Exploring Digital Cultural Heritage through Browsing

Mark M. Hall is a Lecturer in Computing & Communications at the Open University, UK. His research interests focus on the intersection between computation and the human user, particularly in information retrieval and the digital humanities. He is particularly interested in helping users explore large data-sets that they are unfamiliar with or where the users lack the expertise to know what they could find and how to find it. As part of this he is also undertaking research into how to evaluate interfaces that support such open-ended exploration.

David Walsh is a Senior Lecturer in Computer Science at Edge Hill University, UK. He is also a PhD student in the Information School at the University of Sheffield. His research interests include information seeking and user experience design, to find potential solutions to human-centred problems, particularly focussing on the digital cultural heritage sector. His PhD aims to identify and understand the user groups of digital cultural heritage including their preferences for seeking information and establishing a generous interface that enables all user groups to effectively seek DCH information.

Abstract

Digitisation of our cultural heritage has created vast digital archives, many of which are publicly accessible via the web. However, publicly accessible does not necessarily mean that the content of the collection is truly accessible to the wider public. The reason for this is that the standard interface for accessing the collections is the search box. This is very effective for experienced users who have the information seeking skills and domain knowledge to formulate appropriate queries, but for the less knowledgeable users the white search box represents a significant hurdle. This group of users require a generous interface that provides them with an overview over the collection and the ability to explore it, without having to explicitly enter a search keyword. In this chapter we discuss a range of existing

approaches that have been used to provide less knowledgeable users with such interfaces and then present the Digital Museum Map, an algorithm and interface for automatically generating a virtual museum interface for an unstructured collection, that can then be explored by browsing through the virtual museum.

Introduction

Digitisation of our cultural heritage by galleries, libraries, archives, and museums (GLAM) has created vast digital archives, many of which are publicly accessible via the web. These archives should, in theory, widen access to our digital cultural heritage (DCH), however, in practice, GLAM websites frequently experience bounce rates of over 60%, meaning they lose more than half of their visitors after the first page (Hall et al., 2012b; Walsh et al., 2020). This mirrors a common complaint in the wider field of digital libraries: “So what use are the digital libraries, if all they do is put digitally unusable information on the web?” (Borgmann, 2010).

Users visit GLAM websites for a wide range of reasons, ranging from professional work goals to pure leisure activities. They may be planning a physical visit to the GLAM, to find out about the institution itself, to buy something from its online shop, or to explore the GLAM’s digital holdings. Some may be visiting to find something specific, some may be looking for more general inspiration, and some purely to spend some time. The visitors’ degrees of background knowledge and expertise will also vary significantly. Supporting this vast range of potential visitor requirements is of course very difficult, however the very high bounce rates experienced by GLAM websites indicates that there is a significant fraction of the potential visitors, for whom the current provisions do not work.

Out of the range of user characteristics and goals for visiting GLAM websites this chapter will focus on supporting access to the GLAM’s digital holdings, in particular for users who have a less focused goal or less domain expertise. The reason for this is that focused search and high-expertise users are already

served quite well by the most common interface for accessing the digital holdings: the search box. However, Koch et al. (2006), Ruecker et al. (2011), and Walsh et al. (2020) show that the majority of users prefer to use browsing a navigation structure as the way to access the information they are seeking. Browsing, as a concept, covers a wide range of behaviours (Bates, 1989; Rice et al., 2001; Bates, 2007), but for the purposes of this chapter we use a very broad definition of a browsing-based interface as one that allows the user to interact with a collection without having to explicitly enter a search keyword. A search system may be running in the background, but this must either not be visible to the user or interaction with the search system must be an optional extra.

Our focus is on browsing because for users with less domain expertise or less focused goals the blank search box is a known and significant barrier (Belkin, 1982; Whitelaw, 2015). These kinds of users are also more likely to visit in a leisure, rather than a work context (Wilson and Elsweiler, 2010) and thus tend to follow more exploratory behaviour (Mayr et al., 2016). Supporting these more open-ended goals and less experienced users in their interactions with search interfaces in general is traditionally seen as the domain of exploratory search (Marchionini, 2006; White et al., 2009). Exploratory search interfaces are generally designed to provide the user with guidance as to which keywords will produce search results and to help them narrow down their search, using features such as query suggestion and search facets. While these provide some indications as to which search terms will produce results, they generally do not provide an overview of the collection as a whole and still assume that users come to the system with at least a partially defined goal. Addressing this gap has been the task of exploratory interfaces developed under the labels of rich prospect browsing (Ruecker et al., 2011) and generous interfaces (Whitelaw, 2015). While they differ in some aspects, which will be explored in more detail later, both labels place a strong focus on providing the user with an initial overview over the collection and on allowing the exploration of said collection without having to enter a search query. These ideas have spurred the development of a range of interesting and innovative interfaces, however none of

