases present in the data within an information retrieval system. Many studies have shown that popular user-curated wikis, such as Wikipedia, are found to contain biased language with respect to gender, race, and religion (Wagner, Garcia, Jadidi, & Strohmaier, 2015; Graells-Garrido, Lalmas, & Menczer, 2015; Greenstein & Zhu, 2012; Knoche, Popović, Lemmerich, & Strohmaier, 2019). With respect to gender biases, a study by Wagner, et al. (2015) found that Wikipedia presented gender-related words, as well as relationship-related words, more frequently for articles about females than for articles about males. A separate analysis of bias in Wikipedia discovered a greater prominence of gendered, familial, and artistic words in articles about females, while articles about males contained a greater use of sport-related vocabulary (Graells-Garrido, Lalmas, & Menczer, 2015). Graells-Garrido et al. also found that articles about males more frequently contained words about cognitive processes, whereas articles about females more frequently contained vocabulary that related to their sexuality. Political bias was also found in Wikipedia articles by Greenstein and Zhu (Greenstein & Zhu, 2012) who detected a greater use of "democrat's phrases" between 2002 and 2003, but a decline in this over time (despite still being prominent in 2011). A study by Knoche, et al. (2019) tried to identify the biases present in three wikis: Conservapedia, RationalWiki and Wikipedia. Knoche, et al. (2019) explain that Conservapedia traditionally adopts a conservative viewpoint whereas RationalWiki is traditionally more liberal. In their findings, Knoche, et al. (2019) observed biases in Wikipedia and Conservapedia in the form of strong associations between females and art, while males had strong associations with science. These associations were found to be much weaker in the traditionally liberal RationalWiki. Knoche, et al. (2019) found that Conservapedia also had a strong bias for unpleasant words to be associated with black people's names, whereas white people's names were strongly associated with pleasant words. Lastly, Knoche, et al. (2019) found that religious biases were also detected, where Conservapedia associated Christianity strongly with pleasant, yet Islam with the unpleasant; these associations were weaker in the other two wikis. Their research does well to highlight the potential for biases to be introduced through user-curated data, subconsciously and consciously. It is then reasonable to expect that the metadata for precolonial African heritage data, which has been annotated and curated by individuals who do not necessarily share the same religion, gender and cultural history as those who were colonised, contains biases from the curators of these artefacts. These biases need to be accounted for in information retrieval systems that contain African historical data, and techniques to mitigate these biases need to be explored.

2.5 User-control for user experience and bias mitigation

A potential solution to the problem of bias in retrieval systems is the notion of control to assist users with overcoming algorithmic errors and biases (Burrell, Kahn, & Jonas, 2019). While much early research regarding user-control of retrieval systems was conducted to improve retrieval effectiveness, mainly in the form of

Interactive Query Expansion (IQE) (Nemeth, Shapira, & Taeib-Maimon, 2004; Efthimiadis, 1993; Ruthven, 2003; Fonseca, Golgher, Possas, & Ribeiro-neto, 2005), there has been less of an emphasis on it for mitigating bias in recent times. In fact, when dealing with bias and algorithmic fairness, research tends to go in the opposite direction of user-control by opting to develop and automate the process based on "expert" qualifications and technical validation (Burrell, Kahn, & Jonas, 2019).

For the early research about IQE, much of the justification for it was the control that it gives to the user, as the user has their criteria for relevance and thus is more capable of making decisions about what will be useful to their search (Ruthven, 2003) and that their interaction with the system is important if performance is to be improved (Anick, 2003). The problem with IQE seemed to be the conflicting findings with its ability to improve retrieval effectiveness, with some research noting improvements of up to 52% in precision (Ruthven, 2003; Fonseca, et al., 2005) and others finding no significant difference in performance between IQE and Automatic Query Expansion (Nemeth, Shapira, & Taeib-Maimon, 2004; Beaulieu, 1997). Explanations for this were that users required adequate instruction on how to use IQE effectively as searchers were found to not frequently make performant expansion decisions (Ruthven, 2003). Other possibilities for user-control in search engines are Advanced Search Features, however research has also shown this has a low amount of usage from users (Nemeth, Shapira, & Taeib-Maimon, 2004; Spink, et al., 2001). Despite the inconclusive effect on retrieval effectiveness, it was found that users still had a preference for the inclusion of IQE in search engines and expressed that it made their overall search experience easier and would want it integrated into their frequently used search engines (Nemeth, Shapira, & Taeib-Maimon, 2004). This is supported by the notion that strengths of user-control were found to be the increased self-efficacy felt by users as well as the increased levels of trust and freedom when using the systems (Orji, Oyibo, & Tondello, 2017). This is also consistent with research done in 2020 that experimented with user-control in privacy dashboards and found that greater levels of trust can be found with greater levels of user-control (Herder & van Maaren, 2020).

Applications of user-control are also demonstrated in other research areas such as user interface design (Rahdari & Brusilovsky, 2019), educational systems (Orji, Oyibo, & Tondello, 2017), adaptive systems (Tsandilas & Shcraefel, 2003) and recommender systems (Spink, et al., 2001). Research comparing system-controlled and user-controlled personalisation revealed the need for users to control personalisation to ensure the feeling of freedom, especially in the instances where there is a lack of trust in the personalisation system (Orji, Oyibo, & Tondello, 2017). An experiment for a user-controlled interface for the recommendation of research papers using interactable sliders also showed a positive extensive use of the sliders added to the system (Rahdari & Brusilovsky, 2019). An investigation of Twitter users' perception of the "Twitter algorithm" revealed that many users attempt to explicitly and implicitly manipulate the algorithms through their behaviour on the platform (Burrell, Kahn, & Jonas, 2019). It was found that many users had developed strategies so that they were able to restrict their community on the platform, or to ensure their message was broadcasted to as wide an audience as possible, depending on where they felt the platform had failed them (Burrell, Kahn, & Jonas, 2019). This research ultimately led to the conclusion that there is an overlap in the concepts of transparency, fairness and control, and that user-control of systems has a desirable purpose, which provides social utility, enforces values such as human autonomy and encourages productive human-machine partnerships (Burrell, Kahn, & Jonas, 2019). The research went as far as to say that

the lack of capacity for human intervention in algorithmic decision making is a "threat to human dignity" (Jones, 2017).

2.6 Decolonisation of retrieval

In 2015, student-led protests throughout universities in South Africa for the #RhodesMustFall movement brought recognition and promoted discourse in South African mainstream media regarding the inequalities that were systematically built into tertiary education institutions (Kessi, Marks, & Ramugondo, 2020). This further propagated prior discourse surrounding decolonisation, and the conversation of decolonisation has since spread to the United States and the United Kingdom, where students want cultural and systematic changes to institutions and their respective curricula (Gill, 2018). Decolonisation is the process of unlearning unjust practices and ideals and dismantling institutions with the intention of facilitating other perspectives of knowledge (Kessi, Marks, & Ramugondo, 2020). Through decolonising, society can recognise that historically marginalised groups are agents of their own experiences and pasts (Parker, 2016; Memmi, Greenfeld, Sartre, & Gordimer, 2013; Fanon, Sartre, & Farrington, 1963) and can provide the foundation to redeem African theory across diverse disciplines and domains (Kessi, Marks, & Ramugondo, 2020).

Four dimensions of decolonisation have been identified as important to the initiative, these being: structural, epistemic, personal and relational (Kessi, Marks, & Ramugondo, 2020). Of these, what should be of interest to information sciences and technology is the epistemic dimension. Epistemic decolonisation revolves around the notion that Euro-American knowledge is not necessarily objective or universal and instead, at the very core, concepts such as positionality, subjectivity and situatedness are of importance as well (Kessi, Marks, & Ramugondo, 2020). As such, there is the encouragement to reclaim alternate theory and perspective that is not entrenched in Euro-American theory and to explore alternate techniques to solve problems (Kessi, Marks, & Ramugondo, 2020; Parker, 2016; Bhabha, 1983; Gill, 2018; Brian, 2018).

Given that museums are epistemic spaces (Mbembe, 2015), it is reasonable to suggest that digital retrieval systems containing the very heritage data present in museums are also in need of undergoing the process of decolonisation. Research conducted on the Google search engine's retrieval integrity for the query term: "black girls" (Noble, 2013) was noted for undertaking "decolonising work" due to Noble's questioning of the structure and nature of web searches and results in relation to social justice (Parker, 2016) and questioned the inability for users to be able to "see Google's algorithm" to understand the associations made by the search engine. A core concept of decolonisation – interrogation – strongly encourages questioning what qualifies as "expert", why certain knowledge is presented, and why certain knowledge is omitted (Kessi, Marks, & Ramugondo, 2020). The research on the Google search engine strongly proposed the reevaluation and reimplementation of the digital tools that students are prompted to use for research and proposed that there is value in the conversation surrounding the examination of information retrieval results in relation to bias and misrepresentation (Noble, 2013; Parker, 2016).

2.7 Summary

There have been global efforts towards the preservation of heritage data. This data is highly valued and, without preservation efforts, could be soon lost entirely due to artefact degradation and the lack of emphasis placed on oral history. Unfortunately, in the case of African heritage data preservation, research needs to be steered away from the direction being followed in the global North-West and instead needs to be focused on addressing the practical concerns that African heritage collections are faced with, such as artefact deterioration.

These challenges can be addressed using information retrieval systems that allow for the heritage data to be browsed through the use of queries that return ranked lists of results. Through advancements in information retrieval technology, we are no longer restricted to the use of unimodal queries to retrieve unimodal results and are now able to combine data of different modalities and retrieve them through multimodal queries. This allows for the preservation and exploration of both image and text heritage data. A problem with using information retrieval systems is the inherent bias present in both the algorithms and metadata. This is especially concerning when dealing with sensitive data such as heritage data, where misattribution or misrepresentation of the data can have serious social and technical consequences. This problem of bias can potentially be solved through the use of user-control, which allows the user to manipulate a system in a manner that produces their desired output. While this has been used in many fields, there has not been much recent work adopting user-control for the mitigation of bias in information retrieval systems, as research in this regard typically moves in the opposite direction of automation. Despite this, the notion of user-control strongly aligns with the principles of decolonisation, and this warrants the exploration of it to mitigate bias in information retrieval systems containing African heritage data.

Chapters 3 and 4 outline the two experiments conducted in an attempt to find answers to the research questions. These chapters also contain the findings from these experiments and are followed by a discussion of these findings, answers to the research questions, and the subsequent conclusions in Chapter 5.

3. Experiment 1: Offline evaluation of pre-processing algorithms for text and image retrieval

3.1 Overview

The intention behind Experiment 1 was to provide a baseline offline evaluation of pre-processing algorithms for text and image retrieval to determine if algorithmic variation has the potential to have a significant difference on retrieval effectiveness. A suitable number of sample queries and relevance judgements needed to be obtained for use during the experiment given the lack of a conveniently available testbed containing African heritage data queries and results. An information retrieval system was built using Apache Solr containing multiple indexes to support multiple pre-processing algorithm implementations. The image retrieval engine, built using LIRE, had an accuracy parameter that needed to be optimized for each of the image-retrieval algorithms to ensure that algorithms were compared at their respective best-performing levels. A multimodal information retrieval system, using the image retrieval engine, as well as providing text retrieval functionality, needed to be developed with an index of the FHYA dataset containing African heritage data. Participants were also asked to provide relevance judgements for results retrieved by various retrieval algorithms so that retrieval metrics across all algorithms in question could be calculated. The following sections below outline the processes followed to achieve this, as well as the findings from the experiment.

3.2 Testbed

This section provides a description of the Five Hundred Year Archive[7] dataset that was indexed into the multimodal information retrieval system used in this study. The process followed to create a collection of information needs and text- and image-based queries is also outlined.

3.2.1 *The Five Hundred Year Archive*

The Five Hundred Year Archive project, an initiative of the Archive and Public Culture (APC) at the University of Cape Town, is a digitised archive containing artefacts mainly from the KwaZulu-Natal region in South Africa from between the periods of 1750 to the late nineteenth century (The Five Hundred Year Archive, n.d.). The name for this initiative is derived from the 500 years of European presence in Southern Africa during the Apartheid era and attempts to bring to light the five hundred years, pre-colonisation, of African heritage (The Five Hundred Year Archive, n.d.). The dataset of the FHYA consists of multimodal data pertaining to botany, oral history, ethnology, archaeology, ethnography and more (The Five Hundred Year Archive, n.d.). These artefacts were curated based on the variety of technical challenges they would present to a heritage digitisation initiative (The Five Hundred Year Archive, n.d.). Artefacts in the archive that were produced pre-colonisation consist of objects that were either excavated, or collected, and texts that were written in the precolonial times (The Five Hundred Year Archive, n.d.). There is also material that, while produced after the commencement of formal colonisation,

[7] http://www.apc.uct.ac.za/apc/research/projects/five-hundred-year-archive

refer to precolonial times (The Five Hundred Year Archive, n.d.). The FHYA also consists of transcribed oral material, as well as audio files, although no audio files were included in this study.

The FHYA dataset provided for use in this study consists of artefacts that have an image (in some instances multiple images) and related metadata. These images were supplied in a JPEG format, and the metadata supplied in XML format. The metadata consisted of descriptive information of the artefacts such as the name of the parent collection, the title of the artefact, a material designation (e.g.: textual record, object, sound recording etc.), a unique identifier, a list of events with associated dates (e.g.: curation, custody, collection, making etc.) as well as a list of associated image or audio media file names. There also existed a corresponding list of titles for the associated media file names. Some artefacts, along with possible image and audio media, also had PDF files referenced. Some of these PDF files were over 20 pages in length. For efficiency of indexing, storage, and retrieval, only the first 5 pages of any PDF file were used in this study.

This archive initiative is still ongoing in 2020, where the system is being tested with the intention of improving the user experience through improving accessibility and the user interface (The Five Hundred Year Archive, n.d.). The expectation is that this research initiative will further motivate for the digitisation of other materials from this region and greater southern Africa by promoting the value of investigation of precolonial African heritage (The Five Hundred Year Archive, n.d.). The APC have graciously allowed for the FHYA dataset to be used in the study as the document collection in the multimodal information retrieval system. The subset of the FHYA used for this study consisted of 1345 text documents, containing detailed descriptions of various heritage African artefacts, as well as a total of 5708 images that were also indexed for use in this study. The reason for the significantly higher number of image documents compared to text documents is because of the fact that many artefacts contained multiple image views. Unfortunately, permission was not granted to include any FHYA data in any research papers. However, examples of the metadata and image data can be seen in Fig. 3.1, which were taken from the public FHYA website.

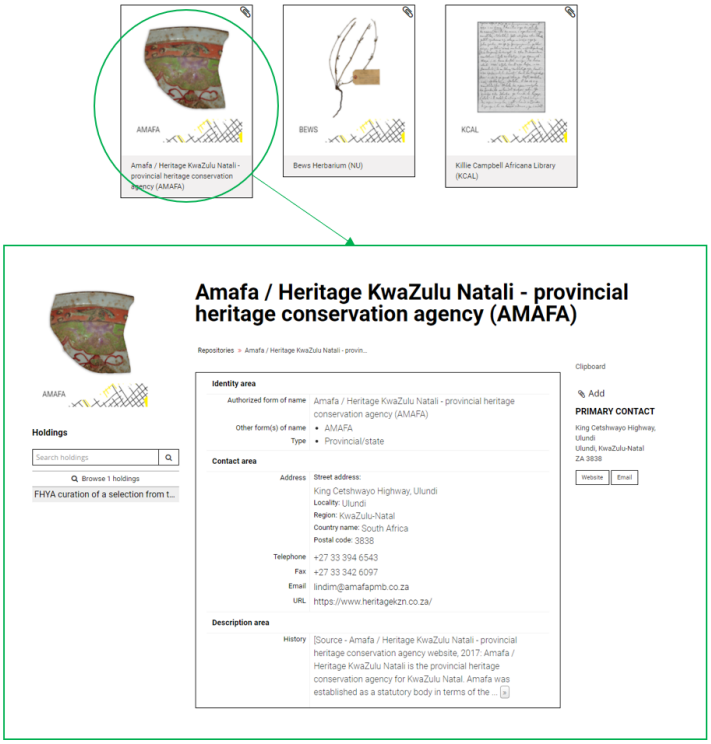

Figure 3.1: Example metadata and image data from the Five Hundred Year Archive[8]

3.2.2 *Information Needs*

Given that the data used in the information retrieval system was precolonial African heritage data from the Five Hundred Year Archive, the experiment required sample queries, as well as relevance judgements for results to evaluate retrieval effectiveness.

The Google Trends tool was used to obtain a list of candidate information needs and associated queries. This was done by submitting topics, such as “History of Africa”, into the tool and storing the related queries for the topic, as well as parsing any related topics and their respective related queries. See Figure 3.2 for an example of the related topics and queries for the topic “Zulu people”.

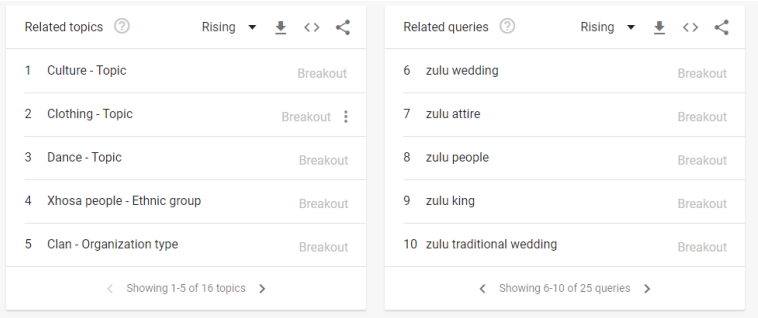

Figure 3.2: The related topics and queries for the topic “Zulu people” on Google Trends

[8] https://fhya.org/amafa-heritage-kwazulu-natali-provincial-heritage-conservation-agency-amafa

This produced a list of 57 candidate information needs and 318 candidate queries to be used for the study. The candidate information needs, and queries, were graded based on their respective relevance to the subject domain of precolonial African heritage data by the Archive & Public Culture Research Initiative. This was done to ensure that any queries used for the experiments were obtained from information needs that were expertly verified as being relevant to the subject domain. The results of this can be seen in Table 3.1. The 57 candidate information needs were thus filtered to the 54 that were identified as being at least relevant (if not highly relevant) to the subject domain.

Table 3.1: Graded relevance of candidate information needs by the Archive & Public Culture Research Initiative

Information Need	**Type**	**Relevance Judgement (mark with an X)**		
		Not Relevant	**Relevant**	**Highly Relevant**
Tradition	Belief			x
Beadwork	Topic		x	
Zulu people	Ethnic group			x
Ancient history	Topic			x
History of Africa	Topic			x
Religion	Topic		x	
Traditional African religions	Belief		x	
Xhosa people	Ethnic group			x
Sotho people	Ethnic group			x
Xhosa language	Spoken language			x
Hymn	Composition type		x	
Gospel music	Musical genre	x		
Worship	Topic		x	
Leadership	Topic			x
Community	Topic			x
Individual	Topic		x	
Person	Topic		x	
Old age	Topic		x	
Health care	Professional field		x	
Society	Topic		x	
Human	Primate		x	
Nature	Topic		x	
Anatomy	Field of study		x	
Safety	Topic		x	
Health	Topic		x	
Eating	Topic		x	
Physical fitness	Topic		x	
Healing	Topic		x	
Religious text	Literary genre		x	
Fruit	Topic		x	
Meat	Food		x	
Community development	Topic		x	
Innovation	Topic		x	
Skill	Topic		x	
Praise	Topic			x
Clan	Organization Type			x
Music of Africa	Musical Genre			x
Dance	Topic			x
Africans	Topic			x
Black people	Topic			x
Symbol	Topic			x
Religion in South Africa	Topic	x		
Culture	Topic			x
Indigenous Peoples	Topic			x
Zulu language	Spoken language			x
Zulu kingdom	Topic			x
Statue	Visual art form	x		

Design	Topic		x	
Northern Sotho language	Spoken language		x	
Tswana people	Ethnic group		x	
Venda language	Spoken language		x	
Tsonga people	Ethnic group		x	
Southern Ndebele people	Ethnic group		x	
Swazi people	Ethnic group			x
Healer	Topic			x
Spirituality	Topic		x	
Painting	Topic		x	

3.2.3 *Text Query Formulations*

Now that a list of 54 relevant information needs was acquired, and associated relevant queries, these associated queries needed to be reduced to a collection that was suitable for use in the experiment. This entailed submitting all these candidate queries to the information retrieval system (to be discussed in section 3.3) and storing the retrieved results for each respective query. Any query where less than 10 results were retrieved was removed from being a candidate for use in the experiments. This was to confine the query collection to only queries that were able to retrieve sufficient results for analysis and relevance judgements. From this process, a final list of 70 sample text-based queries were obtained (see Appendix A).

3.2.4 *Image Query Formulations*

In order to obtain a suitable image-based query collection, an online digital library containing such data was sourced. The Smithsonian National Museum of African Art (NMAfA)[9] was identified as being a suitable candidate, given it contains traditional African art and artefacts, as well as having suitable accessibility. The NMAfA allow for image files to be used for educational use under "fair use", thus permission was not required to be obtained from NMAfA. NMAfA strictly refuse the publishing of Smithsonian Institution images for commercial gain.

The image query collection was thus built by crawling the NMAfA collection and storing links to images in the collection. The links to 298 available images (at the time) in the NMAfA collection were stored in an HTML file, such that each image was displayed beneath one another when opened. This was done due to the convenience of being able to open this HTML file in a browser to view all these images quickly and efficiently. This HTML file was sent to 3 participants (participant information to be outlined further in 3.5), and participants were asked to respond with a list of the URLs for all of the images that they believed were not relevant to the subject domain: "Precolonial South African Heritage". All images that were identified by any single participant as not being relevant to the subject domain were removed from being candidate images to be used in the experiment. From this, a total of 133 filtered images remained as candidates for use in the experiment. 70 images were then randomly sampled from this list of 133 images (for consistency with the number of text queries obtained in 3.2.3) to be used in the experiment. The images were all maintained at their original size and resolution from the NMAfA source. An example of an image from the collection that was crawled for use in this study can be seen in Figure 3.3.

[9] https://africa.si.edu/collections/collections

Figure 3.3: Image data of an artefact crawled from the Smithsonian Institution National Museum of African Art[10]

3.3 System Design and Implementation

The creation of a multimodal information retrieval system for precolonial, African heritage data was required for this study. The retrieval system was developed using Apache Solr[11] and LIRE, an open-source, Java plug-in to facilitate image retrieval and search within Solr (Lux, Riegler, Halvorsen, & MacStravic, 2017). The following sections outline the process followed to develop the multimodal information retrieval system, as well as the process to optimize the accuracy parameter of the image-retrieval engine across various algorithms.

3.3.1 *Text retrieval*

Solr has proven functionality and capability with respect to text retrieval, and as such was chosen to be the foundational basis of the information retrieval system and was responsible for the handling of text retrieval for the system, given a text query. The version of Solr used throughout the duration of this study was v7.7.2. Solr allows for document collections, in a variety of formats (e.g.: XML, JSON etc.), to be indexed to individual indexes, which Solr refers to as cores. It is possible to define the pre- and post-processing algorithmic procedures for each of these indexes in their schema configuration files. It was chosen to make use of this behavior in Solr by creating individual indexes for each algorithm to be investigated. This meant that, if all indexes contained the same document collection, the algorithms could be compared by querying all indexes with the same sets of queries. This, in essence, is the approach that was used to provide an offline evaluation of algorithms to determine if there can be a significant difference in retrieval effectiveness across algorithms.

For text retrieval, the retrieval system supported 3 text-retrieval algorithms: stemming, stopping and synonyms (i.e. a thesaurus). The interest behind exploring the effects of stemming is due to the inconsistency in the language used between the curators of the artefacts in the heritage data and the modern researchers using heritage data retrieval systems. There is also a mixture of languages in the data, as well as the fact that many of the languages are low-resource and do not contain a standardized vocabulary. While the queries used in this study were all in English, many of the documents contained annotations, or the names of places and people, borrowed from African languages such as isiZulu. Given the language-based nature of stemming, this could result in isiZulu words being

[10] https://africa.si.edu/collections/collections
[11] https://lucene.apache.org/solr/

stemmed and incorrectly matched with the stems of the English query phrases. It is therefore of interest to examine the effects of stemming in a dataset that contains English, as well as multiple low-resource languages. Stopping, or stopword removal, is also of interest as this could potentially improve retrieval effectiveness by promoting retrieval based on the matching of rarer, shorter words that add more meaning to documents and queries than frequently used stopwords. Lastly, this research is interested in exploring whether the use of thesauri, or query expansion via synonyms, can address the potential issue of differences in language as a result of modern researchers being in a different era, belonging to a different culture, or speaking language differently, to the curators of the artefacts. A further description and motivation for these algorithms can be found in 2.2. For stemming, Porter's stemming algorithm (using `PorterStemFilterFactory`) was used for pre-processing at index-time and query-time, because the textual data was nominally in English. Stopword removal (using `StopFilterFactory`) was also conducted for the stopping algorithm at index- and query-time. The list of stopwords were included in the stopwords.txt configuration file that is used by the `StopFilterFactory` for stopword removal. This list was obtained from the supported stopwords in Lucene's `StopAnalyzer`[12]. For the synonyms algorithm, the Wordnet thesaurus[13] was used for query-expansion only (using `SynonymGraphFilterFactory`). The WordNet thesaurus is free to use for these purposes and this was one of the main reasons for its inclusion in this study. Using the `SynonymGraphFilterFactory` requires an offline store of the thesaurus in a synonyms.txt file for the query expansion behavior, where each line entry in the file is a comma separated list of words that are to be considered synonyms of one another. Thus, if a word in the query exists in the synonyms.txt file, all the other words in the same entry as the query word will be appended to the query. Combinations of these algorithms were also added to the retrieval system. This included the following combinations (the names of the pre-processing algorithms that are used for the rest of this paper are in brackets following each description): Stemming and Stopping (stemmingAndStopping), Stopping and Synonyms (stoppingAndSynonyms), Synonyms and Stemming (synonymsAndStemming), Stemming, Stopping and Synonyms (allConfig) and a pre-processing algorithm using none of the techniques (NoConfig). Each of these pre-processing algorithms were present in the retrieval system in the form of their own core, or index, with their respective schema reflecting their index- and query-time processes. Figure 3.4 is a snippet from the managed-schema for the stemmingAndStopping core and shows how stopword removal and stemming is being done at index-time and the same for query-time. The text data from the FHYA collection were provided as several XML files. These were converted to JSON, and then indexed into every core mentioned above.

[12] https://alvinalexander.com/java/jwarehouse/lucene/src/java/org/apache/lucene/analysis/StopAnalyzer.java.shtml
[13] https://wordnet.princeton.edu/

```
<fieldType name="text_en" class="solr.TextField" positionIncrementGap="100" multiValued="true">
  <analyzer type="index">
    <tokenizer class="solr.WhitespaceTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <filter class="solr.NGramFilterFactory" minGramSize="1" maxGramSize="300"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
    <filter class="solr.PorterStemFilterFactory"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.WhitespaceTokenizerFactory"/>
    <filter class="solr.PorterStemFilterFactory"/>
    <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
  </analyzer>
</fieldType>
```

Figure 3.4: The managed-schema configuration for the stemmingAndStopping core

3.3.2 *Image retrieval*

LIRE is an open-source Java plug-in for Solr to facilitate image retrieval and search (Lux & Marques, 2013). LireSolr is an implementation of the LIRE library, built on top of a Solr search server, which makes use of deep learning processes to extract features from images (Lux, et al., 2017). 6 of the available image-retrieval algorithms in LIRE were included in the system. These were: Edge Histogram (Sun Won, Kwon Park, & Park, 2002), Pyramid Histogram of Oriented Gradients (PHOG) (Bosch, Zisserman, & Munoz, 2007), Auto Colour Correlogram (Huang, Kumar, Mitra, Zhu, & Zabih, 1997), Colour and Edge Directivity Descriptor (Chatzichristofis & Boutalis, 2008), Colour Layout (Kasutani & Yamada, 2001) and Joint Composite Descriptor (Chatzichristofis & Boutalis, 2009). Table 3.2 provides a description of each of these algorithms. Examples of images that some of these algorithms consider similar can also be seen in Table 3.3. These algorithms were selected from LIRE due to their diverse implementations and expertise. More specifically, exploring the differences between shape-based pre-processing algorithms, image-based pre-processing algorithms, and pre-processing algorithms based on both colour and shape, was of interest to determine if these elements are influential in the retrieval effectiveness of this African heritage data. The image dataset was found to contain many images that were comprised mainly of brown colours and Earthy tones. Many of the artefacts had also experienced degradation, meaning their shapes were no longer wholly intact. Lastly, there were also many images of cards that just described artefacts, many of which were simple pieces of white paper with handwriting or typed content. These elements of the image dataset largely contributed to the interest in evaluating the difference in performance between colour-based, shape-based, and colour-and-shape-based image retrieval pre-processing algorithms. The EH and PH algorithms provide the opportunity to explore the value of using shape-based image retrieval pre-processing algorithms. The AC and CL algorithms are both colour-based image retrieval pre-processing algorithms, and lastly, the CE and JC pre-processing algorithms combine both colour- and shape-detection.

Integration of LIRE with Solr was done by following the installation instructions provided with LireSolr. First, a new core was created (in this study the core was named “LIRE”). LireSolr provided a gradle task, `distForSolr`, which assisted with the creation of jar files that are required. This task was run, and the jar files were copied to the *…/opt/solr/server/solr-webapp/webapp/WEB-INF/lib/* location, following which, the custom “/lireq” `RequestHandler` and “lirefunc” `ValueSourceParser` were registered in the solrconfig.xml file.

The Solr managed schema was then edited to include the fields required for LIRE indexing (i.e. imgurl, id, title etc.). Then, the `ParallelSolrIndexer`, supplied with LIRE, was used to create an index of the JPEG image files in the FHYA collection that were then added to the *LIRE* core in the Solr system and committed.

Table 3.2: Overview of LIRE image-retrieval algorithms included in the multimodal retrieval system

Name	Abbreviation	Descriptive Name (for users)	Description
Edge Histogram	EH	Local Edge Distribution	Optimises matching performance by initially using global and semi-local edge histograms. Then determines similarity by combining global, semi-global, and local histograms. This is then compared against the MPEG-7 descriptor of the local histogram (Sun Won, Kwon Park, & Park, 2002).
Pyramid Histogram of Oriented Gradients (PHOG)	PH	Shape Similarity and Spatial Layout	Determines spatial layout information by tiling the image, across multiple resolutions, into various regions. Local shape information is also determined by iterating over regions and acquiring the distribution of edge orientations. Shape and layout similarity is thus determined by the distance between two PHOG descriptors (Bosch, Zisserman, & Munoz, 2007).
Auto Colour Correlogram	AC	Spatial Correlation of Colour	A colour correlogram – a table that represents the probability of finding a pixel of a particular colour in a specified distance from a pixel of another colour – is efficiently computed. A relative distance measure, instead of an absolute distance measure, is then used to determine similarity (Huang, et al., 1997).
Colour and Edge Directivity	CE	Colour and texture	Colour information is captured through the use of fuzzy rules with a Fuzzy-Linking histogram. Digital filters then assist with the capturing of texture information within subregions of an image. The colour and texture information is then captured in a single histogram (Chatzichristofis & Boutalis, 2008).
Colour Layout	CL	Spatial distribution of colour	Captures spatial colour distribution by determining the dominant colour in subregions of an image (Kasutani & Yamada, 2001).
Joint Composite Descriptor	JC	Fuzzy colour, edge information and texture	A combination of two descriptors, CEDD and FCTH (fuzzy colour and texture histogram). FCTH makes use of fuzzy colour with detailed edge descriptions (Lux & Marques, 2013).

Table 3.3: Examples of images found to be similar by image-retrieval algorithms in LIRE[14]

Algorithm	Images found to be similar
Edge Histogram	
Pyramid Histogram of Oriented Gradients	
Auto Colour Correlogram	

3.3.3 *LIRE Accuracy Parameter Optimization*

LIRE requires use of an accuracy parameter when making retrieval requests using "/lireq". This parameter is a double in [0.05, 1] and defaults to 0.33 (i.e. 33%). There exists a trade-off with this parameter, where higher values suggest a higher level of accuracy but with slower processing times. Given that processing times were not a major concern, the process to instead ensure that all algorithms were being used with their most retrieval-effective accuracy parameter value was undertaken. This entailed following an experiment where, for all algorithms, the respective results for 7 different image-based queries were collated across different accuracies (0.5, 0.6, 0.7, 0.8 and 0.9). This produced 7 'supersets' of results that contained all of the retrieved images for each query across the different accuracies. These supersets were then distributed to 5 participants (further information regarding participants can be found in section 3.5), who were asked to grade the relevance of each result to the respective query image using a graded relevance scale (i.e. not relevant, relevant or highly relevant). The supersets were provided in CSV format, with 3 blank columns alongside a link to each of the retrieved results. The 3 columns provided a space for participants to grade the relevance of each retrieved result (i.e. not relevant, relevant, or highly relevant) through the addition of an "X" in the appropriate column. These relevance judgements were then used to evaluate the retrieval effectiveness of the image-retrieval algorithms, across each of the individual candidate accuracy values, so that the best performing accuracy value for each image-retrieval algorithm could be determined (for the FHYA dataset). Table 3.4 shows the results of this process, with the selected accuracy value for each algorithm indicated in bold. Given that F-measure is the harmonic mean of both recall and precision, the accuracy values that produced the highest respective F-measure metric were selected for each algorithm. The only

[14] http://image-similarity.ait.ac.at/solr/lire_eval.html

anomaly observed during this process was that the Color Layout image-retrieval algorithm retrieved the same set of results for all accuracy values. The reasons for this are not known and this algorithm potentially does not make use of a parameter to configure its processing (i.e. it is not parameterised).

Table 3.4: Results from Cranfield-style experiment to determine best performing accuracy value (bold) for image-retrieval algorithms

Core	Accuracy	Recall	Precision	F-Measure
EH	**50%**	**0.910**	**0.345**	**0.487**
	60%	0.853	0.326	0.460
	70%	0.862	0.331	0.466
	80%	0.881	0.338	0.476
	90%	0.881	0.338	0.476
PH	50%	0.507	0.117	0.176
	60%	**0.512**	**0.119**	**0.179**
	70%	0.497	0.114	0.172
	80%	0.497	0.114	0.172
	90%	0.459	0.105	0.157
AC	50%	0.771	0.105	0.172
	60%	0.851	0.112	0.185
	70%	0.823	0.110	0.181
	80%	**0.958**	**0.206**	**0.323**
	90%	0.863	0.194	0.303
CE	**50%**	**0.946**	**0.133**	**0.214**
	60%	0.900	0.124	0.199
	70%	0.918	0.126	0.203
	80%	0.918	0.126	0.203
	90%	0.954	0.131	0.211
CL	50%	1.000	0.293	0.439
	60%	1.000	0.293	0.439
	70%	1.000	0.293	0.439
	80%	1.000	0.293	0.439
	90%	1.000	0.293	0.439
JC	50%	0.571	0.083	0.136
	60%	0.933	0.205	0.305
	70%	**0.914**	**0.210**	**0.310**
	80%	0.705	0.088	0.146
	90%	0.705	0.088	0.146

3.4 Experiment Design

The experiment was then conducted to provide a baseline offline evaluation of algorithms for text and image retrieval. This entailed using a similar process to the one followed to determine the best performing accuracy values (in section 3.3.3) in order to evaluate the theoretical retrieval effectiveness of the text and image retrieval algorithms. 70 text-based and 70 image-based queries (obtained from the process outlined in 3.2) were submitted to the retrieval system across all text and image retrieval pre-processing algorithms. Afterwards, the unique results from each pre-processing algorithm for a given query were recorded and a superset of results for each query was compiled. Supersets were compiled in CSV format such that the file name was the respective query, and each of

the retrieved documents were listed beneath one another. Alongside the link to each of the retrieved documents were 3 columns for relevance of the document to be graded (i.e. not relevant, relevant, highly relevant). 10 participants were then allocated 14 supersets each (7 image supersets and 7 text supersets) and were asked to grade the relevance of results in the superset to its respective query by adding an "X" in the column that best represented the relevance of the document. This can be seen in Figure 3.5, which is a snippet of a participant's response. This in turn was used to evaluate the performance of each pre-processing algorithm.

Query: african names			
Document	Not Relevant	Relevant	Highly Relevant
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=HAM-4-1		X	
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=CUL-1-3	X		
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=NU0025862-0		X	
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=NU0039266		X	
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=NU0039406		X	
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=MG-2-1		X	
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=JSA-6-19	X		
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=JSA-1-20	X		
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=JSA-1-26	X		
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=JSA-2-17	X		
http://pumbaa.cs.uct.ac.za/~soham/result.html?id=JSA-2-22	X		

Figure 3.5: Snippet of a participant's grading of relevance of retrieved documents to the query "african names"

3.5 Participants

This section describes the processes followed to recruit participants for participation in the two pre-experiments (the filtration of image queries and optimisation of the LIRE accuracy parameter), as well as for the experiment itself. Demographic information regarding participants is also provided. For the pre-experiments and formal experiment, an invitation for participation was sent out via email and on social media platforms inviting anyone aged over 18 to participate. The below sections describe how the responders of these invitations were allocated to the different experiments.

3.5.1 *Image query filtering*

3 participants were sampled using the Convenience sampling method for this pre-experiment. The participants had all responded positively to the invitation for participation sent out on social media platforms. They were selected for participation for this particular pre-experiment given their immediate availability for participation. The participants were 2 males and 1 female all between the ages of 23 and 30. One of the participants was a university student at the time, while the two other participants were working as software developers.

3.5.2 *LIRE accuracy optimisation*

The Judgement sampling method was used for this pre-experiment. 5 participants were recruited, all of whom had responded positively to the invitation for participation sent out on social media platforms. They were selected for participation for this particular pre-experiment given their immediate availability for participation, but more

importantly, because all participants had completed a Computer Science (or related) degree. This was done intentionally given it was the first time participants in this study were being asked to grade the relevance of results on a multi-graded scale (i.e. Not Relevant, Relevant or Highly Relevant), and the participants' prior experience in this field could assist them in performing these tasks efficiently and correctly. The participants were all males between the ages of 26 and 32. While all of the participants had prior university experience and qualifications with a Computer Science (or related) degree, only one of the participants was still actively pursuing their tertiary education (i.e. pursuing an MSc degree). The other 4 participants were all employed in the software development field.

3.5.3 *Formal experiment*

10 participants were recruited for this experiment using the Convenience sampling method. These participants had also responded positively to the invitations sent out requesting participation. There were no formal requirements of prior history with Computer Science (or a related field) to promote a diverse background of participants grading the relevance of results. Participants were 70% male and 30% female, with all participants aged between 23 and 35 years of age. Of the 10 participants, 4 were currently enrolled at a university, while the remaining 6 were employed.

3.6 Data Analysis

Following the experiment outlined in 3.4, a total of 140 graded superset CSV files were obtained. A program was written in C# that parsed the relevance judgements contained in all 140 superset files and was then able to calculate various retrieval metrics to provide an evaluation of all the algorithms. Metrics obtained from this evaluation assist with determining whether there is a statistically significant variation in retrieval performance between pre-processing algorithms for this dataset. Statistical tests were performed with these metrics using R and RStudio. The results of this analysis are outlined in the section 3.7.

3.7 Metrics

The metrics used to evaluate the retrieval effectiveness of the different pre-processing algorithms included recall, precision, F-Measure, NDCG, DCG and the number of documents retrieved. This section briefly explains these metrics and how they are calculated.

Recall is the proportion of relevant documents, available in the collection for a given query, that are returned as results by the information retrieval system. It is calculated as the number of relevant documents retrieved divided by the total number of relevant documents in the collection (Ting, 2011). Thus, a recall measure of 1.0 would indicate that all relevant documents, for a given query, in the document collection have been retrieved, whereas a recall measure of 0 would indicate that none of the relevant documents have been retrieved. Precision is the proportion of documents that are returned as results to a given query, that are relevant to that query. This is calculated as the number of relevant documents retrieved divided by the total number of documents that were

retrieved (Ting, 2011). A precision measure of 1.0 would thus indicate that all retrieved documents are relevant, whereas a precision measure of 0 would indicate that none of the retrieved documents are relevant. The metric, F-Measure, is the harmonic mean of the recall and precision metrics (Zhang & Zhang, 2009). An F-Measure of 1.0 indicates recall and precision were both perfect.

While the above metrics can provide insight into the retrieval effectiveness of an information retrieval system, they do not consider ranking performance. For that, the retrieval metrics DCG and NDCG are used. Discounted Cumulative Gain (DCG) is used to measure ranking quality (i.e. highly relevant documents are ranked higher than less relevant documents) and is calculated by summing the relevance scores for retrieved documents (Järvelin & Kekäläinen, 2009). These scores are discounted, such that documents ranked lower are given lower weights so that they contribute less to the value of the metric. This is usually calculated up to some position K in the list of results, and is reported as DCG@K. Normalised Discounted Cumulative Gain (NDCG) normalises DCG values by dividing them by the best possible DCG for a result list, known as the Ideal Discounted Cumulative Gain (IDCG) (Järvelin & Kekäläinen, 2002). That is, IDCG is the DCG for the ranking of results where the results are ranked perfectly with respect to their relevance to the query. Thus, NDCG values will always be between 0 and 1. An NDCG value of 1 would indicate perfect ranking performance.