these have seen any major uptake outside the institutions they were developed for. This is in part because developing or even just adapting such interfaces is significantly more complex than deploying an off-the-shelf search interface (Ruecker et al., 2011), but also because, particularly for museums, the question of what a digital, virtual, or online presentation should be is still a contested area (Biedermann, 2017; Meehan, 2020).

The remainder of the chapter is structured as follows: first we will investigate the current state of interfaces for accessing GLAM's digital holdings in more detail. We will look at the kind of data available to users, the boundaries experienced by users in accessing the data, the types of interfaces that have been developed to overcome these boundaries, and techniques for automatically structuring collections to support access. The second major section will introduce the Digital Museum Map (DMM). The DMM demonstrates how to address the issues with exploring large, digital collections, by providing a generous, browsable interface, that is based on an automatically generated organisational hierarchy, and that can be applied to any collection without major human input.

Background

Many GLAMs house large heterogeneous collections and, through digitisation, have created digital collections covering parts of their physical collections. In many cases, the GLAMs have then made these digital collections available online. One limitation of the digitisation process is that it is very resource intensive, and as a result, most institutions have an ongoing, rolling digitisation programme (Denbo et al., 2008). This continuous digitisation process means that any curated online presentation of the collections needs to be continuously updated, creating an ever-growing digital environment.

Where these digital collections have been made available online, they generally consist of images of the items together with meta-data describing the item. The meta-data are typically drawn from the institution's own catalogues. In the case of libraries, these catalogues are generally created to enable

access to the holdings by the reader, however in galleries, museums, and archives the catalogues focus primarily on providing museum staff or professional researchers (Eklund, 2012; Vane, 2020) with details of the artefacts such as provenance, descriptive, and organisational information (such as dates, condition, material, style, rights, acquisition, genre, …). These meta-data were generally created when the institution acquired the object, and while in the digitisation process, the data are sometimes cleaned and standardised, Agirre et al. (2013) show that the meta-data of items are often limited and incomplete. This represents less of a technical issue for off-the-shelf faceted search systems, which easily deal with limited or missing data, although lower meta-data quality will obviously impact how successfully users can use the search system. However, for more complex interfaces that go beyond search, preprocessing the data-set is necessary. This preprocessing ranges from simpler tasks, such as normalising spellings or date formats, to automatically structuring the collections, where no or no consistent structuring is available.

Accessing the Collections

The GLAM websites that house the online collections also provide other types of information, including information about the institution, how to visit the physical institutions, selected items from the institution's holdings, and potentially an online gift shop. Due to GLAMs continued focus on their physical spaces, these other information services tend to receive the majority of attention, with the full collections access often treated as more of an afterthought.

Initially, where collections were made available, the interface for accessing the collection tended to be either a single white search box (basic search) or a set of search boxes, where each supported search in a specific meta-data field (advanced search). These search interfaces provided a fast and efficient entry point into the collection for the users familiar with either the collection itself or with the kind of data the collection contained. These users generally have a high degree of both specific and general CH

knowledge (Academics, Museum Staff, Research professionals, …) and typically have a particular information need, which enables them to successfully convert their need into the appropriate search terms to find what they are looking for (Marchionini, 2003; Falk, 2009; Skov and Ingwersen, 2008).

For users who have lower levels of CH knowledge (Marchionini, 2003; Falk, 2009; Walsh et al., 2020) or have a less focused information need (Casual users, General public, Non-professional users, …), the blank search box represents a significant barrier to accessing the collection (Belkin, 1982; Whitelaw, 2015). For online collections, this is a particular problem, as Walsh et al. (2020) showed that the majority (almost 70%) of a national museum's online audience was from this lower CH knowledge group. Additionally, users from this group are less likely to be frequent visitors (most are first-time visitors) and have a strong preference for browsing based access. The lack of prior experience means that these users need more help and information when getting started with the collection. Without this supporting information, this group of users is likely to give up and move on relatively quickly (Hall et al., 2012b).