3.8 Findings

Table 3.5 shows the results of nonparametric, asymptotic Friedman tests for the retrieval metrics for the image algorithms. A description of these algorithms can be found in Table 3.2. The Friedman test revealed that there was a statistically significant difference in the recall of the different image algorithms, $\chi^2(5) = 29.708$, $p < .001$. Significant differences were also found for precision, F-measure, DCG and NDCG (@5, @10) and number of documents retrieved. This would suggest that algorithmic variation for heritage image data retrieval can significantly improve retrieval effectiveness for image retrieval across many metrics. This led to further detailed inspection of the differences between algorithms. Figure 3.6 illustrates the differences between image algorithms with respect to precision and recall, indicating that AC, EH and PH outperformed other algorithms for these two metrics. Table 3.8, which contains mean values for image algorithms across all recorded metrics, also shows that these three algorithms had the best ranking performance given their high relative NDCG (@5, @10) values. The results of two-sample Wilcoxon Signed Rank tests containing further pairwise comparisons of retrieval metrics between image algorithms (i.e. pairwise comparisons of all algorithms compared to one another) can be seen in Table 3.7. P-values were adjusted using the "BH" (Benjamini, Hochberg) method due to the multiple comparisons performed. Table 3.7 shows significant differences between AC and EH (the two best-performing algorithms) for all other algorithms except between each other. The presence of EH and PH in the top 3 performing algorithms would suggest that shape-based image pre-processing algorithms potentially perform better for this heritage data retrieval compared to colour-based pre-processing algorithms (which is further supported by the worst-performing algorithm being CL). The significant differences found between AC and EH and all other algorithms further reinforces the finding that algorithmic variation can improve retrieval effectiveness of this heritage image data. However, it should be noted that while PH was found to be the third best performing image pre-processing algorithm (see Figure 3.6), there were no significant differences found between PH and any algorithm it

outperformed for recall and precision (see Table 3.7), which points to the need for possible further exploration into shape-based vs colour-based image retrieval pre-processing algorithms.

While significant differences were recorded between image pre-processing algorithms, this was not the case for the text pre-processing algorithms. For text retrieval pre-processing algorithms, Table 3.6 shows that significant differences were only observed for precision ($\chi^2(7) = 47.801$, $p < .001$), F-measure and number of documents retrieved. This would suggest that algorithmic variation of text pre-processing algorithms is not able to significantly improve retrieval effectiveness of heritage text data for as many metrics as is possible with image retrieval. Despite the lack of significant differences, some interesting observations were recorded for text retrieval, which suggest the need for further exploration. Figure 3.7 shows the precision and recall performance among the text retrieval pre-processing algorithms. It can be seen from the graph that the addition of stemming to the baseline algorithm (noConfig) leads to an improvement in recall and precision performance. This would suggest that stemming processing on heritage text data can potentially lead to the improvement of text retrieval effectiveness, although there are no significant differences. This is also potentially supported by the fact that every pre-processing algorithm that contained stemming processing (i.e. stemming, stemmingAndStopping, synonymsAndStemming & allConfig) had better mean recall performance compared to the baseline and all except synonymsAndStemming and allConfig had better precision and NDCG performance compared to the baseline (see Table 3.9). Conversely, the addition of thesaurus processing appears to worsen retrieval effectiveness when compared to the baseline (see Figure 3.7). Thesaurus processing only seems to be effective when it is combined with stemming processing (i.e. allConfig, synonymsAndStemming), however, that still seems to be less effective than stemming processing in isolation. This suggests that thesaurus processing has the potential to negatively impact the retrieval effectiveness of this heritage text data. This requires further inspection as to whether it is an issue with the thesaurus used for this research (i.e. the WordNet thesaurus) or thesaurus processing in general. Lastly, stopword removal appears to have no significant impact on retrieval effectiveness of this heritage text data. The stopping pre-processing algorithm recorded lower mean recall and NDCG values than the baseline but higher mean precision values (see Table 3.9). Figure 3.7 visualizes the impact of stopword removal on retrieval effectiveness with respect to recall and precision, and it is visible that there is no clear improvement or degradation in retrieval effectiveness from the addition of stopword removal processing. This could be a result of the lack of stopwords in the queries used for the evaluation (~11%).

Table 3.5: Asymptotic Friedman results for image algorithms

	df	χ^2	Sig
Recall	5	29.708	.000
Precision	5	27.883	.000
F-Measure	5	28.915	.000
DCG@5	5	25.232	.000
DCG@10	5	35.005	.000
NDCG@5	5	25.232	.000
NDCG@10	5	35.005	.000
Number Docs Retrieved	5	25.041	.000

Table 3.6: Asymptotic Friedman results for text algorithms

	df	χ^2	Sig
Recall	7	10.865	.145
Precision	7	47.801	.000
F-Measure	7	25.585	.000
DCG@5	7	4.523	.718
DCG@10	7	7.387	.390
NDCG@5	7	4.523	.718
NDCG@10	7	7.387	.390
Number Docs Retrieved	7	34.282	.000

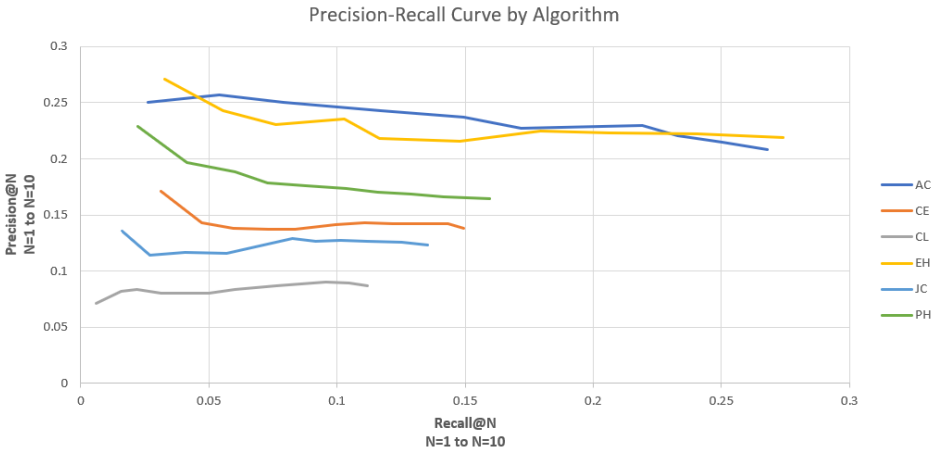

Figure 3.6: Precision-Recall curve by image algorithms

Table 3.7: Wilcoxon Signed Rank tests for pairwise comparison between image algorithm (*= statistical significance)

	Recall	Precision	F-Measure	DCG@5	DCG@10	NDCG@5	NDCG@10	Number Docs Retrieved
AC-CE	0.029*	0.085	0.046*	0.037*	0.041*	0.029*	0.021*	0.105
AC-CL	0.001*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.143
AC-EH	0.844	0.514	0.666	0.642	0.779	0.507	0.937	0.319
AC-JC	0.008*	0.028*	0.012*	0.011*	0.004*	0.018*	0.005*	0.038*
AC-PH	0.048*	0.172	0.101	0.176	0.195	0.102	0.075	0.042*
CE-CL	0.195	0.077	0.080	0.036*	0.036*	0.029*	0.041*	0.937
CE-EH	0.011*	0.051	0.012*	0.062	0.030*	0.102	0.021*	0.029*
CE-JC	0.661	0.385	0.493	0.371	0.289	0.507	0.324	0.248
CE-PH	0.594	0.438	0.437	0.215	0.325	0.367	0.498	0.004*
CL-EH	0.001*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.074
CL-JC	0.396	0.166	0.187	0.085	0.165	0.090	0.198	0.445
CL-PH	0.088	0.055	0.046*	0.011*	0.022*	0.012*	0.021*	0.011*
EH-JC	0.008*	0.028*	0.012*	0.036*	0.006*	0.064	0.006*	0.011*
EH-PH	0.017*	0.085	0.046*	0.371	0.165	0.334	0.075	0.226
JC-PH	0.415	0.381	0.363	0.084	0.154	0.232	0.216	0.002*

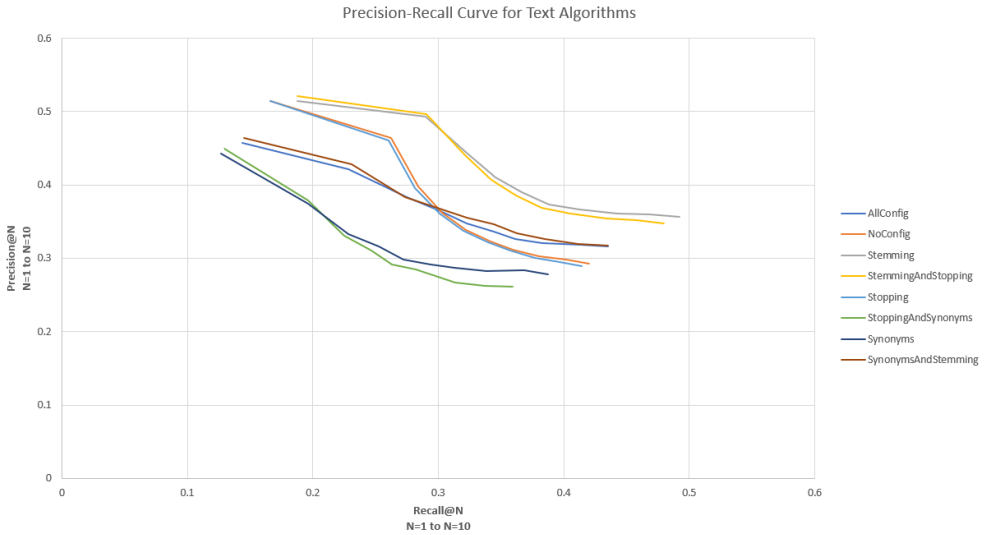

Figure 3.7: Precision-Recall curve by text algorithms

Table 3.8: Mean and standard deviation for retrieval metrics for image algorithms

	Recall		Precision		F-Measure		DCG@5		DCG@10		NDCG@5		NDCG@10		Number of Docs Retrieved	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AC	0.268	0.272	0.211	0.219	0.202	0.194	2.057	2.352	2.891	3..081	0.315	0.323	0.330	0.289	9.771	0.543
CE	0.150	0.185	0.145	0.183	0.125	0.136	1.208	1.511	1.852	2.270	0.198	0.246	0.216	0.246	9.614	0.666
CL	0.112	0.170	0.093	0.129	0.089	0.116	0.675	1.259	1.113	1.637	0.103	0.204	0.129	0.186	9.629	0.802
EH	0.274	0.261	0.221	0.219	0.214	0.193	1.950	2.259	2.966	3.015	0.284	0.311	0.337	0.299	9.857	0.427
JC	0.135	0.174	0.129	0.164	0.116	0.123	1.077	1.352	1.618	2.042	0.177	0.229	0.188	0.222	9.529	0.756
PH	0.159	0.166	0.165	0.207	0.147	0.163	1.576	1.920	2.281	2.769	0.222	0.248	0.234	0.252	9.943	0.234

Table 3.9: Mean and standard deviation for retrieval metrics for text algorithms

	Recall		Precision		F-Measure		DCG@5		DCG@10		NDCG@5		NDCG@10		Number of Docs Retrieved	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AllConfig	0.619	0.347	0.338	0.337	0.369	0.318	3.157	2.809	4.469	4.045	0.562	0.390	0.574	0.359	17.629	4.694
NoConfig	0.558	0.384	0.389	0.335	0.387	0.316	3.210	2.685	4.346	4.014	0.589	0.386	0.591	0.379	14.357	7.629
Stemming	0.684	0.338	0.416	0.335	0.438	0.307	3.570	2.689	5.051	4.094	0.646	0.359	0.667	0.343	16.143	6.427
StemmingAndStopping	0.649	0.343	0.442	0.348	0.440	0.314	3.553	2.703	4.976	4.100	0.642	0.363	0.657	0.349	15.214	7.013
Stopping	0.539	0.382	0.421	0.347	0.391	0.313	3.197	2.680	4.309	4.016	0.586	0.386	0.585	0.380	13.414	7.972
StoppingAndSynonyms	0.505	0.375	0.319	0.326	0.321	0.296	2.747	2.447	3.813	3.614	0.499	0.382	0.504	0.368	16.243	6.001
Synonyms	0.521	0.376	0.314	0.316	0.326	0.296	2.781	2.424	3.966	3.624	0.505	0.381	0.524	0.368	16.457	5.702
SynonymsAndStemming	0.621	0.336	0.341	0.340	0.372	0.311	3.220	2.765	4.509	4.022	0.568	0.375	0.577	0.351	18.043	3.876

3.9 Summary

This chapter detailed the methodology followed for an experiment that provided an offline evaluation of pre-processing algorithms for text and image retrieval. The chapter first described origins of the Five Hundred Year Archive (FHYA) project, as well as the subset of the FHYA dataset that was provided by the APC for use in the experiment and consisted of 1345 text documents and 5708 images containing precolonial African heritage data. This chapter then outlined the process followed to verify the relevance of information needs, which were used to collect a total of 70 sample text-based queries for use in the experiment. The methodology followed to crawl 70 sample image-based queries from the Smithsonian Institution National Museum of African Art[15] was also described. This chapter then outlined the creation of a multimodal information retrieval system using Apache Solr and LIRE, which made use of multiple cores, indexed with the abovementioned subset of the FHYA dataset, to allow for algorithmic variation. The pre-experiment for optimizing the LIRE accuracy parameter for each of the 6 image retrieval pre-processing algorithms was also described. The procedure for the formal experiment was then detailed in this chapter, where supersets of results for the 70 text-based and 70 image-based queries were obtained, and the relevance judgements of the results were determined by 10 participants. These relevance judgements were lastly used to calculate various retrieval metrics across all of the algorithms and were presented at the end of the chapter. These results showed that significant differences can be found between the retrieval effectiveness of image retrieval pre-processing algorithms across recall, precision, F-measure, DCG, NDCG and number of documents retrieved. It was found that the shape-based image retrieval pre-processing algorithms performed better for this dataset than colour-based pre-processing algorithms, with Edge Histogram being the top performing image retrieval pre-processing algorithm. Despite these findings for image retrieval algorithms, significant differences were not recorded for text retrieval pre-processing algorithms. Despite this, the results do suggest the benefit of stemming pre-processing for textual retrieval of this data, whereas retrieval is worsened by the addition of synonym processing.

[15] https://africa.si.edu/collections/collections

4. Experiment 2: Investigating user experience and bias mitigation

4.1 Overview

This experiment served the purpose of investigating the user experience with a retrieval system that allowed users to compare pre-processing algorithms, query methods, and various result formats. This entailed developing a user interface for the multimodal information retrieval system, developed in Experiment 1, and asking users to perform multiple tasks while continuously providing feedback via a survey. The below sections outline the process followed to develop the system required for use in the experiment, demographic statistics about the participants in the experiment, as well as the procedure participants were asked to follow during the experiment. The counterbalancing measures taken to prevent any order effects are also detailed below. Lastly, this chapter is concluded with the findings from the experiment.

4.2 System Design and Implementation

While Experiment 1 required the development of a multimodal information retrieval system that supported algorithmic variation, Experiment 2 required a multimodal information retrieval system that provided a user interface that allowed a user to be able to explicitly make changes to the pre-processing algorithms used, query methods and result formats. This section details the design and implementation of a user interface for the multimodal information retrieval system developed for Experiment 1 (containing the FHYA document collection) that is able to facilitate these variations. The tasks that were selected for use in the experiment are also described below, as well as how the information retrieval system was hosted on the Internet for accessibility to participants.

4.2.1 *Algorithms*

This experiment investigated the effects of algorithmic variation of text and image retrieval algorithms. For text algorithms, users were asked to compare the following 3 algorithms (explained in previous sections): stemming, stopping, and synonyms (i.e. a thesauri). For image retrieval algorithms, the 3 best-performing algorithms from Experiment 1 were selected for comparison. These were: Edge Histogram (Sun Won, Kwon Park, & Park, 2002), Autocolour Correlogram (Huang, et al., 1997) and Pyramid Histogram of Oriented Gradients (Bosch, Zisserman, & Munoz, 2007). Users were all designated a fixed text and image algorithm pair (explained in the research block section 4.3.3) and were asked to compare the results of these algorithms for the same query tasks.

4.2.2 *Experimental Tasks*

While preparing for Experiment 1, 70 sample text- and image-based queries were respectively obtained and verified by experts and other participants. Experiment 2 required a pair of sample queries for each of the formats (text and image) to be used as tasks during the experiment. Thus, 2 images and 2 text queries were randomly sampled from the pool of sample queries obtained during Experiment 1. These queries can be found in Table 4.1 and were used as tasks by participants during the experimental process outlined in the coming sections. As can be

seen in Table 4.1, the two image tasks consist of an image of beadwork and an image of a wooden object. Going forward, this paper will refer to these two tasks as the "bead image task" and the "wooden image task".

Table 4.1: Randomly sampled queries used as tasks for Experiment 2

Format	Query
Text	Beaded necklaces
Text	Praise and worship
Image	 [16]
Image	 [17]

4.2.3 ***User Interface***

In the web interface of the retrieval system, used for the user experiment, users were able to conduct searches using both text- and image-based queries, specify the pre-processing algorithm for relevance judgement and view results as text only, images only or a hybrid approach containing both text and images for results. This configurability is the addition of user-control to the retrieval system. The multimodal query format and result format support was added with the thinking that it would assist users with overcoming misattributed metadata, and with the potential to present more diverse information and results to users. The support for alternating among pre-processing algorithms can potentially also assist users with being able to find data that is relevant to queries from different perspectives. This interface was developed using JavaScript and HTML. Upon first opening of the interface, users are shown their first two tasks; this can be seen in Figure 4.1. Participants can either be shown pane (A), which presents the text tasks, or pane (B), which shows the image tasks. This was dictated by the research block the user was allocated to (to be explained later). Figure 4.1 shows that users are restricted to only selecting the next available task (i.e. if they have not completed task 1, task 2 is greyed out and is labeled unavailable). Figure 4.1 is also a capture of the entire interface upon first entry. The elements of the interface are as follows:

(1): Instruction window that was included to guide the user throughout the experiment. This window has a blue glow effect that animates whenever the current instruction is updated.

(2): The aforementioned task pane.

[16] https://africa.si.edu/collections/collections
[17] https://africa.si.edu/collections/collections

(3): The feedback window that displays the user's feedback as they provide it to the system during the experiment (this will be discussed later).

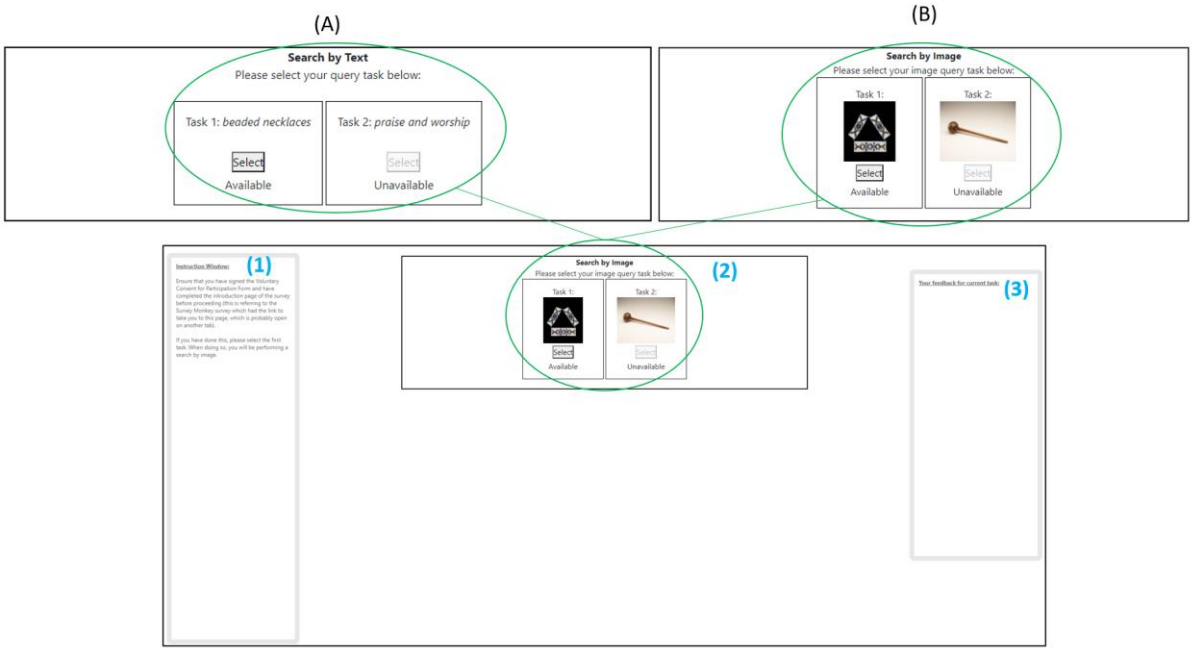

Figure 4.1: The opening screen of the interface where the user has been allocated task groups (A) or (B)

Upon selection of a task, the query associated with the task is asynchronously submitted, using jQuery and async await, to the default algorithm assigned to the user in the Solr-based information retrieval system. When the asynchronous response that contained the retrieved results is received, the result panes are then shown. This can be seen in Figure 4.2, which shows the following elements:

(1): The query for the current task.