To help those users who have less knowledge of the collection or who have less clear information needs, White (2009) suggested exploratory search systems. The most common interface for supporting exploratory search is the faceted search interface mentioned earlier. The advantage of the faceted search interface is that, in addition to the search box, for a selection of the collection's meta-data fields, the interface shows the user a choice of the most common values. Instead of being required to enter a search term, the user can select a value from the facet list and see results for that value. The exact facets used depend on the available meta-data, but commonly available facets include dates, locations, categories, materials, and techniques. Letting the user select from these facets reduces the chance of getting a zero-result (Hearst, 2006; Russel-Rose and Tate, 2012), which particularly aids non-expert users, who can, in that way, learn what search terms will lead to results.

The main limitation of faceted search interfaces are that the number of different values that can be

shown in the facets is restricted by the space available in the interface (Lang, 2013). This is problematic for collections access, as the heterogeneous nature of DCH collections means that every facet tends to have a long tail of values that do not occur very frequently. As faceted search interfaces will generally only show 10 or 20 of the most common values, none of these infrequent values will be accessible through the facet interface and thus remain undiscoverable to the user. Nevertheless, faceted search systems represent the most common interface for collection level access in DCH.

From Searching to Browsing

Faceted search, while assisting users with determining appropriate search keywords and thus reducing the barriers to using these interfaces for non-expert users, is still designed around search, even though non-expert users prefer browsing-based interfaces (Walsh et al., 2020). When developing browsing-based interfaces, there are two main requirements. First, they need to provide an initial overview of the collection (Greene et al., 2000; Hibberd, 2014). Second, through the browsing and visualisation interface, they must support the user in exploring the collection and gradually building up a more detailed understanding of the content (Giacometti, 2009; Mauri et al., 2013).

Where a browsing-based interface is provided by the GLAM institution, currently the most common interface is the manually curated digital exhibition (Coudyzer and van den Broek, 2015), although some libraries also provide browsing access via existing classification systems such as Dewey Decimal Classification (Vizine-Goetz, 2009; Lardera et al., 2018). The manually curated exhibitions generally provide an overview of the collection and very detailed information on a curated set of high-importance items. The main issue with these is that they require manual curation and creation and this does not scale to the amount of data in modern collections and struggles to keep up with the ongoing digitisation processes. While there have been attempts at automatically combining explanatory text with items selected from the collections (Hall et al., 2012b) to overcome the scaling limits and dynamically create

exhibitions on a topic selected by the user, the results have not seen widespread uptake.

Rich Prospect Browsing and Generous Interfaces

Instead the focus has been on more informative, supportive, and scalable browsing interfaces labelled as either “Rich Prospect Browsing” (Ruecker et al., 2011) or more recently “Generous Interfaces” (Whitelaw, 2015). The two terms were developed independently, but essentially describe the same core idea of providing an interface that does not require a-priori expertise of either the interface or the collection in order to use the interface successfully.

The driving principle behind Ruecker et al. (2011)’s rich prospect browsing is Schneiderman (1996)’s interaction pattern of “overview first, zoom and filter, then details on demand.” The core requirement of rich prospect browsing is that upon entering the collection, the user should be provided with a meaningful representation of every item in the collection. The user should then be able to manipulate this representation to explore the collection.

An example that tries to get as close to the meaningful representation of every item is Foo (2016)’s interface for a public-domain release of about 178 000 items from the New York Public Library. The initial screen shows a grid with a small thumbnail of every item, as expected from a rich prospect browsing interface. Users can click on the images to get more detail about them and navigate between them. The images on the initial screen are organised by time, but can also be arranged by genre, collection, or colour. The big question the interface raises is whether the tiny thumbnails, each thumbnail is only a few pixels large, represent a meaningful representation of the individual items, as, apart from colour differences, it is very difficult to discern anything about the items.

Whitelaw (2015)’s generous interface is a slightly more generic concept, which is less prescriptive regarding specific interface elements. A generous interface should also provide an initial view of the collection. Unlike with rich prospect browsing, the assumption is that the initial screen shows a sample

drawn from the collection, rather than all the items. The sample provides the starting point and clues that assist the user in then exploring the collection.