(2): The options for dynamically changing result formats.

(3): The current result format.

(4): A column of the results, referred to in this study as the "fixed column", for the query task retrieved using the pre-processing algorithm named at (5)

(5): The designated pre-processing algorithm name. Note that for ease of understanding for users, the interface referred to pre-processing algorithms as "configurations", and pre-processing algorithms were displayed using a descriptive name (see Table 3.2).

(6): An empty column of results, referred to in this study as the "dynamic column". This column can be dynamically populated with results by selecting a new pre-processing algorithm from (7).

(7): Dropdown menu containing alternative pre-processing algorithms to populate results in the dynamic column.

(8): A link to a page explaining the behavior of the different pre-processing algorithms.

(9): The instruction in the instruction window has updated and the window is glowing to indicate this to the user.

Figure 4.2 also shows how users are able to select an algorithm to retrieve results and populate them in the dynamic column. The order with which pre-processing algorithms in the dynamic column are selected and

displayed is controlled and alternated, as can be seen in Figure 4.2, where an algorithm is greyed out because the user needs to view the results using another algorithm first. This was done to prevent order effects from constantly viewing results of pre-processing algorithms in a particular order. For the next task, the user would be presented with the same two pre-processing algorithms in the dropdown menu, but in the reverse order. The instruction window is also currently explaining to the user how and why results have been populated in the fixed column and that they need to select an algorithm from the dropdown menu to populate results in the dynamic column.

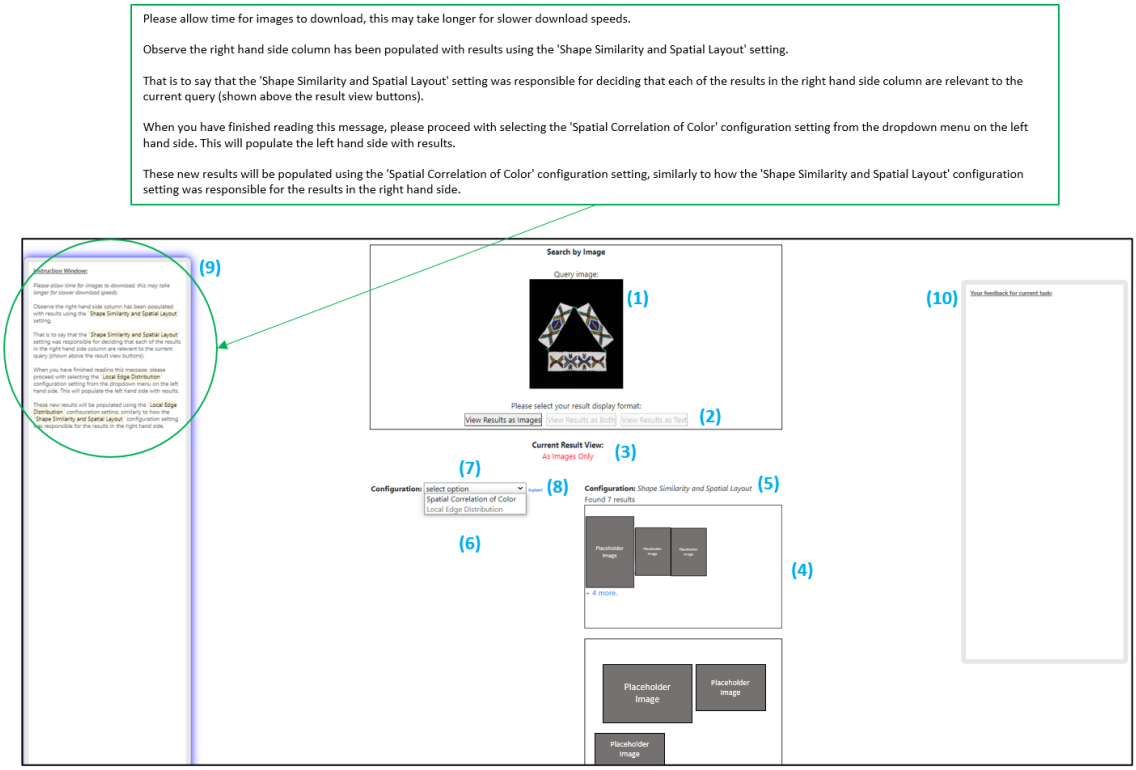

Figure 4.2: The interface upon initial showing of results and the dropdown menu for algorithmic variation

Upon selection of a pre-processing algorithm from the dropdown menu, the empty dynamic column is populated with results for the same query task, except for them being retrieved by a different pre-processing algorithm to that used for the fixed column (i.e. it will now use the pre-processing algorithm selected from the dropdown menu). This can be seen in Figure 4.3. This fixed vs dynamic column design was deliberately implemented to provide a simple mechanism for users to compare two potentially different sets of results that were retrieved by two different pre-processing algorithms. Having the results of two pre-processing algorithms side-by-side prevents the need for users to have to memorize results from previous pre-processing algorithms and instead focus directly on the difference between the two columns. This is also being explained to the user via the instruction window, which is telling the user to “Take note of differences in the results in the left and right hand side columns” and “If the two columns contain different results, take note of which column contains results you prefer relative to the query. You will be asked shortly to state which of the 3 settings produced the set of results that you prefer the most and prefer the least”. Any reference to a pre-processing algorithm in the instruction window is also highlighted for ease of visibility. The designated pre-processing algorithm for the fixed column remained consistent for the duration of a user’s session with the system, and the reason for this is specified in the section

about research blocks (section 4.3.1). Given that 3 algorithms for each modality were being compared in this experiment, and the fixed column had the designated pre-processing algorithm, the dropdown menu for the dynamic column always contained the other two pre-processing algorithms for the user to select.

Figure 4.3: The interface upon showing of dynamic results

When the results for both of the algorithms in the dropdown menu have been viewed, the system freely allows for alternation of these algorithms so that users are able to confidently form an opinion of the algorithms they prefer the most and prefer the least. This is indicated to them in the instruction window (see (2) in Figure 4.4). At this point, note that results have only so far been viewed in one single format. The system prompts users, upon deciding their most and least preferred pre-processing algorithms, to provide their feedback by clicking the button at the bottom right corner of the interface (see (1) in Figure 4.4). This button only appears when the system has detected that users have viewed the results for all pre-processing algorithms, following which, an alert dialogue is shown to the user asking them to provide their most preferred pre-processing algorithm (by typing in either A, B, or C) (see Figure 4.5). Verification was done on this input to ensure it was correctly entered. An identical dialogue is shown to the user immediately afterwards asking the user to provide their least preferred algorithm as well. Both dialogs present a confirmation dialogue to the user, allowing them to change their entry before it is recorded to prevent accidental feedback.

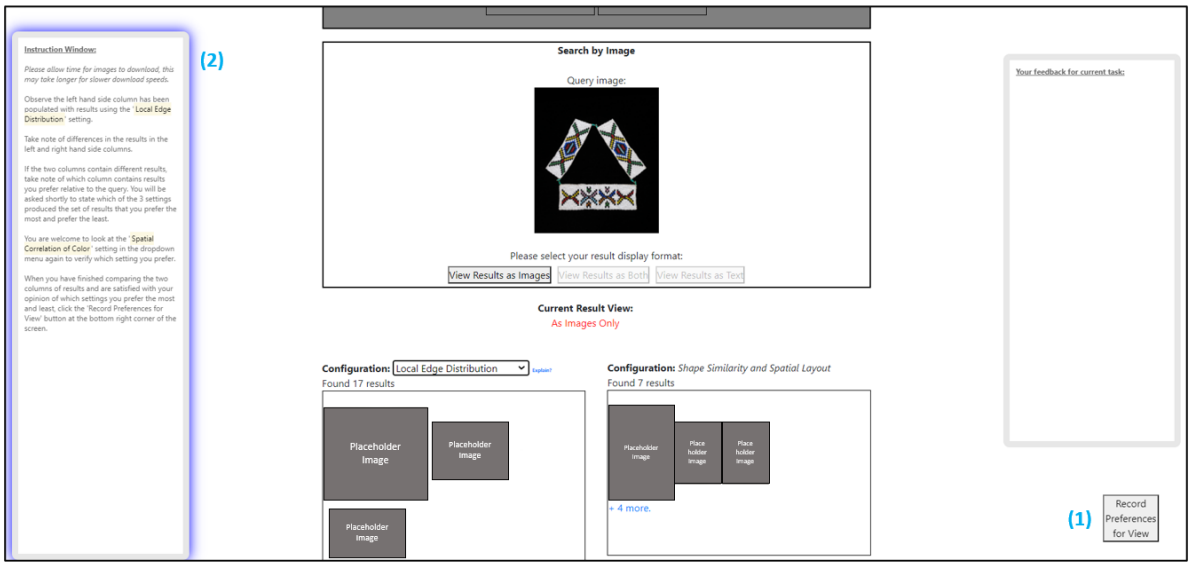

Figure 4.4: Interface upon completion of viewing all algorithms for one result view

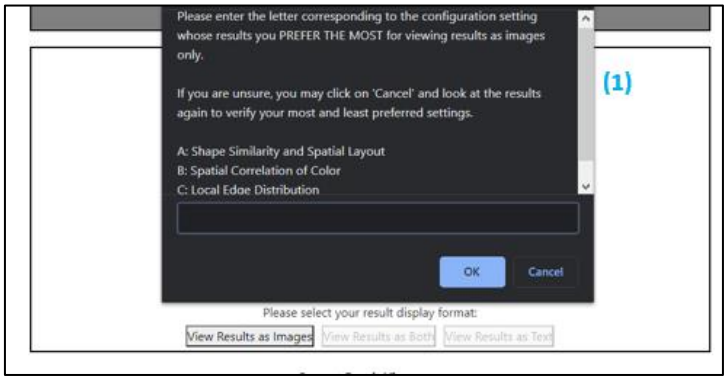

Figure 4.5: Alert prompting users' feedback regarding algorithm preferences

When the user has provided their feedback regarding their most and least preferred pre-processing algorithm, the system generates a list representing their ranking of pre-processing algorithms (for the current result format) and displays it to the user in the feedback window (see (1) Figure 4.6). This was done so that the users were able to refer back to their preferences when answering survey questions. The initial implementation was for users to explicitly provide a ranking of 1st to 3rd, although it was found to be more efficient to ask users to simply provide their most and least preferred algorithms and to derive a ranking from this. At this point, a new result format is also enabled (see (2) in Figure 4.6), allowing the user to now compare results across pre-processing algorithms for this task in a different result format. Upon selecting the new result format, the previous one is disabled, and all results are displayed using the new format (see (1) in Figure 4.7).

Figure 4.6: User feedback of algorithm ranking and enabling of new result format

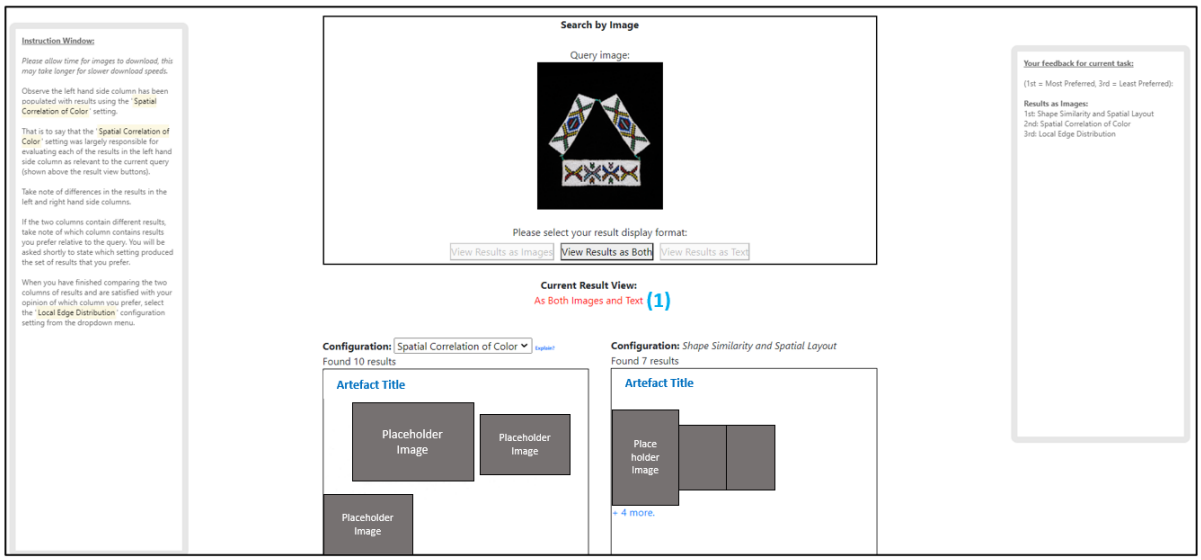

Figure 4.7: Change in result format

Upon progression into a new result format, users now repeat the process of selecting pre-processing algorithms from the dropdown menu, comparing results, and providing feedback for their most and least preferred pre-processing algorithms for each of the 2 remaining result formats. This is done to assess whether users prefer different pre-processing algorithms when the result format changes, to determine if result format has any effect on the perceived retrieval effectiveness of different pre-processing algorithms. When they have completed providing the feedback for all result formats, they are shown the interface as seen in Figure 4.8 As can be seen at (1), feedback has been recorded for all of the different result formats. There is now the option for them to unlock all of the different result formats using the button at (2). When this is done, all of the result format options at (3) become enabled, allowing the user to freely alternate between result formats and to also change pre-processing

algorithms. The instruction window, at this point, prompts them to do this and compare result formats so that they can now provide feedback regarding which is their most and least preferred result format (see Figure 4.9).

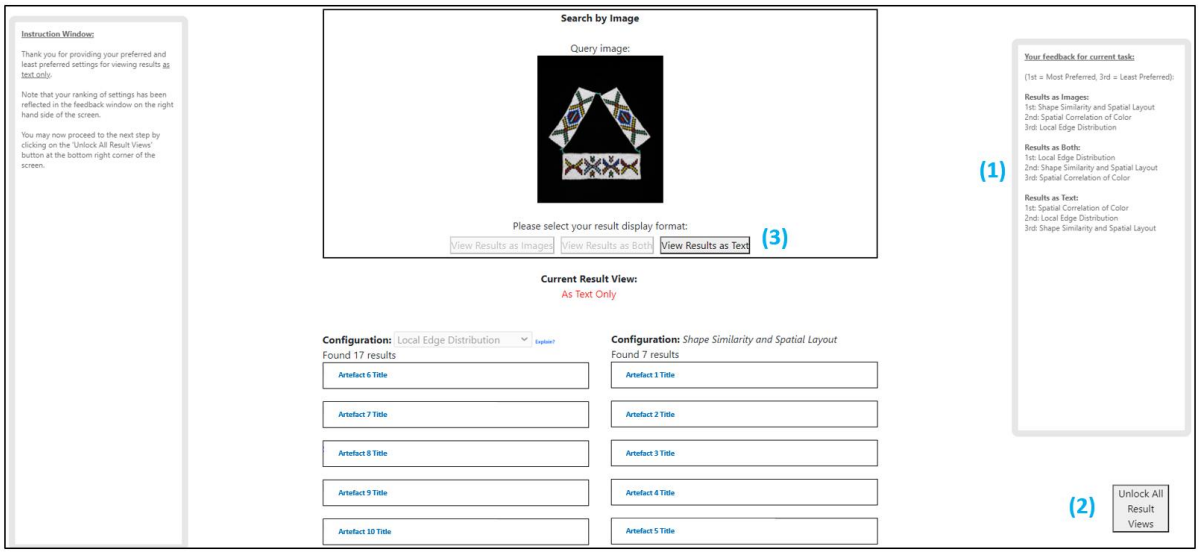

Figure 4.8: All result format feedback recorded

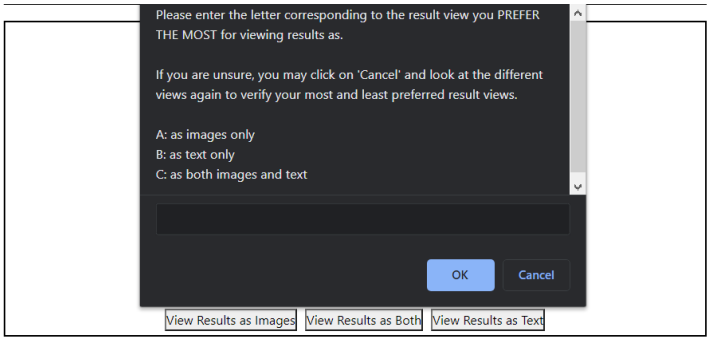

Figure 4.9: Alert prompting users' feedback regarding result format preferences

At this point, the user has now successfully viewed results for the same query, across all pre-processing algorithms, and for every supported result format. They have also recorded their feedback regarding their algorithmic preferences for each result format, as well as their preferences for result formats. This can be seen at (1) in the Figure 4.10. The user has now concluded their first task and can progress to the next task by clicking the button at (2). Before this, the instruction window instructs the user to proceed with answering the survey questions for this task before proceeding to the next task. Upon selection of (2), an alert dialogue is also presented to the user to verify whether they have completed all of the necessary steps in the task and have answered the survey questions for the task. The process is repeated for the second task (that has the same query format as the first task), following which they will repeat the process for tasks 3 and 4, which have a query of the opposite format to the queries in the first two tasks.

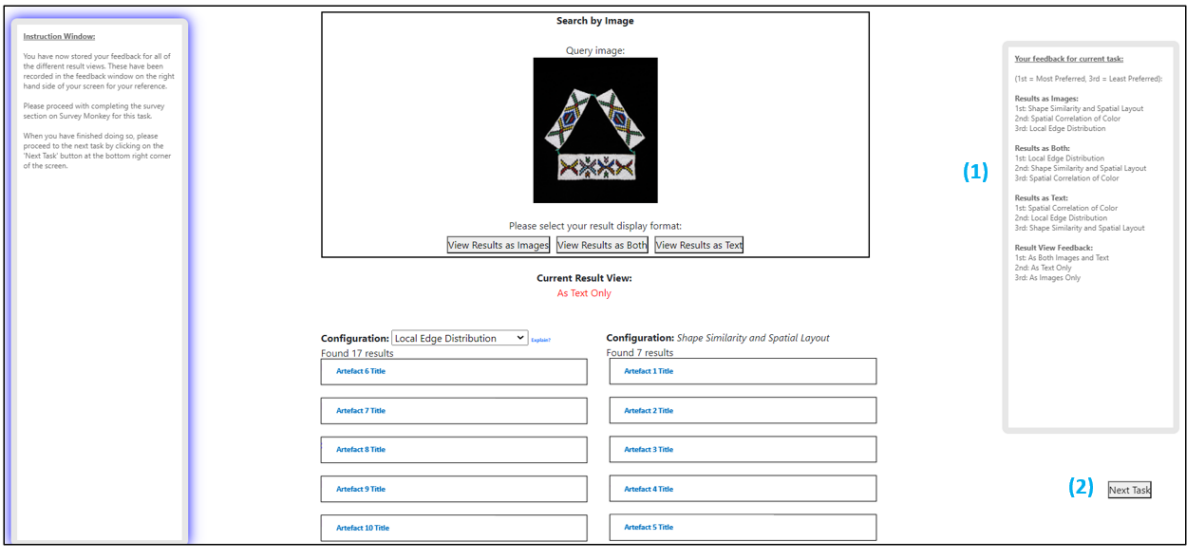

Figure 4.10: All algorithmic and result format preferences recorded

Upon completion of all 4 tasks, participants are shown a confirmation dialogue verifying that they have answered the survey questions for the final task, that they have answered the survey questions in the conclusion section, and that they have submitted their survey (see Figure 4.11).

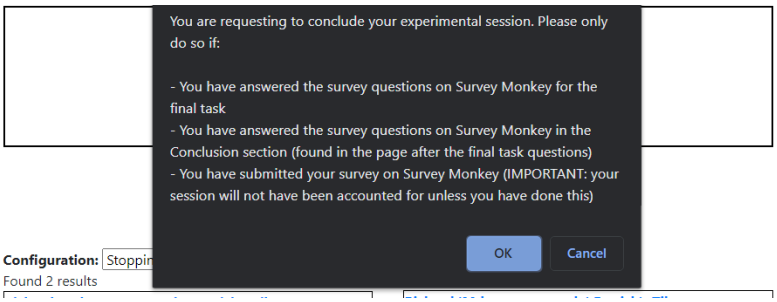

Figure 4.11: Dialogue verifying users have completed and submitted the survey

Another feature of the system included the ability for users to view individual results in greater detail. They were free to do this in order to acquire more information about a result retrieved by an algorithm to determine whether it is relevant. This is done by clicking the tile of a single result entry and displays associated metadata for that entry, as well as all images available for that result (see Figure 4.12).

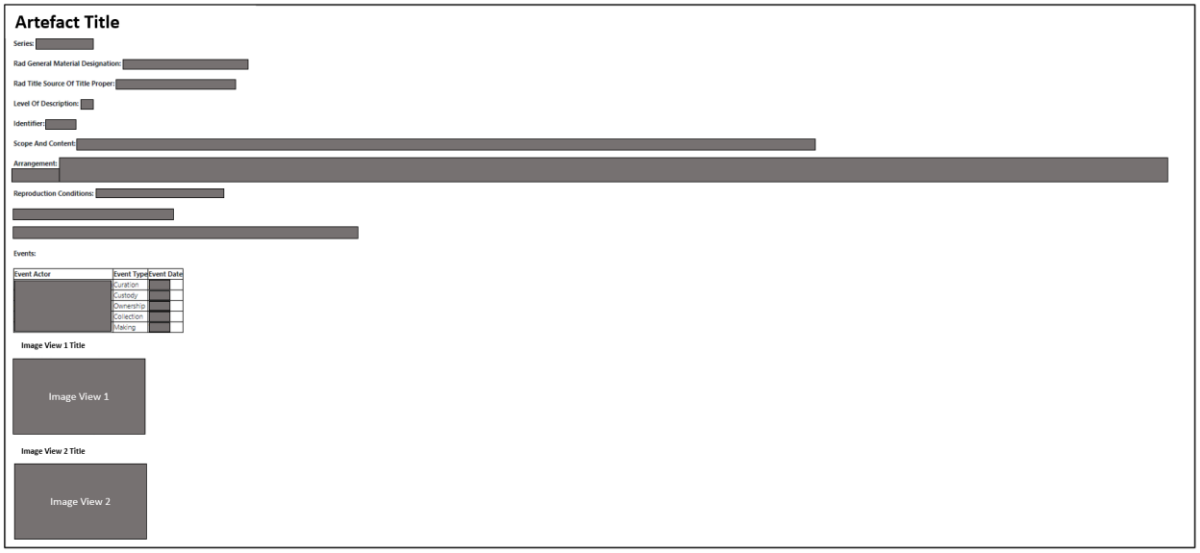

Figure 4.12: Detailed result page for a retrieved artefact (anonymised)

Lastly, two "Explain?" pages (for text and image respectively) were available for users to provide a more granular explanation of the pre-processing algorithms they were being asked to compare. Users were able to navigate to this by selecting the button at (7) in Figure 4.2. These explanation pages can be found in Appendix B.

4.2.4 *Hosting*

The Solr-based multimodal information retrieval system and user interface needed to be hosted in a location where it was freely and easily accessible to potential participants. Access was provided to a server belonging to the department of Computer Science at the University of Cape Town to host the system. The entire Solr system, associated image data files from FHYA, and the HTML pages needed for the user interface, were all stored in a directory in the server. The Solr system was started on port 9129, and URL rewriting was used to easily direct all requests to the Solr server (from the UI) to the path: http://pumbaa.cs.uct.ac.za/1/solr/#/. This made access to the system simple, as participants only needed to be provided a link to the experiment.html page (the UI) located on the server to participate in the experiment.

4.3 Experiment Design

4.3.1 *Pre-experiment procedure*

Participants were all sent an email, derived from a standard template (see Appendix C), which outlined the details regarding what participation in this experiment entailed. Participants were advised in this email that the experiment had been designed to be completed within an hour, although 1½ hours should be allocated for safety. The email then explained to participants that all links provided to them for the experiment should be opened in Google Chrome, as support for the experimental system was only tested for Google Chrome. The email provided the participants with 3 links: a link to the survey, a link to the introductory video (located on YouTube), and a link to the search engine to be used for the experiment. Participants were asked in the email to keep all tabs open at all

times for the duration of the experiment and to ensure that the experiment was completed in its entirety in a single session. Lastly, the email contained the due date for completion of the experiment for the respective participant. Participants were typically provided a 2-week window for completion, although extensions were granted under abnormal circumstances (18 participants were provided extensions beyond the 2-week window).

4.3.2 ***Experiment procedure***

Upon opening of the survey, participants were first asked to read, complete, and sign the Informed Voluntary Consent to Participate in Research Study form (see Appendix D). Participants were then shown an introductory video that was uploaded to YouTube (https://www.youtube.com/watch?v=K-zD6B7SW3M&feature=youtu.be) (and embedded in the survey). The introductory video was created and used in order to standardise the information that was provided to every participant to mitigate the effects of the researcher biasing individual participants differently. The introductory video reminded participants to read and sign the consent form in the survey, explained what participants will be asked to do for the experiment, the aims of the research, and showed an example of how to complete a task in the system whilst completing the survey simultaneously. Most importantly, the introductory video explained the context of bias in information retrieval to the participant. The use of a referrable definition of bias was avoided to prevent influencing the users' perception of the concept of bias. Instead, an example of bias was illustrated to participants in the example video and participants were left to form their own interpretation of bias from this. The illustration of bias in the introductory video was the following:

Let's say there is a person named Jane. Jane's job is to write about movies on the Internet. Jane writes about a large variety of movies of all genres, but Jane's favourite genre of movie is Action. Whenever Jane writes about Action movies, Jane describes them with positive words like 'exciting', 'fun', 'enjoyable'. Jane also happens to hate Comedies. When Jane has to write about a Comedy, she usually describes it as 'boring' or 'annoying'. If someone was to search through Jane's movie write-ups online, it would be more likely that they see Jane's Action write ups, rather than Jane's comedy write-ups, due to how Jane describes Action films more positively than Comedies. In other words, Jane has allowed her bias of liking Action films over Comedy films, to affect how likely it is for Comedy film write-ups to be found in the search engine. This is an illustration of how bias in a search engine exists and can be applied to many types of information far beyond the scope of movie write-ups.

Participants were then asked to complete the introductory page of the survey (see Appendix E), which acquired general demographic information such as their age, self-reported gender, whether or not they believe there is any bias in the results shown to them by search engines and lastly, whether they believe that by changing the settings of a search engine, they can reduce the amount of bias in the search engine. Following this, participants were asked to complete 4 of the tasks described in the user interface description in 4.2.3 (2 tasks for searching via text query and 2 tasks for searching via image query). For each task, participants had to perform the following sequence of actions:

1. Select the task.
2. Observe results for the query for the task have been populated in the fixed column using their designated pre-processing algorithm.

3. Select the first available pre-processing algorithm from the dropdown menu to populate results for the same query (using the new pre-processing algorithm) in the dynamic column.
4. Take note of which column contains results they prefer relative to the query and what any major differences between the two columns are.
5. Select the second available pre-processing algorithm from the dropdown menu to populate results for the same query (using the new pre-processing algorithm) in the dynamic column.
6. Take note of which column contains results they prefer relative to the query and what any major differences between the two columns are.
7. Record their most and least preferred pre-processing algorithms, for the current task and current result format.
8. Switch to the new result format.
9. Repeat steps 3-8.
10. Record their most and least preferred result format for the current task.
11. Answer the survey questions for the task

Participants were guided through these steps by the instruction window in the system, which regularly updated to inform them what is expected of them next. Their feedback regarding algorithmic and result format preferences were also stored and displayed in the feedback window for easy referral when completing the survey. In step 11, the participants were asked questions in the survey about their experience with the task (see Appendix E). This included asking them to provide their ranking of pre-processing algorithms for the task, which pre-processing algorithms they found to be of value, which pre-processing algorithms they believed retrieved biased results, their ranking of result formats, and which result formats they believe to have biased presentation of results.

Upon completion of all 4 tasks, participants had to lastly complete the conclusion section of the survey (see Appendix E). This asked participants about their preferred methods of search (i.e. via a text-based query or an image-based query), whether they believed the different methods of search contained bias, and asked them once again whether they believed that changing the settings of a search engine can reduce bias (in order to determine if the experiment had changed their response from the first answering of the question). Lastly, participants were asked to provide suggestions for improvement or general comments.

4.3.3 ***Research blocks and counterbalancing***

This experiment followed a repeated measures design, where participants were asked to compare all pre-processing algorithms, query formats, and result formats. For each search method (i.e. search by text query and search by image query), there were 3 pre-processing algorithms, and participants were each designated a text- and image-based pre-processing algorithm (that populated results in the fixed column) to compare against the pre-processing algorithms in the dynamic column. This meant that there were 9 possible pre-processing algorithmic pairs (3×3) that participants could be assigned to. These algorithmic pairs could've been confounding factors in the outcome of the participant's responses regarding preferences and perceived bias. As a result of this, a blocked randomization experimental design was adopted, whereby the experiment was blocked on unique pre-processing algorithm pairs. This would allow for the control of variance in results due to the potential interaction effects from

various designated pre-processing algorithmic pairs (i.e. a certain pair may bias participant responses more than others and any frequent use of that pair will have an effect on the results of the experiment). Each of these 9 blocks contained 2 different treatments, these being whether they conducted searches via text first, or whether they conducted searches via image first. This was to mitigate any order effect from searching by one method before the other. Thus, exactly half of participants would have done the experiment by searching with text first, whilst the other half would have done the experiment by searching with images first (i.e. full counterbalancing). This meant a total of 18 treatments (i.e. 2 treatments × 9 research blocks). Thus, to conform to the generalised randomised block design (which requires a minimum of 3 allocations per treatment per block), the minimum number of participants required for this design was 54 (3 participants × 18 treatments). Figure 4.13 illustrates the structure of this blocked experimental design.

Other factors that could have influenced the results from order effects or other potential biases include the order of execution of the tasks, the order of viewing of result formats, and the location (i.e. left or right) of the dynamic and fixed columns. Given that there were 2 tasks per search method, this was fully counterbalanced, such that half of the participants performed searches with the tasks in one order (per query format) and the other half of participants did the reverse order for both respective query formats. The order of viewing the 3 result formats was also fully counterbalanced across participants. There were 6 unique orders to view result formats in, and each participant was assigned 1 of the 6 unique sequences (and each of the sequences were assigned to an equal number of participants). Lastly, the positions of the fixed and dynamic columns were also fully counterbalanced, such that half of the participants started the experiment with the fixed column on the left, while the other half of the participants started the experiment with the fixed column on the right.

To support this counterbalancing, query strings were used in the URLs of the system, provided to participants, that the system used to determine the appropriate experience to provide the participant. The query string was able to provide the system the following information: the designated text pre-processing algorithm, the designated image pre-processing algorithm, the first search method, the order to view result formats in, the order to view text tasks in, the order to view image tasks in, and the starting position of the dynamic column. An example of this can be seen in the URL in Figure 4.14. Given that participants had to answer survey questions after each task, this also meant that there needed to be different versions of the survey that corresponded with the task orders. Thus, there were 18 different surveys (i.e. one for each research block treatment) and participants were sent the link to the corresponding survey for the research block treatment they were assigned to.

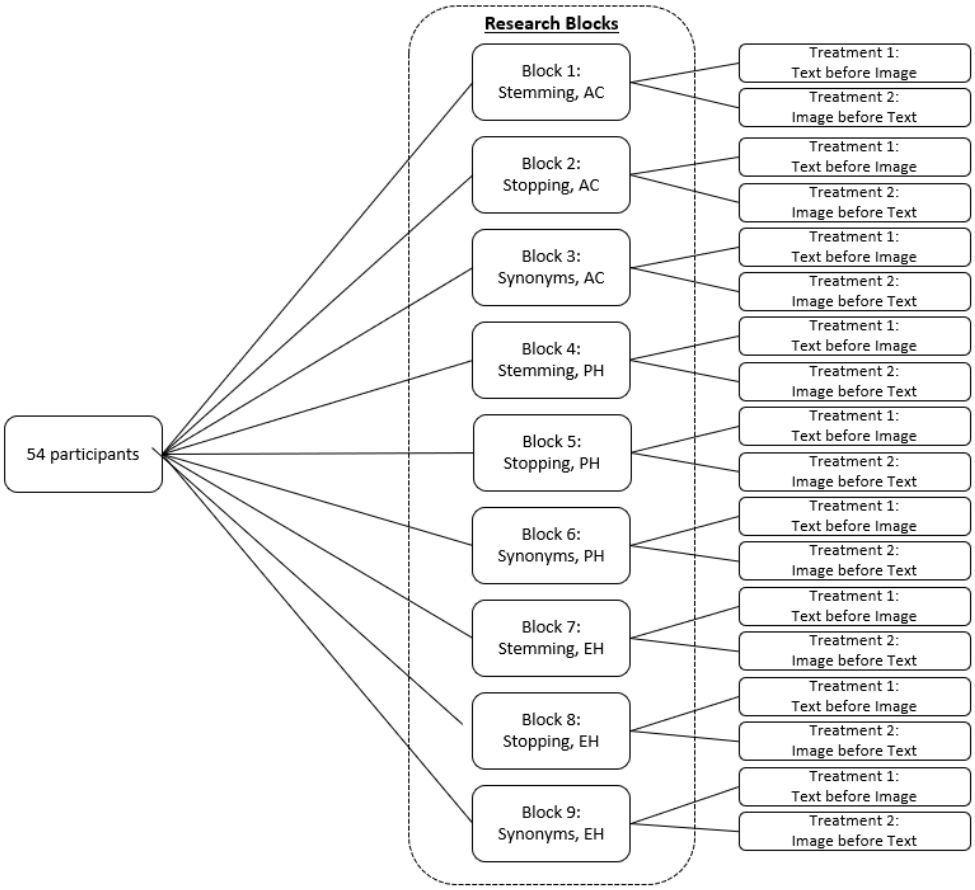

Figure 4.13: Research design blocked on 9 algorithmic pairs with 2 treatments per block

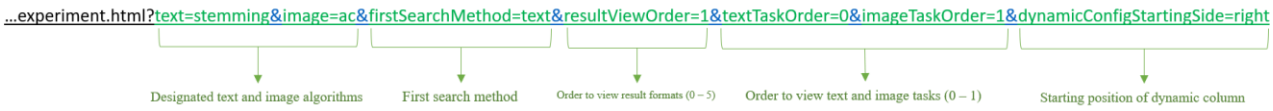

Figure 4.14: System URL with query strings to support blocked research design

4.3.4 *Pilot study*

Prior to conducting the formal experiment, two pilot rounds were conducted to ensure the process was feasible and would be able to provide the data required to conduct proper analysis. The first round was distributed to 4 participants. 2 of the participants completed the entire experiment and provided feedback, 1 participant completed the experiment and provided feedback but forgot to submit their survey, and the final participant did not participate. At this point, a major difference with the interface was that it did not contain a feedback dialog to record user preferences. The introductory video also had a very technical, formal definition of bias in information retrieval systems and was uploaded to Vimeo. The position of the fixed and dynamic columns were also regularly alternated between tasks. Feedback from the initial round indicated that participants needed a better example of bias in the introductory video, had technical difficulties with the embedded Vimeo video, struggled to recall their preferences, found it confusing that the columns kept alternating, and felt the experiment took too long (the

average time of completion was close to 3 hours). Participants felt that with clearer instructions, minimizing time to recall preferences, and a clearer definition of bias, the experiment would be shorter and easier to complete. To address these issues, the feedback dialog was added to the interface, the illustration of bias was added to the introductory video, and the instruction window provided participants with far more detailed instructions to minimize the time they would need to spend thinking about what to do next. A second round of the pilot study was then conducted with 3 participants, all of whom completed the experiment successfully in under 1½ hours and did not raise any issues with the design or process.

4.4 Participants

This section describes the processes followed to recruit participants for participation in the pilot study for the experiment, as well as for the experiment itself. Demographic information regarding participants is also provided.

4.4.1 *Pilot study*

For the first round of the pilot study, 4 participants were recruited using the Judgement Sampling method. This was done intentionally to choose participants who have a demonstrated history of conducting university-level research, preferably in the field of Computer Science, so that they could draw on their past experiences to provide valuable feedback for improving the experiment. Of the 4 participants, only 3 completed the experiment and provided feedback. The 3 participants were 1 male and 2 females, all between the ages of 25-38, with two of the participants actively working towards completing their PhD in Computer Science at the University of Cape Town, while the 3rd participant was completing their MSc in a Humanities-related degree at the University of KwaZulu-Natal.

For the second round of the pilot study, 3 participants were recruited using the Convenience Sampling method. The aim of this round was to determine whether participants would have the same problems as those raised by the participants for the first round, thus there was no longer a specific need for participants with experience with university-level research. The 3 participants that were recruited were 2 males and 1 female and all individuals were between the ages of 23 and 32.

4.4.2 *Formal experiment*

As mentioned in section 4.3.3, a minimum of 54 participants were required to conform to the generalised randomised block design. To obtain this number of participants, both the Convenience and the Snowball sampling methods were used. An email invitation was sent to the Department of Computer Science at the University of Cape Town's Grads mailing list, as well as to the UCT research invitation mailing list (after obtaining permission), inviting participants over the age of 18 with Internet-access. The researcher also sent invitations out through a variety of social media platforms in an attempt to recruit as many participants as possible (to meet the minimum of 54 participants). Those who responded positively to the invitations were then sent an instruction email (see Appendix C) providing them with the necessary information and links to participate, as well as a due date for completion. Participants were also asked to spread the word about the research and extend the invitation to anyone

they felt would be interested in participating. A total of 54 participant responses were recorded for this experiment, and the demographic statistics for the experiment can be seen Table 4.2. 29 of the 54 recruited participants were university students (at the time), while the remaining 25 had completed university and were employed.

Table 4.2: Demographic Statistics for the formal round of Experiment 2

Variables	Levels	Frequency	Percentage
Age	<= 25	33	61.11%
	26-30	14	25.93%
	31-35	3	5.56%
	>= 36	4	7.41%
Gender	Male	33	61.11%
	Female	21	38.89%

4.5 Data Analysis

User experience and perception of bias mitigation was modeled through an analysis of the participants' responses in the user experiment survey. Participants were asked to identify pre-processing algorithms, result formats and query formats that they perceived to contain bias in their results or that they found to be of value in the retrieval system. This allowed for inspection of the different phases of retrieval to identify which can be considered contributors of bias from the perspective of the participants. The survey that participants were asked to complete can be seen in Appendix E. The results of the analysis of all participant responses are presented in section 4.6.

4.6 Findings

For all pairwise comparisons of user preferences for pre-processing algorithms and result formats, listed in this section, these results were determined from the explicit rankings of pre-processing algorithms and result formats that participants provided in response to the survey questions (see Appendix E). For example, if a participant ranked the following algorithms (in order from most preferred to least preferred): stemming, synonyms, stopping, then this research implicitly deduces that stemming is preferred over synonyms and stopping, and that synonyms is preferred only over stopping. From these rankings, pairwise preferences of pre-processing algorithms and result formats compared to others were implicitly deduced. The "Frequency of preferences over other algorithm" values (in Figure 4.15 for example) denote the values derived from these pairwise deductions.