An interesting example of a generous interface is Coburn (2016)'s "Collections Dive", using items from the Tyne and Wear museum's archive (http://www.collectionsdivetwmuseums.org.uk/). Here the user is initially presented with a random sample of related items and can then, by scrolling down, request more items. Depending on the speed at which the user scrolls, the additional items are more (slow scrolling) or less (fast scrolling) similar to the previously visible items. In this way, the user can explore the collection. The user can then also select items to see further details about them. As Speakman et al. (2016) show, the interface is very engaging, but because the user can only scroll and has no control over what kind of items are shown next, only how similar they are to what they saw previously, it does not achieve extended engagement.

Visualisations for Browsing

The two interfaces nicely demonstrate that to develop browsing-based interfaces that scale to the size of current GLAM collections, the display of items has to be augmented with controls that allow the user to move between different parts of the collection. The most common approaches to this are visualisations and hierarchical navigation structures. Of the two, visualisations are used more frequently and Windhager et al. (2018) provide a detailed overview of the possible visualisation methods. Common visualisation methods include timelines, spatial (map) displays, network diagrams, and word clouds. The main advantage of these is that while they may require the collection's meta-data to contain specific fields (for example time or location information), if these requirements are met, they can be used to visualise and provide access to any kind of collection.

Timeline visualisations work by showing the user a horizontal or vertical timeline and the items in the

collection are then placed on this timeline based on their date(s). By interacting with the timeline the user can easily restrict the items they see based on the time-period they are interested in. Glinka et al. (2017) demonstrate the use of a timeline as the primary method for organising and browsing a collection, where all or at least the vast majority of items have temporal meta-data. Their timeline visualisation shows not only when the items in the collection were created, but also shows how many items can be found at each point in time, providing additional guidance to the user. The interface also includes the functionality to restrict the timeline by keyword. This illustrates one limitation of timelines, which is that time on its own is a very limited organisational principle, and that an additional structuring principle is often required. While timelines are usually shown as linear features, Hinrichs et al. (2008) demonstrate that other displays are also possible, using concentric “tree-trunk” visualisation that represents a range of time periods.

Düring et al. (2015) demonstrate the use of network diagrams as a visual interface to a collection. In a network diagram, the items and meta-data values are shown as nodes, with edges between items and their meta-data values. DCH collections are generally well suited to network diagrams, as items often share where they were created, who created them, or what kind of thing they are, predisposing them to a network display. The power of network diagrams is that they allow for very efficient horizontal navigation through the collection. At the same time, the main limitation of network diagrams is that they do not scale that well. In particular, as the number of edges increases, it becomes difficult to visually distinguish which nodes are linked, as the network diagram degenerates into a black mess of lines.

Spatial displays are the most varied visualisation method. In the most common case, a 2-dimensional map is used to visualise the spatial meta-data (Simon et al., 2016). The advantage of this kind of map is that the user can zoom out to see an overview over the collection and zoom back in to see individual items. They can also easily be combined with a temporal visualisation. The most significant limitation of maps is that they require that the items have spatial meta-data and that it is in a computer-readable

form that gives an exact location, as current web-based maps cannot handle vague spatial information. Where spatial meta-data is only available as complex natural language descriptions such as “found next to the river Nile” or very imprecise descriptions such as “printed in Germany”, current map interfaces are not able to represent these locations accurately or at all, making these items inaccessible via the map visualisation.

The 2-dimensional map can also be used to display other information. For example, Descy (2009) describes the use of map interfaces to visualise search result clusters. Similarly, Hall and Clough (2013) present an interactive map visualisation that enables the exploration of a hierarchical structure used to organise a collection of about 500k items taken from Europeana. In that interface, the elements on the map no longer represent real-world geography, but instead a virtual geography, where the map elements are concepts from the hierarchy. This combines the value of a hierarchy for structuring things and the map as a known interface for exploring the world.

Word clouds (Feinberg, 2010) represent another visualisation technique. A word cloud is generated by extracting keywords from all items in the collection and then displaying the most frequent keywords (Sinclair and Cardew-Hall, 2008; Wilson et al., 2012). Users can then select keywords in the word cloud and see the items associated with that keyword. Various visual modifications, such as font size or colour, can be applied to the displayed keywords and used to provide additional information such as the relative frequency of the displayed keywords (Lohmann et al., 2009), guiding users in their exploration. They are relatively similar to the facets in a faceted search system and share many of their advantages, such as ease of generation, and disadvantages, such as not scaling well to the large number of diverse keywords common in heterogeneous DCH collections.