Figure 4.15 shows pairwise comparisons of pre-processing algorithms based on user preferences. A description of these image pre-processing algorithms can be found in Table 3.2. If we combine the preferences of image pre-processing algorithms across both image tasks, we observe that PH was preferred over AC and EH, and, that AC was preferred over EH. The overall preference for PH further supports the earlier findings in Experiment 1 that shape-based pre-processing algorithms perform better for heritage image data retrieval than colour-based pre-processing algorithms. 2 of PH's top 3 results for the "bead image task" contained beads in the retrieved images, and all of PH's top 3 results for the "wooden object task" were wooden objects in the same orientation as the

query image. The "wooden object task" demonstrated how colour-based pre-processing algorithms can perform poorly on heritage image data retrieval, where AC retrieved images of boxes, seemingly because they were similar in colour to the query image. Stemming was also unanimously preferred (if preferences are combined across both text tasks) over Synonyms and Stopping, which is consistent with the earlier findings that stemming processing can improve retrieval effectiveness compared to the baseline. The only time a pre-processing algorithm was preferred over stemming processing was thesaurus processing for the query "praise and worship". This seems reasonable given that one would expect stemming processing to be more valuable for the query "beaded necklaces" given the opportunity to stem the individual terms (i.e. bead, necklace) compared to the query where thesaurus processing was preferred (i.e. praise, worship). Interestingly, stopword removal was never preferred over any other algorithm for the query that contained the stopword ("praise and worship"). This is consistent with the earlier findings that were unable to establish concrete improved performance from stopword removal.

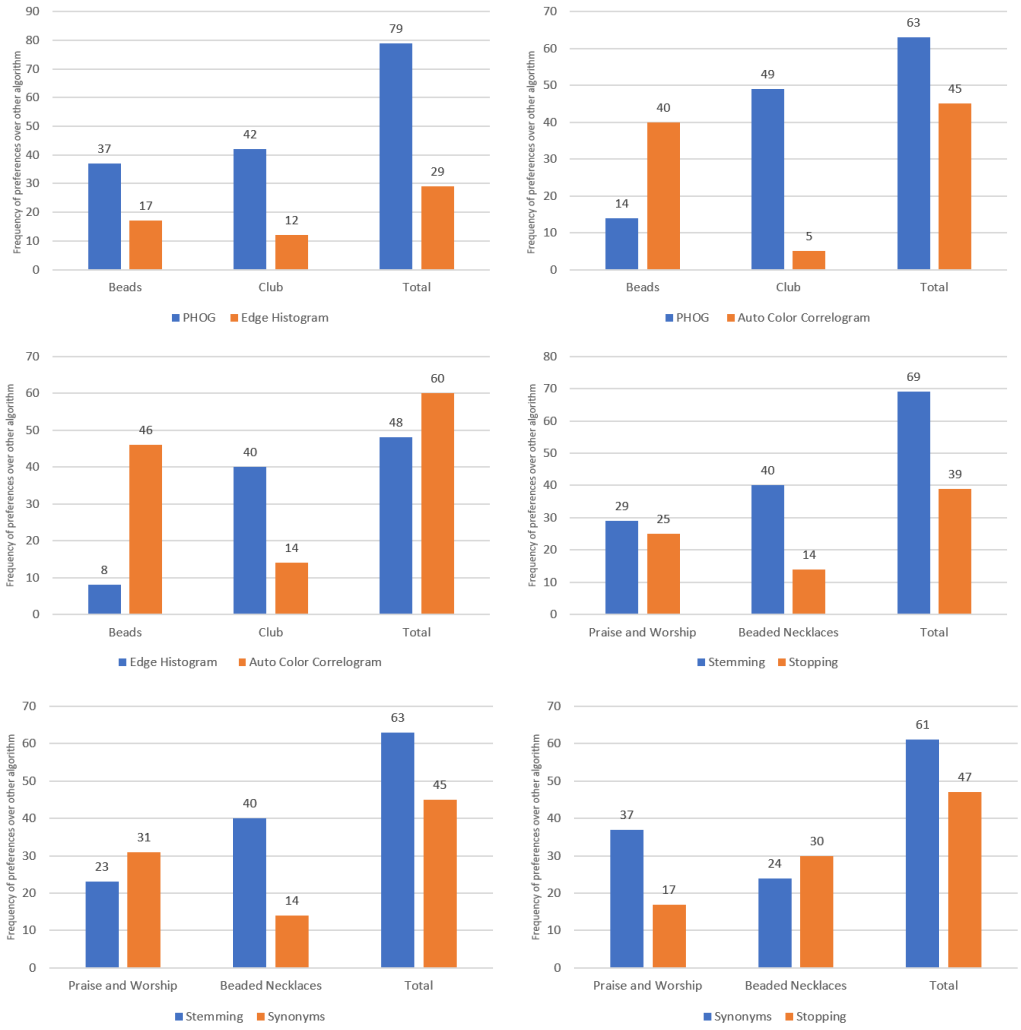

Figure 4.15: Pairwise comparisons of preferences for algorithms

Figure 4.16 shows participant preferences for result formats. It is clear to see that viewing results "As Both Images and Text" was strongly favoured over the other two result formats. The multimodal result view was preferred over viewing results "As Text Only" 92% of the time and was preferred over viewing results "As Images Only" 91%

of the time. This suggests that users prefer result formats that present a greater diversity of information with regards to heritage data and is consistent with research that highlights the importance of interface design integration and diverse results with user-control (Thomas, Billerbeck, Craswell, & White, 2019; Ruthven, 2003). When asked to explain their preference of the multimodal format, one participant responded saying that they preferred that it "provided a more informative view". Other justifications for the preference of the multimodal result format were that it "gives you a broad understanding of the results" and that unimodal result formats "creates more room for interpretation and bias". 58% of participants preferred viewing results "As Images Only" rather than "As Text Only". A participant justified their preference for images over text by saying "bias is more easily conveyed using words. Pictures are more open to interpretation and as a result [are] more difficult to contain bias". Another justification for the preference of image over text was that the participant found the text metadata to be "quite vague" and that "imagery helps to understand the text". Future research in this aspect would be informative to provide insight into this statistic, as this could potentially be as a result of inaccuracies or biases in the textual metadata, which viewing results "As Images Only" is able to overcome.

Participants were unable to unanimously express a preference for query format (see Figure 4.17). 50% of participants preferred image-based queries while 48% of participants preferred text-based queries. Some justifications for preferring image-based queries were that it "gives the reader an immediate visual to decide which results are more accurate quicker", that "you can select what you want to search for instead of trying to describe it" and "it was easier to compare the results to the object I was searching [with]". The second and third responses particularly highlight the struggle with accurately describing heritage data and how image-retrieval could potentially overcome this difficulty. Justifications for the preference of text-based queries were that "I have grown [accustomed] to searching with text" and "concepts like "praise" and "worship" in documents will be tough to query with images". The second justification is reasonable and demonstrates a scenario where image-based querying may fail in the retrieval of heritage data for more abstract concepts. The fact that users were able to provide reasonable justifications and preferences for both query formats would suggest that control of query format would be useful for participants to be able to execute any given query. Participants appear to prefer image-based queries for objects that are difficult to describe, and text-based queries for abstract concepts that are difficult to represent in an image.

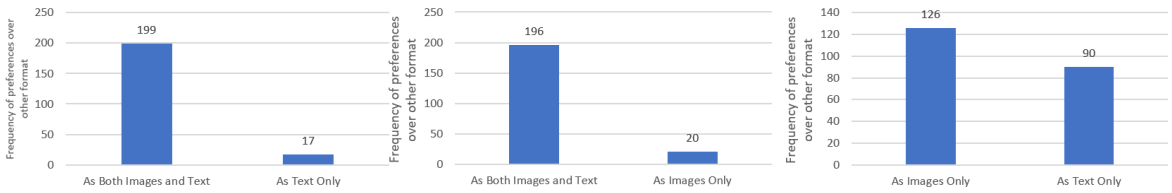

Figure 4.16: Pairwise comparisons of preferences for result views

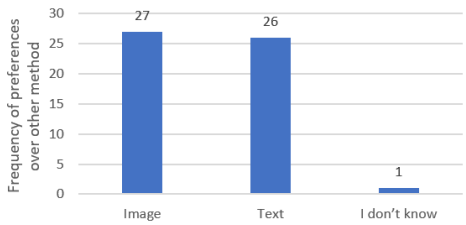

Figure 4.17: Participant preferences for query format

Participants were also asked to identify whether query formats, pre-processing algorithms or result formats can be potential contributors of bias in the system. Despite being aware of bias, and identifying pre-processing algorithms, result formats and query formats to contain bias, there were no general trends in the data. Figure 4.18 shows that no single image retrieval pre-processing algorithm was unanimously identified for containing the most bias across the pre-processing algorithms. This is a result from the analysis of the survey question asking participants to select the pre-processing algorithms they believed to contain biased results. PH, however, was recorded the least for containing the most bias and this correlates with the overall preference for that pre-processing algorithm. This is also observed with the results for the text retrieval pre-processing algorithms in Figure 4.19, where stemming was also identified the least (in total) for containing the most bias and was the most preferred text retrieval pre-processing algorithm. This suggests a potential relationship between user preference of pre-processing algorithms and user perception of algorithmic bias.

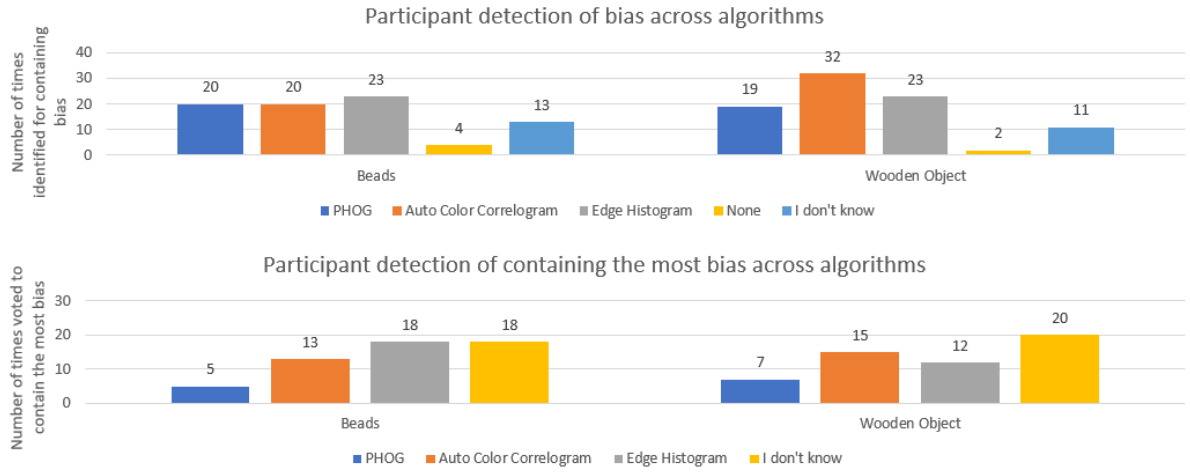

Figure 4.18: Participant identification of bias in image algorithms

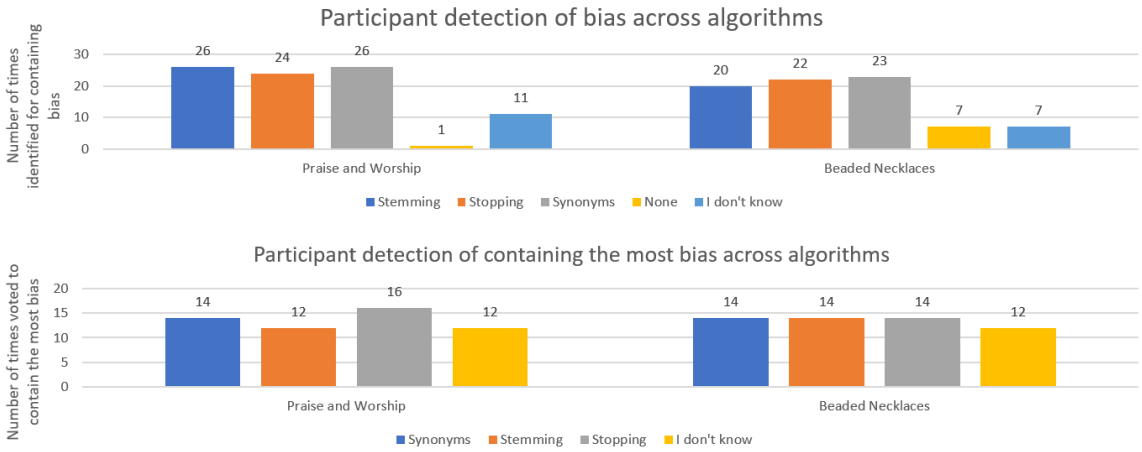

Figure 4.19: Participant identification of bias in text algorithms

Much like how preferred retrieval pre-processing algorithms were identified less frequently for containing bias, the preferred result format was also identified the least for containing the most bias among result formats (see Figure 4.20). Both unimodal result formats were identified 52% - 57% more frequently than the multimodal result view for containing the most bias. Similarly, to how a preference for query format was unable to be established, participants were also unable to unanimously identify a query format for containing more bias than the other (see Figure 4.21). Figure 4.22 groups the manner in which participants detected biases across search methods. Of the 85% of participants who detected biases in the search methods, 37% detected biases in both search methods, 35% detected biases when searching by image, and 28% detected biases when searching by text. Lastly, it was found that the number of participants who believed algorithmic variation can mitigate bias in retrieval grew from 65% before the experiment to 81% after the experiment (see Figure 4.23).

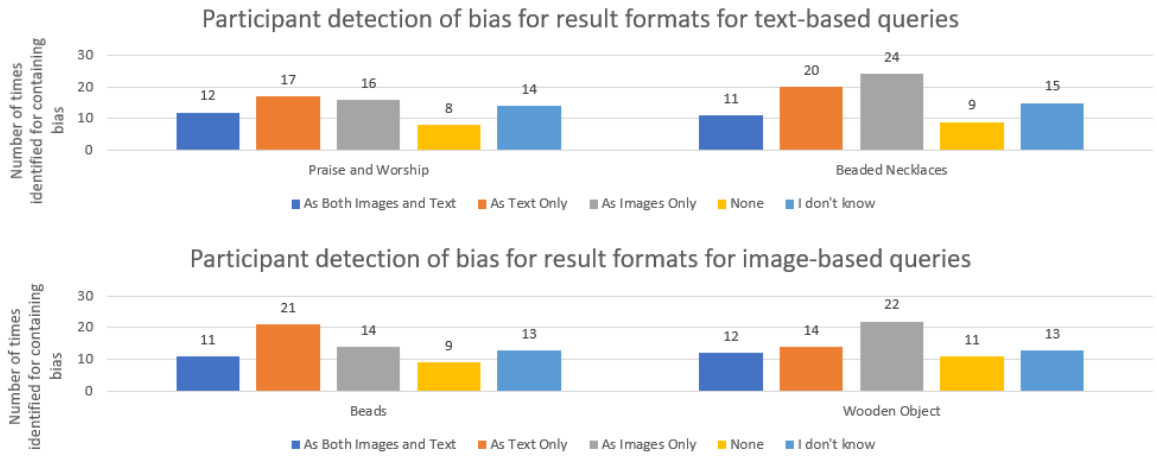

Figure 4.20: Participant identification of bias in result views

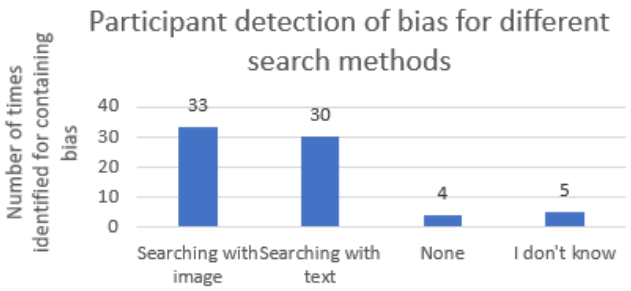

Figure 4.21: Participant identification of bias in query format

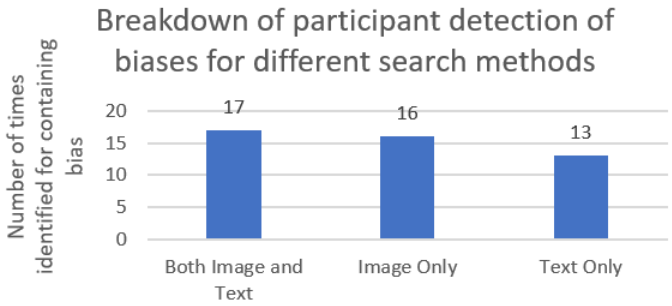

Figure 4.22: Breakdown of participant detection of biases for different search methods

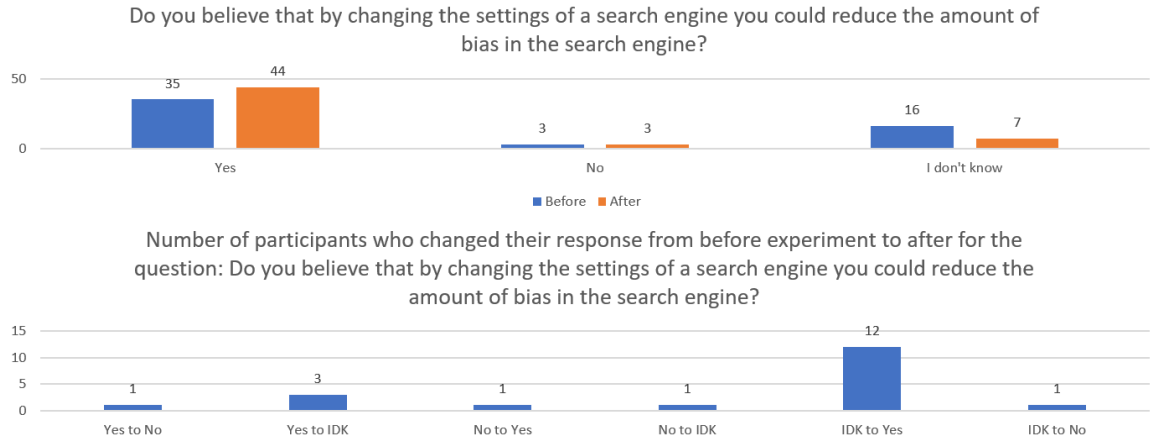

Figure 4.23: Participant change in response for whether algorithmic variation can mitigate bias in retrieval for before and after experiment

4.7 Summary

This chapter details an experiment conducted to investigate user experience and bias mitigation through the addition of user-control and algorithmic variation to a multimodal information retrieval system. A front-end for the multimodal information retrieval system built in Experiment 1 (see Chapter 3) was developed and hosted on a server for ease of access for participants. The experiment procedure was discussed in detail, as well as the

randomised block design that allowed for the control of potential extraneous variables and the counterbalancing measures taken to prevent any order effects. This chapter also details the pilot studies that were conducted to ensure the experiment design was feasible going into the formal experimentation phase. This chapter then detailed the results from 54 participants who were recruited for the experiment and were asked to perform 2 query tasks for each of the 2 different modalities, and, continuously engage with the experimental system by varying pre-processing algorithms and result formats. The results of participant feedback from survey responses were analysed, presented, and briefly discussed in this chapter. The results show that participants preferred the shape-based image retrieval pre-processing algorithm, PHOG, more than the other available pre-processing algorithms. Stemming was also identified as the most preferred text retrieval pre-processing algorithm. The results also showed that participants prefer multimodal result formats over unimodal result formats but were unable to identify a preferred query method. Despite this, participants provided reasonable justifications for the inclusion of each unimodal query method. With respect to bias, participants were unable to identify contributors of bias in the query method, algorithmic selection or result presentation phases of the experiment. However, most participants, following the experiment, believed that they can reduce the bias in a search engine from having control over the settings of the system. Conclusions that were reached on the basis of these findings are discussed further in the next chapter.

5. Conclusions

The aims of this research were to explore the effects on user experience and bias mitigation from the addition of user-control and algorithmic variation to a multimodal information retrieval system containing heritage data. This was investigated by conducting two experiments.

The first experiment provided a baseline offline evaluation of various pre-processing algorithms for text and image retrieval, in order to quantifiably determine whether algorithmic variation has an effect on retrieval effectiveness. This entailed submitting queries to a multimodal information retrieval system using multiple pre-processing text and image retrieval algorithms, and then calculating retrieval metrics for each of the pre-processing algorithms using relevance judgements for the result sets.

The second experiment was a user experiment to investigate the user experience with a retrieval system that allowed them to compare pre-processing algorithms. Users were asked to perform tasks using a search engine that allowed them to conduct queries via text or image, select the retrieval pre-processing algorithm and alternate among various result views. Users then provided feedback via a survey regarding their experience of controlling the system as well as their perceptions of bias in the system.

The results from both experiments assist with answering the research questions for the study. These answers, as well as conclusions reached on the basis of the findings, are outlined below.

5.1 Answers to research questions

This research presents several key findings, the first of which assists with answering research question 1: *What is the impact of text vs. image querying on user experience and retrieval effectiveness of unbiased results?*

Participants recorded preferences for both query formats and identified reasonable scenarios where both formats would be needed in order to satisfy effective retrieval, provide more information, and make for an easier retrieval experience. Mainly, image-based querying was preferred for queries that were difficult to describe and text-based querying was preferred for abstract query concepts.

Participants, however, seemingly struggled to unanimously identify contributors of bias in the retrieval systems. Despite being aware of bias, and identifying pre-processing algorithms, result formats and query formats to contain bias, general trends could not be extracted from the data. There was no single pre-processing algorithm, result format or query format, of either modality, identified unanimously for containing bias. This research reveals potential trends in user preference for algorithms and user perception of algorithmic bias, although more research in this regard is required.

The findings of this study also assist with answering research question 2: *What is the impact of variation of retrieval pre-processing algorithms on retrieval effectiveness of unbiased results?*

It was found that algorithmic variation can potentially help users improve retrieval effectiveness for image-based queries better than for text-based queries. Significant differences were found between image pre-processing algorithms for all retrieval metrics in the study, whereas this was only the case for precision, F-measure and number of documents retrieved for text pre-processing algorithms. The best performing image pre-processing algorithm with respect to recall – Edge Histogram – improved recall over the worst performing image pre-processing algorithm by 245%. This is consistent with prior research that suggests the value that different algorithms bring on a per-query basis and the value algorithmic variation brings, to allow users to decide which algorithm is best suited for their needs since they are the ultimate deciders of relevance (Fonseca, et al., 2005). This research also found that shape detection is potentially better performing for image retrieval than colour-based pre-processing algorithms for precolonial African heritage data. EH and PH significantly outperformed other pre-processing algorithms for this dataset, with AC being the only colour-based algorithm in the top 3 performing pre-processing algorithms. Participants also expressed an overall preference for the PH pre-processing algorithm. It was also found that synonym post-processing in retrieval has seemingly poor performance on precolonial heritage data. The thesaurus used in this research was the WordNet thesaurus, and results show that it affected retrieval performance negatively whenever it was used. This suggests that traditional thesauri may potentially not be effective in the context of heritage data.

A key finding of this study is that exposure to algorithmic variation can potentially be effective in persuading users in the belief that algorithmic variation can assist with mitigation bias in retrieval systems. The number of participants who believed that algorithmic variation could mitigate bias in retrieval grew from 65% before the experiment to 81% after the experiment. Of the 16 participants who responded “I don’t know” to that question before the experiment, 75% responded “Yes” at the end of the experiment.

Lastly, the findings that answer research question 3: *What is the impact of result display format on user experience?*

Users seem to prefer multimodal query formats and result views. Viewing results “As Both Images and Text” was preferred over the other two result views. The multimodal view was preferred 92% of the time over viewing results “As Text Only” and 91% of the time over “As Images Only”. This correlates with prior research that proposes that interface design is a core component of user-control and important to the improvement of system uptake and diversification of results (Thomas, et al., 2019).

5.2 Limitations

This research is not without limitations, and one of several that must be identified is the number of participants. With 54 participants and 18 research blocks, the research does satisfy the generalized random block design, but stronger results may be obtained with more participants. Participants were also largely sampled using the Convenience sampling method, a method prone to volunteer bias, as a result of the difficulty in finding willing participants during the COVID-19 pandemic and national lockdown in South Africa. This research, replicated with probability sampling methods instead, would produce more reliable results.

Secondly, all experiment sessions were conducted remotely due to COVID-19 and social distancing restrictions in South Africa. This prevented strict control of the experiment environment and conditions and many participants could have potentially stretched out their session across multiple days, or left the experiment intermittently to fulfil other commitments and responsibilities.

It must also be acknowledged that ancestral information about all participants was not captured. Given the nature of the data used in this study (precolonial African heritage data), and the sensitivity around this for those who are direct descendants of the cultures captured in the data, it must be considered that the perceptions of bias and efficiency of the information retrieval processing of this data may vary significantly among South African individuals of different descent. Lastly, due to time constraints, participants in the user experiment were only subjected to two tasks for each respective query format. Ideally, participants would have had to evaluate more tasks per query format to gain a better grasp of the differences between the two formats.

5.3 Conclusions

If there is an effort to decolonise the epistemic spaces in modern society, we cannot overlook the digital retrieval systems containing heritage data. Biases in the data and algorithms of information retrieval systems are well documented and researched, but their severity when dealing with heritage data is of notable importance. The misattribution of heritage data in these information retrieval systems has the potential to misinform users, reinforce stereotypes, or render the data irretrievable. It was attempted to adopt decolonisation principles in the digital retrieval of precolonial African heritage data through the addition of algorithmic variation of retrieval in the effort to mitigate bias, improve retrieval effectiveness and improve user experience. This research provides many encouraging results to assist with the development of digital retrieval systems containing heritage data. Especially if there is a concern of decolonisation, algorithmic variation should be potentially considered as a valuable addition to any digital retrieval system, but more pertinently for a retrieval system containing precolonial heritage data. Participants expressed a belief that algorithmic variation can assist with the mitigation of bias in retrieval systems, and the results also showed the ability of algorithmic variation to assist with potentially improving retrieval effectiveness on a per-query basis.

Secondly, our research suggests that users have a potential preference for multimodal query and result formats in retrieval systems. Participants largely identified the multimodal view as their most preferred view, and this should be a potentially important consideration in the design of future retrieval systems that contain multimodal data. Participants also justified the support for multimodal query formats, where image-based querying assists with concepts that are difficult to describe and text-based querying assists with concepts that are abstract.

Lastly, this research builds on prior research that argues for the need for systems to be designed with bias in mind. Whilst participants were unable to identify contributors of bias, the results of our research suggest that participants are aware of bias in query formats, algorithmic processes and result views, and that systems need to be designed in a manner that acknowledge and are transparent to their users about the presence of bias in these components.

Our study provides an entry into research towards the adoption of decolonisation principles in retrieval processes and the use of algorithmic variation to facilitate this, with refinement through future research. The findings of this study suggests the importance for future research in the field of heritage data preservation and retrieval to not solely focus on automation, but also appreciate and encourage the inclusion of user control of the heritage data experience.

5.4 Future Work

One aspect of this study that warrants future research is the finding that shape-based image pre-processing algorithms had better retrieval performance than colour-based image pre-processing algorithms. This relationship should be explored in greater detail and preferably with other heritage datasets to determine if there is a general trend with improved retrieval using shape-based image pre-processing algorithms with heritage data. The seemingly poor performance of synonym processing in heritage text data retrieval should also be explored further in research. The WordNet thesaurus was used in this study; it is recommended that the retrieval effectiveness from using other thesauri with heritage data is explored to establish whether the findings of this study are a result of the WordNet thesaurus's incompatibility with this dataset, or, whether synonym processing has a general degrading effect on retrieval effectiveness of heritage text data. Lastly, future research should attempt to further explore contributors of user-perceived bias in heritage data information retrieval systems. While the findings of this study suggest that participants are aware of the presence of bias, they were unable to unanimously identify contributors of bias with respect to query format, result format or pre-processing algorithm.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society, 20*(3), 973-989.

Anick, P. (2003). Using Terminological Feedback for Web Search Refinement - a Log-Based Study. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (pp. 88-95). Toronto, Canada.

ARL Working Group on Special Collections. (2009, March). *Special Collections in ARL Libraries: A Discussion Report from the ARL Working Group on Special Collections.* Retrieved from ARL: https://www.arl.org/wp-content/uploads/2009/03/scwg-report-mar09.pdf

Ashenfelder, M. (2009). 21st Century Shipping Network Data Transfer to the Library of Congress. *D-Lib Magazine, 15*(7/8).

Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, 8-19.

Becker, C., Kolar, G., Küng, J., & Rauber, A. (2007). Preserving interactive multimedia art: A case study in preservation planning. *International Conference on Asian Digital Libraries*, 257-266.

Beer, D. (2019). *The social power of algorithms.* Routledge.

Benson, A. C. (2010). Killed negatives: the unseen photographic archives. *Archivaria, 68*, 1-37.

Bhabha, H. (1983). Difference, discrimination and the discourse of colonialism. *The politics of theory*, 194-211.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. *Proceedings of the 6th ACM international conference on Image and video retrieval*, (pp. 401-408).

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology, 15*(3), 209-227.

Brand South Africa. (2017, May 26). *Mapungubwe: South Africa's lost city of gold*. Retrieved from Brand South Africa: https://www.brandsouthafrica.com/tourism-south-africa/travel/cultural/mapungubwe-south-africas-lost-city-of-gold

Brian, M. (2018). Decolonizing the University, Knowledge Systems and Disciplines in Africa. *African Studies Quarterly*, 121-122.

Burrell, J., Kahn, Z., & Jonas, A. (2019). When Users Control the Algorithms: Values Expressed in Practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1-20.

Caplan, P., & Guenther, R. (2005). Practical preservation: the PREMIS experience. *Library Trends, 54*(1), 111-124.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1721-1730).

Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., & Cord, M. (2018). Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 35-44.

Chatzichristofis, S. A., & Boutalis, Y. (2009). Selection of the proper Compact Composite Descriptor for improving content based image retrieval. *Proc. of the 6th IASTED International Conference*.

Chatzichristofis, S. A., & Boutalis, Y. S. (2008). CEDD: Color and Edge Directivity Descriptor. A Compact Descriptor for Image Indexing and Retrieval . *International Conference on Computer Vision Systems* (pp. 312-322). Berlin, Heidelberg: Springer.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62.

Efthimiadis, E. N. (1993). A User-Centered Evaluation of Ranking Algorithims for Interactive Query Expansion. *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, (pp. 146-159). Pittsburgh, PA, USA.

Fanon, F., Sartre, J.-P., & Farrington, C. (1963). *The Wretched of the Earth.* New York: Grove Press.

Fonseca, B., Golgher, P. B., Possas, B., & Ribeiro-neto, B. (2005). Concept-based interactive query expansion. *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, (pp. 696-703). Bremen, Germany.

Frants, V. I., Shapiro, J., Taksa, I., & Vladimir, G. (1999). Boolean search: Current state and perspectives. *Journal of the American Society for Information Science, 50*(1), 86-95.

Giaretta, D. (2006). CASPAR and a European infrastructure for digital preservation. *ERCIM News, 66*, 47-49.

Gill, J. (2018). Decolonizing Literature and Science. *Johns Hopkins University Press, 26*(3), 283-288.

Gracy, K. F., & Kahn, M. B. (2012). Preservation in the digital age. *Library Resources & Technical Services, 56*(1), 25-43.

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First women, second sex: Gender bias in Wikipedia. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 165-174.

Greenstein, S., & Zhu, F. (2012). Is Wikipedia Biased? *American Economic Review, 102*(3), 343-348.

Guangxin, R., Cai, J., Li, S., Yu, N., & Tian, Q. (2014). Scalable Image Search with Reliable Binary Code. *Proceedings of the 22nd ACM international conference on Multimedia*, 769-772.

Guill, K. L. (2009). Arguing for Space in an User-Focused Environment. *Library & Archival Security, 22*(2), 115-123.

Hamilton, K., Karahalios, K., Sandvig, C., & Eslami, M. (2014). A path to understanding the effects of algorithm awareness. *CHI'14 extended abstracts on human factors in computing systems*, 631-642.

Hare, J. S., & Lewis, P. H. (2013). Explicit diversification of image search. *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 295-296.

Herder, E., & van Maaren, O. (2020, July). Privacy Dashboards: The Impact of the Type of Personal Data and User Control on Trust and Perceived Risk. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 169-174.

Huang, J., Kumar, S., Mitra, M., Zhu, W.-J., & Zabih, R. (1997). Image Indexing Using Color Correlograms. *Proceedings of IEEE Computer Society conference on Computer Vision and Pattern Recognition*, (pp. 762-768).

Hube, C., & Fetahu, B. (2019). Neural based statement classification for biased language. *Proceedings of the twelfth ACM international conference on web search and data mining*, 195-203.

Huffman, T. N. (2000). Mapungubwe and the origins of the Zimbabwe culture. *Goodwin Series*, 14-29.

Imran, H., & Sharan, A. (2009). Thesaurus and query expansion. *International journal of computer science & information Technology (IJCSIT), 1*(2), 89-97.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422-446.

Järvelin, K., & Kekäläinen, J. (2009). Discounted Cumulated Gain. (L. Liu, & M. T. Özsu, Eds.) *Encyclopedia of Database Systems*.

Jones, M. L. (2017). The right to a human in the loop : Political constructions of computer automation and personhood. *Social Studies of Science*, (pp. 216-239).

Kasutani, E., & Yamada, A. (2001). The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)* , (pp. 674-677).

Kessi, S., Marks, Z., & Ramugondo, E. (2020). Decolonizing African Studies. *Critical African Studies, 12*(3), 271-282.

Knoche, M., Popović, R., Lemmerich, F., & Strohmaier, M. (2019). Identifying biases in politically biased wikis through word embeddings. *Proceedings of the 30th ACM conference on hypertext and social media*, 253-257.

Kowalski, G. J. (1997). Introduction to Information Processing Systems. In *Information Retrieval Systems* (Vol. 1). Boston, MA: Springer. Retrieved from Springer: https://link.springer.com/chapter/10.1007%2F978-0-585-32090-8_1

Laenen, K., Zoghbi, S., & Moens, M.-F. (2018). Web search of fashion items with multimodal querying. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 342-350.

Lancaster, F. (1968). *Information retrieval systems; characteristics, testing, and evaluation.*

Lashkari, A. H., Mahdavi, F., & Ghomi, V. (2009). A Boolean Model in Information Retrieval For Search Engines. *2009 International Conference on Information Management and Engineering*, 385-389.

Lo, R. T.-W., He, B., & Ounis, I. (2005, January). Automatically Building a Stopword List for an Information Retrieval System. *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), 5*, 17-24.

Long, M., Cao, Y., Wang, J., & Yu, P. S. (2016). Composite correlation quantization for efficient multimodal retrieval. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 579-588.

Lux, M., & Marques, O. (2013). Visual information retrieval using java and lire. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 5*(1), 1-112.

Lux, M., Riegler, M., Halvorsen, P., & MacStravic, G. (2017). LireSolr: A Visual Information Retrieval Server. *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 466-469.

Mbembe, A. (2015). Decolonizing Knowledge and the Question of the Archive.

Memmi, A., Greenfeld, H., Sartre, J., & Gordimer, N. (2013). *The colonizer and the colonized.* Routledge.

Meng, L., Tan, A.-H., Leung, C., Nie, L., Chua, T.-S., & Miao, C. (2015). Online multimodal co-indexing and retrieval of weakly labeled web image collections. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 219-226.

Minicka, M. (2006). Safegaurding Africa's Literary Heritage : Timbuktu rare manuscripts project. *Proceedings LIASA WCHELIG Winter Colloquium : Collaboration for success, Cape Town, South Africa*.

National Archives of Australia. (2010). *Digital Preservation Software Platform*. Retrieved from https://sourceforge.net/projects/dpsp/

National Library of New Zealand. (2007). *Metadata Extraction Tool*. Retrieved from National Library of New Zealand: http://meta-extractor.sourceforge.net/

Nemeth, Y., Shapira, B., & Taeib-Maimon, M. (2004). Evaluation of the real and perceived value of automatic and interactive query expansion. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 526-527).

Noble, S. U. (2013). Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. *InVisible Culture, 19*.

Oliver, G., & Harvey, R. (2016). *Digital curation.* American Library Association.

Olojede, A., & Suleman, H. (2015). Investigating Image Processing Algorithms for Navigating Cultural Heritage Spaces using Mobile Devices. *International Conference on Asian Digital Libraries* (pp. 215-224). Springer.

Orji, R., Oyibo, K., & Tondello, G. F. (2017). A Comparison of System-Controlled and User-Controlled Personalization Approaches. *Adjunct publication of the 25th conference on user modeling, adaptation and personalization*, (pp. 413-418). Bratislava, Slovakia.

Otterbacher, J., Bates, J., & Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. *Proceedings of the 2017 chi conference on human factors in computing systems*, 6620-6631.

Parker, K. R. (2016). Introduction decolonizing the university: A battle for the African mind. *CLA Journal, 60*(2), 164-171.

Petras, V., Hill, T., Stiller, J., & Gäde, M. (2017). Europeana–a search engine for digitised cultural heritage material. *Datenbank-Spektrum, 17*(1), 41-46.

Pickover, M., & Peters, D. (2002). DISA: An African Perspective on Digital Technology. *Innovation, 24*, 14-20.

Pogacar, F. A., Ghenai, A., Smucker, M. D., & Clarke, C. L. (2017). The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, (pp. 209-216).

Rahdari, B., & Brusilovsky, P. (2019). User-controlled hybrid recommendation for academic papers. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, (pp. 99-100).

Roy, A., Ghosh, K., Basu, M., Gupta, P., & Ghosh, S. (2018). Retrieving information from multiple sources. *Companion Proceedings of the The Web Conference 2018*, 43-44.

Ruthven, I. (2003). Re-examining the Potential Effectiveness of Interactive Query Expansion. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 213-220). Toronto, Canada.

Salton, G., Buckley, C., & Fox, E. A. (1983). Automatic query formulation. *Journal of the American society for information science, 34*(4), 262-280.

Seah, B.-S., Bhowmick, S. S., & Sun, A. (2014). Summarizing social image search results. *Proceedings of the 23rd International Conference on World Wide Web*, 369-370.

Shenton, H. (2009). Virtual reunification, virtual preservation and enhanced conservation. *Alexandria, 21*(2), 33-45.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98*, 277-284.

Simpson, T. W. (2012). Evaluating Google as an Epistemic Tool. *Metaphilosophy, 43*(4), 426-445.

Skotnes, P., & Bleek, W. H. (2007). *Claim to the country: the archive of Lucy Lloyd and Wilhelm Bleek.* Jacana Media.

Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology*, 226-234.

Suleman, H. (2011). An African Perspective on Digital Preservation. *Multimedia Information Extraction And Digital Heritage Preservation*, 295-306.

Sun Won, C., Kwon Park, D., & Park, S.-J. (2002). Efficient Use of MPEG-7 Edge Histogram Descriptor. *ETRI journal, 24*(1), 23-30.

Sweetnam, M. S., Agosti, M., Orio, N., Ponchia, C., Steiner, C. M., Hillemann, E.-C., . . . Lawless, S. (2012). User needs for enhanced engagement with cultural heritage collections. *International Conference on Theory and Practice of Digital Libraries*, 64-75.

Tala, F. Z. (1999). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.

The Five Hundred Year Archive. (n.d.). *About*. Retrieved from FHYA: https://fhya.org/about

Thomas, P., Billerbeck, B., Craswell, N., & White, R. W. (2019). Investigating Searchers' Mental Models to Inform Search Explanations. *ACM Transactions on Information Systems, 38*(1), 1-25.

Ting, K. M. (2011). Precision and Recall. (C. Sammut, & G. Webb, Eds.) *Encyclopedia of Machine Learning*.

Tsandilas, T., & Shcraefel, M. C. (2003). User-controlled link adaptation. *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, (pp. 152-160). Nottingham, UK.

University of Cape Town. (2007). *Percival Kirby Musical Instruments*. Retrieved from UCT Libraries Digital Collections: https://digitalcollections.lib.uct.ac.za/percival-kirby-musical-instruments

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1).

White, R. (2013). Beliefs and biases in web search. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, (pp. 3-12).

Whitelaw, M. (2009). Visualising archival collections: The visible archive project. *Archives and Manuscripts, 37*(2), 22.

Wilkie, C., & Azzopardi, L. (2017, November). Algorithmic bias: do good systems make relevant documents more retrievable? *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2375-2378.

Williams, K., Manilal, S., Molwantoa, L., & Suleman, H. (2010). A Visual Dictionary for an Extinct Language. *The Role of Digital Libraries in a Time of Global Change* (pp. 1-4). Berlin, Heidelberg: Springer.

Wong, S. M., & Raghavan, V. (1984). Vector space model of information retrieval: a reevaluation. *SIGIR, 84*, 167-185.

Xu, J. (1997). Solving the word mismatch problem through automatic text analysis. *Doctoral dissertation, University of Massachusetts at Amherst*.

Yang, K., & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1-6.

Zhang, E., & Zhang, Y. (2009). F-Measure. (L. Liu, & M. T. Özsu, Eds.) *Encyclopedia of Database Systems*.

Zhao, K., Wei, E., Sui, Q., Zhu, K. Q., & Lo, E. (2013). CISC: clustered image search by conceptualization. *Proceedings of the 16th International Conference on Extending Database Technology*, 749-752.

Appendices

Appendix A: 70 Text-based queries used for Experiment 1

Information Need	Query
Africans	africa
Tradition	african attire
Traditional African religions	african culture
Music of Africa	african dance
Africans	african dresses
Gospel Music	african gospel
History of Africa	african history
Music of Africa	african instruments
Africans	african languages
Africans	african names
Painting	african paintings
Africans	african people
Religion	african religion
Music of Africa	african songs
Beadwork	beaded necklaces
Beadwork	beads
Clan	clan names
Clan	clan praises
Clan	dlamini clan
Eating	eating
Fruit	fruit and vegetables
Fruit	fruit diet
Fruit	fruit trees
Gospel Music	gospel music songs
Healing	healing prayer
Culture	heritage
Indigenous Peoples	indigenous dance
Indigenous Peoples	indigenous plants
Praise	joyous celebration
Clan	khumalo clan
Health Care	life
Religious Text	love scriptures
Culture	material culture
Meat	meat
Clan	mthembu clan
Clan	nkosi clan
Worship	praise
Praise	praise and worship
Praise	praise god
Praise	praise poem
Healing	prayer
Healing	prayer for healing
Religious Text	prayer scriptures
Meat	red meat

Religion in South Africa	religions in south africa
Religious Text	scriptures about faith
Religious Text	scripture about love
Dance	south african dance
Indigenous Peoples	south african indigenous games
Religion in South Africa	south african religions
Spirituality	spiritual father
Spirituality	spiritual warfare prayers
Praise	spirit of praise
Statue	statues in south africa
Statue	statue meaning
Statue	status
Symbol	symbolic meaning
Dance	traditional dance
Tradition	traditional designs
Design	traditional dresses designs
Healer	traditional healers in south africa
Healing	traditional healing
Tradition	traditional outfits
Society	traditional society
Praise	woman in praise
Clan	xhosa clan names
Healer	xhosa traditional healers
Beadwork	zulu beadwork
Clan	zulu clan names
Dance	zulu dance

Appendix B: System image and text algorithm explanation pages

Text Configurations:

Stemming:
It is very likely that documents in this retrieval system are going to use different forms of a word, such as *reproduces, reproducing,* and *reproduced*

The goal of stemming is to reduce various forms of a word to their common base form. For instance, the result of applying stemming to the following text would be as follows:

the boy's cars are different colors -> the boy car are differ color

This system makes use of the Porter's Stemming Algorithm.

Stopping:
Stopwords are words that are frequently occuring in a language, but which might not provide much meaning to a document in this system.

Enabling Stopping will make the system enforce Stopword Removal, that is, to remove stopwords from the query.

The following are the words of the English language considered as Stopwords by this system:

a, an, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, s, such, t, that, the, their, then, there, these, they, this, to, was, will, with

Synonyms:
The documents in the system may not contain the exact word from your query, but rather, a synonym of the word you have entered.

When enabling Synonyms, the query is expanded to include synonyms of the words from the original query, thereby diversifying the potential number of words to be matched against documents in the collection.

This system makes use of the WordNet thesaurus, found here.

Image Configurations:

Spatial correlation of color:
A configuration which matches images based on the spatial correlation of colors. It is able to tolerate changes in the aesthetics of images as a result of changes in viewing positions or zooms.

Below are images considered to be similar using this configuration:

Local edge distribution:
A configuration which matches images based on changes in the frequency and directionality of brightness changes in the images. It does so through examining local edges in the images, that is, edges which occur within regions of the images.

Below are images considered to be similar using this configuration:

Shape similarity and spatial layout:
A configuration which matches images based on their shape correspondence. This makes use of two attributes of the images:

- Local image shape - shapes, structures and/or patterns found in different regions within the image
- Spatial layout - the location of these local shapes and features within the image

Below are images considered to be similar using this configuration:

Appendix C: Experiment participation instruction template

Hello {INSERT NAME HERE},

Thank you for choosing to participate in my experiment investigation. The experiment has been designed to be completed within an hour, although it would be safest to allocate 1½ hours for completion.

Please only open the experiment link below when you are planning on starting and completing the entire experiment session. If you open this before and do not complete it, it will result in duplicate entries.

The **required browser** for the duration of this experiment is Google Chrome. **Please open all links using Google Chrome as the browser**. Other browsers may not be supported and may cause issues during the experiment, resulting in an invalid session. If you do not have Google Chrome installed, you may find it here: https://www.google.com/chrome/ (this is a verified and secure link directly from Google Chrome).

When completing the survey, you will notice that questions that require a response are indicated with an asterisk (*) whereas optional questions do not have an asterisk. It would be preferable if optional questions are answered, but these are not required.

Other guidelines to assist you:

- Please ensure that **at all times the survey and search engine tabs are open** (links found below) for the duration of the experiment. You will be required to alternate between these two tabs during the entire session.
- It is recommended that the introductory video is open in a separate tab as well (link found below).
- Please ensure that on completion of the survey that you **submit the survey and receive confirmation of submission from SurveyMonkey**.
- Please be patient with the search engine if it is unresponsive or if images take a while to download. This may be an issue for slower download speeds.

If you come across any issues, or require further assistance or clarification, do not hesitate to contact the researcher on sohamsingh@live.co.za or on +27 76 302 6424.

Please do not share any of the following links with anyone, they are made on a per-participant basis. If you wish to refer someone to the experiment kindly ask them to contact the researcher using the details above.

You may find the link to the survey
here: https://www.surveymonkey.com/r/stoppingEhImage

You may find a link to the introductory video here (this is also available in the survey):
https://www.youtube.com/watch?v=K-zD6B7SW3M

You may find a link to the search engine you will be using here (this is also available in the survey): http://pumbaa.cs.uct.ac.za/~soham/experiment.html?text=stopping&image=eh&first

SearchMethod=image&resultViewOrder=4&textTaskOrder=0&imageTaskOrder=1&dynamicConfigStartingSide=right

Thank you once again for your participation, it is greatly valued to the researcher. When you have submitted your survey, please inform the researcher via any means of communication that you have completed your session.

Please can I ask that you submit your survey by the **{INSERT DUE DATE HERE}**. If this is an issue, please contact the researcher as soon as possible so that a new deadline date can be arranged.

Kind regards,

Soham Singh

sohamsingh@live.co.za

+27 76 302 6424

Appendix D: Informed Voluntary Consent to Participate in Research Study form

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF CAPE TOWN
PRIVATE BAG X3
RONDEBOSCH 7701
SOUTH AFRICA

RESEARCHER/S: Soham Singh
TELEPHONE: +27-76-302-6424
E-MAIL: sohamsingh@live.co.za

Informed Voluntary Consent to Participate in Research Study

Investigating User Experience of a Multi-Modal Search Interface

Invitation to participate, and benefits: You are invited to participate in a research study conducted with potential end-users for a search engine containing heritage data. The study aim is to evaluate the effects on user experience from the addition of a number of key features that have been implemented into the search engine system. I believe that your experience would be a valuable source of information, and hope that by participating you may gain useful knowledge.

Procedures: During this study, you will be asked to perform a series of tasks using a search engine, and then provide feedback on these tasks through an electronic survey.

Recording: We may record audio/video as part of the study. This will be used as correlatory information to add an extra layer of information beyond that which you will provide in the survey. If you object to this, please indicate below.

Risks: There are no potentially harmful risks related to your participation in this study.

Feedback: You will receive feedback about the results of this research upon completion of the project, in the form of an electronic research poster outlining the research and the results of this experimental process.

Disclaimer/Withdrawal: Your participation is completely voluntary; you may refuse to participate, and you may withdraw at any time without having to state a reason and without any prejudice or penalty against you. Should you choose to withdraw, the researcher commits not to use any of the information you have provided without your signed consent. Note that the researcher may also withdraw you from the study at any time.

Confidentiality: All information collected in this study will be kept private in that you will not be identified by name or by affiliation to an institution. Confidentiality and anonymity will be maintained as pseudonyms will be used.

What signing this form means: By signing this consent form, you agree to participate in this research study. The aim, procedures to be used, as well as the potential risks and benefits of your participation have been explained verbally to you in detail, using this form. Refusal to participate in or withdrawal from this study at any time will have no effect on you in any way. You are free to contact me, to ask questions or request further information, at any time during this research.

I agree to participate in this research (tick one box) ☐ Yes ☐ No _____ (Initials)
I agree to be audio-recorded ☐ Yes ☐ No _____ (Initials)
I agree to be video-recorded ☐ Yes ☐ No _____ (Initials)
I agree to the use of properly anonymized audio recordings/videos to be used as observational information for this research. ☐ Yes ☐ No _____ (Initials)

_____	_____	_____
Name of Participant	Signature of Participant	Date
SOHAM HANUMAN SINGH	_____	_____
Name of Researcher	Signature of Researcher	Date

Appendix E: Survey for Experiment 2

Investigating User Experience of a Multi-Modal Search Interface

Introduction

Please watch the following introductory video **before proceeding with the survey** (contains audio):

It is recommended that you also open this video in a separate tab so that you may refer to it later on in the session. You can do so by clicking here.

* What is your age?

What is your gender?

* In a search engine, there exists the potential for the system to contain bias. Refer to the introductory video above for an explanation of how this may occur.

As a result of this, it may be possible for results a user would find relevant, to not be shown to the user when searching.

Do you believe that there is any bias in the results shown to you by the search engines that you have used before participating in this experiment?

- Yes
- No
- I don't know

Please explain your response to the above question.

* Do you believe that by changing the settings of a search engine (if they were available to you) you could reduce the amount of bias in the search engine?

- ◯ Yes
- ◯ No
- ◯ I don't know

You have completed the introduction. To navigate to the search engine you will be using for the experiment, please click **here**.

PLEASE NOTE: You must keep this current survey tab open at all times alongside the search engine page. You will be asked to work in both windows throughout the experiment. Clicking the link above will automatically open the search engine in a new tab/window.

You may now proceed to the next phase of the survey.

4

Investigating User Experience of a Multi-Modal Search Interface

Searching with Text - First Task

Please refer to the instruction window in the search engine for how to proceed with the task. When prompted to do so, return back to this tab and answer the survey questions for this task. Keep both tabs open at all times for the duration of the experiment.

Note: when you have completed the steps for each task, do not click on the 'Next Task' button in the search engine until you have completed answering the survey questions for the task. This will allow you to go back to the system and confirm your responses should you have difficulty remembering the results.

Here is the example of bias that was explained in the introduction video, for your reference (if you recall this then you do not need to read this again, it is to save time for those trying to find it in the video again):

"There is a person named Jane. Jane's job is to write about movies on the internet. Jane writes about a large variety of movies of all genres, but Jane's favourite genre of movie is Action. Whenever Jane writes about Action movies, Jane describes them with positive words like 'exciting', 'fun', 'enjoyable'. Jane also happens to hate Comedies. When Jane has to write about a Comedy, she usually describes it as 'boring' or 'annoying'.

If someone was to search through Jane's movie write-ups online, it would be more likely that they see Jane's Action write ups, rather than Jane's comedy write-ups, due to how Jane describes Action films more positively than Comedies. In other words, Jane has allowed her bias of liking Action films over Comedy films, to affect how likely it is for Comedy film write-ups to be found in the search engine. This is an illustration of how bias in a search engine exists and can be applied to many types of information far beyond the scope of movie write-ups."

* Please rank the 3 configuration settings in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the settings as you completed the task.

Stemming

Stopping

Synonyms

Please explain your answer to the above question.

5

* Which of the settings did you find to be of value? (select multiple)

- [ ] Stemming
- [ ] Stopping
- [ ] Synonyms
- [ ] None

Please explain your answer to the above question.

* Which of the settings do you believe contained bias in their results? (select multiple)

- [ ] Stemming
- [ ] Stopping
- [ ] Synonyms
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

* Which setting do you believe contained the most bias in the results?

- ( ) The Stemming results
- ( ) The Stopping results
- ( ) The Synonyms results
- ( ) I don't know

Please explain your answer to the above question.

6

* Please rank the 3 result views in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the result views for this task.

≡ As Both Images and Text

≡ As Text Only

≡ As Images Only

Please explain your answer to the above question.

* Which of the result views do you believe contained bias in their presentation of results? (select multiple)

- [ ] As Both Images and Text
- [ ] As Text Only
- [ ] As Images Only
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

You may now proceed to the next task.

7

Investigating User Experience of a Multi-Modal Search Interface

Searching with Text - Second Task

Please refer to the instruction window in the search engine for how to proceed with the task. When prompted to do so, return back to this tab and answer the survey questions for this task. Keep both tabs open at all times for the duration of the experiment.

Note: when you have completed the steps for each task, do not click on the 'Next Task' button in the search engine until you have completed answering the survey questions for the task. This will allow you to go back to the system and confirm your responses should you have difficulty remembering the results.

Here is the example of bias that was explained in the introduction video, for your reference (if you recall this then you do not need to read this again, it is to save time for those trying to find it in the video again):

"There is a person named Jane. Jane's job is to write about movies on the internet. Jane writes about a large variety of movies of all genres, but Jane's favourite genre of movie is Action. Whenever Jane writes about Action movies, Jane describes them with positive words like 'exciting', 'fun', 'enjoyable'. Jane also happens to hate Comedies. When Jane has to write about a Comedy, she usually describes it as 'boring' or 'annoying'.

If someone was to search through Jane's movie write-ups online, it would be more likely that they see Jane's Action write ups, rather than Jane's comedy write-ups, due to how Jane describes Action films more positively than Comedies. In other words, Jane has allowed her bias of liking Action films over Comedy films, to affect how likely it is for Comedy film write-ups to be found in the search engine. This is an illustration of how bias in a search engine exists and can be applied to many types of information far beyond the scope of movie write-ups."

* Please rank the 3 configuration settings in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the settings as you completed the task.

Stemming

Stopping

Synonyms

Please explain your answer to the above question.

8

* Which of the settings did you find to be of value? (select multiple)

- [ ] Stemming
- [ ] Stopping
- [ ] Synonyms
- [ ] None

Please explain your answer to the above question.

* Which of the settings do you believe contained bias in their results? (select multiple)

- [ ] Stemming
- [ ] Stopping
- [ ] Synonyms
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

* Which setting do you believe contained the most bias in the results?

- ( ) The Stemming results
- ( ) The Stopping results
- ( ) The Synonyms results
- ( ) I don't know

Please explain your answer to the above question.

9

* Please rank the 3 result views in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the result views for this task.

≡ As Both Images and Text

≡ As Text Only

≡ As Images Only

Please explain your answer to the above question.

* Which of the result views do you believe contained bias in their presentation of results? (select multiple)

- [ ] As Both Images and Text
- [ ] As Text Only
- [ ] As Images Only
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

You may now proceed to the next task.

10

Investigating User Experience of a Multi-Modal Search Interface

Searching with Image - First Task

Please refer to the instruction window in the search engine for how to proceed with the task. When prompted to do so, return back to this tab and answer the survey questions for this task. Keep both tabs open at all times for the duration of the experiment.

Note: when you have completed the steps for each task, do not click on the 'Next Task' button in the search engine until you have completed answering the survey questions for the task. This will allow you to go back to the system and confirm your responses should you have difficulty remembering the results.

Here is the example of bias that was explained in the introduction video, for your reference (if you recall this then you do not need to read this again, it is to save time for those trying to find it in the video again):

"There is a person named Jane. Jane's job is to write about movies on the internet. Jane writes about a large variety of movies of all genres, but Jane's favourite genre of movie is Action. Whenever Jane writes about Action movies, Jane describes them with positive words like 'exciting', 'fun', 'enjoyable'. Jane also happens to hate Comedies. When Jane has to write about a Comedy, she usually describes it as 'boring' or 'annoying'.

If someone was to search through Jane's movie write-ups online, it would be more likely that they see Jane's Action write ups, rather than Jane's comedy write-ups, due to how Jane describes Action films more positively than Comedies. In other words, Jane has allowed her bias of liking Action films over Comedy films, to affect how likely it is for Comedy film write-ups to be found in the search engine. This is an illustration of how bias in a search engine exists and can be applied to many types of information far beyond the scope of movie write-ups."

* Please rank the 3 configuration settings in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the settings as you completed the task.

Shape Similarity and Spatial Layout

Spatial Correlation of Color

Local Edge Distribution

Please explain your answer to the above question.

11

* Which of the settings did you find to be of value? (select multiple)

- [ ] Spatial Correlation of Color
- [ ] Shape Similarity and Spatial Layout
- [ ] Local Edge Distribution
- [ ] None

Please explain your answer to the above question.

* Which of the settings do you believe contained bias in their results? (select multiple)

- [ ] Shape Similarity and Spatial Layout
- [ ] Spatial Correlation of Color
- [ ] Local Edge Distribution
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

* Which setting do you believe contained the most bias in the results?

- ( ) The 'Local Edge Distribution' results
- ( ) The 'Spatial Correlation of Color' results
- ( ) The 'Shape Similarity and Spatial Layout' results
- ( ) I don't know

Please explain your answer to the above question.

12

* Please rank the 3 result views in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the result views for this task.

≡ As Both Images and Text

≡ As Text Only

≡ As Images Only

Please explain your answer to the above question.

* Which of the result views do you believe contained bias in their presentation of results? (select multiple)

- [ ] As Both Images and Text
- [ ] As Text Only
- [ ] As Images Only
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

You may now proceed to the next task.

13

Investigating User Experience of a Multi-Modal Search Interface

Searching with Image - Second Task

Please refer to the instruction window in the search engine for how to proceed with the task. When prompted to do so, return back to this tab and answer the survey questions for this task. Keep both tabs open at all times for the duration of the experiment.

Note: when you have completed the steps for each task, do not click on the 'Next Task' button in the search engine until you have completed answering the survey questions for the task. This will allow you to go back to the system and confirm your responses should you have difficulty remembering the results.

Here is the example of bias that was explained in the introduction video, for your reference (if you recall this then you do not need to read this again, it is to save time for those trying to find it in the video again):

"There is a person named Jane. Jane's job is to write about movies on the internet. Jane writes about a large variety of movies of all genres, but Jane's favourite genre of movie is Action. Whenever Jane writes about Action movies, Jane describes them with positive words like 'exciting', 'fun', 'enjoyable'. Jane also happens to hate Comedies. When Jane has to write about a Comedy, she usually describes it as 'boring' or 'annoying'.

If someone was to search through Jane's movie write-ups online, it would be more likely that they see Jane's Action write ups, rather than Jane's comedy write-ups, due to how Jane describes Action films more positively than Comedies. In other words, Jane has allowed her bias of liking Action films over Comedy films, to affect how likely it is for Comedy film write-ups to be found in the search engine. This is an illustration of how bias in a search engine exists and can be applied to many types of information far beyond the scope of movie write-ups."

* Please rank the 3 configuration settings in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the settings as you completed the task.

Shape Similarity and Spatial Layout

Spatial Correlation of Color

Local Edge Distribution

Please explain your answer to the above question.

14

* Which of the settings did you find to be of value? (select multiple)

- [ ] Spatial Correlation of Color
- [ ] Shape Similarity and Spatial Layout
- [ ] Local Edge Distribution
- [ ] None

Please explain your answer to the above question.

* Which of the settings do you believe contained bias in their results? (select multiple)

- [ ] Shape Similarity and Spatial Layout
- [ ] Spatial Correlation of Color
- [ ] Local Edge Distribution
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

* Which setting do you believe contained the most bias in the results?

- ( ) The 'Local Edge Distribution' results
- ( ) The 'Spatial Correlation of Color' results
- ( ) The 'Shape Similarity and Spatial Layout' results
- ( ) I don't know

Please explain your answer to the above question.

15

* Please rank the 3 result views in order from most preferred (first) to least preferred (third).

Note: you may refer to the feedback window on the right hand side of the screen in the search engine page. This will have your rankings of the result views for this task.

As Both Images and Text

As Text Only

As Images Only

Please explain your answer to the above question.

* Which of the result views do you believe contained bias in their presentation of results? (select multiple)

- [ ] As Both Images and Text
- [ ] As Text Only
- [ ] As Images Only
- [ ] None
- [ ] I don't know

Please explain your answer to the above question.

You may now proceed to the next task.

16

Investigating User Experience of a Multi-Modal Search Interface

Conclusion

When answering these questions, consider the search engine you have just used in its entirety.

* Did you find you preferred searching with images, or searching with text?

- ○ Searching with images
- ○ Searching with text
- ○ I don't know

Please explain your answer to the above question.

* Please select which methods of search you believe contained bias when used (select multiple):

- ☐ Searching with images contained bias
- ☐ Searching with text contained bias
- ☐ None
- ☐ I don't know

Please explain your answer to the above question.

* Do you believe that by changing the settings of a search engine (as you did when selecting settings with the dropdown menu) you can reduce the amount of bias in the search engine?

- ○ Yes
- ○ No
- ○ I don't know

Please explain your answer to the above question.

Do you have any suggestions for improvements of the search engine?

17

Do you have any general comments regarding the experiment, survey or system?

You have now completed the survey. Your participation in this study is greatly appreciated.

18