Browsing Navigational Structures

The alternative to the use of visualisations is the provisioning of a navigation structure. This is generally

provided in the form of a hierarchy or taxonomy of concepts. The hierarchical structure can either be used directly for browsing, displaying the hierarchy as a tree or can be visualised in another way, for example, tag-clouds or a map, as demonstrated in the PATHS project (Hall et al., 2014). The difficulty with these is that they require the items to be mapped into an existing hierarchy, either manually or automatically. Libraries are often at an advantage in this, as their collections tend to use a standardised classification hierarchy, which can be browsed on its own or integrated into the search process to enable a mixed search and browse interface (Golub, 2018).

Organising Collections

For many browsing-based interfaces, the items in the collection need to be placed into an organisational structure of some kind. The structure then provides the links that the users use to browse the collection. Methods for undertaking this curation of items can be classified along three primary axes: manual vs. automatic methods, flat vs. hierarchical structures, and purely data-driven methods vs. methods that include external data.

Manual Organisation of Collections

Manual curation (Rao et al., 1995) of an organisational hierarchy is likely to produce the highest quality and most domain-specific curation of the collection. However, it is also the most resource-intensive approach and in general for most GLAM institutions, not a viable approach, even though there is work ongoing on improving tool support for the process (see for example Rehm et al. (2019)). Libraries represent an exception in this case, as most use a standardised classification scheme, such as Dewey Decimal Classification (DDC) (https://www.oclc.org/en/dewey.html), Universal Decimal Classification (http://www.udcc.org/), Library of Congress Subject Headings (https://id.loc.gov/authorities/subjects.html), or BISAC (Book Industry Standards and Communications)

Subject Headings (https://bisg.org/page/BISACEdition) to organise their collections. An in-depth discussion of these schemes is outside the scope of this chapter, but they are generally hierarchical in nature and as such can be used to support a browsing-based interface.

When it comes to manually adding a hierarchical classification scheme to collections that do not use one or do not consistently use one, rather than relying on in-house expertise, crowdsourcing is often seen as a solution to scaling up the process (Sun et al., 2015; Yagui et al., 2019), but as Yagui et al. (2019) show, evaluation and input by domain experts are still required, which means that while the resource bottleneck is reduced, it is not removed.

Automatic Organisation of Collections

Automatic methods for organising collections offer a way of overcoming this bottleneck. At the simplest level, these methods employ basic clustering algorithms to create a flat partition of the collection (Hall et al., 2012a). The limitation of such a pure flat partitioning is that for larger collections, the number of partitions quickly grows to such a degree that navigating these becomes difficult. Algorithms that organise the collection into a hierarchical structure offer to address this. Such algorithms can either be purely data-driven or be based on an existing hierarchy or taxonomy.

The pure data-driven algorithms can use a variety of methods including hierarchical Latent Dirichlet Allocation (LDA) (Griffiths et al., 2003), multi-branch clustering (Liu et al., 2012), co-occurrence (Sanderson and Croft, 1999), or word embeddings (Luu et al., 2016). The advantage of these algorithms is that they do not require any external data and will place all of the concepts and items into a hierarchy. The downside is that while the arrangement of the concepts in the generated hierarchy will be "correct" as far as the algorithm is concerned, the resulting hierarchy is not guaranteed to match what people would consider an appropriate hierarchy. Depending on the algorithm, adding new data may also lead to significant changes to the hierarchy structure, which makes it harder for users to refind

things after such a change. The pure data-driven approaches are also not capable of generalising concepts, so would, for example, be unlikely to group plates and cups under the concept of crockery, unless that concept also existed in the meta-data.

Using existing hierarchies addresses these issues and in previous work a range of hierarchies have been used, including WordNet (Navigli et al., 2003; Stoica et al., 2007), Wikipedia (Milne et al., 2007; Fernando et al., 2012), and DDC (Lin et al., 2017). Other approaches have combined concepts drawn from multiple, existing hierarchies including Library of Congress Subject Headings, DBPedia, Wikidata, or the Art and Architecture Thesaurus (Hall et al., 2014; Charles et al., 2018). While the use of existing hierarchies ensures that the structure follows patterns that are closer to people's expectations of such a hierarchy, if the concepts used in the collection do not exist in the chosen hierarchy, then the affected items cannot be mapped into the hierarchy. The algorithm we present in this chapter addresses these issues and by using a mix of pure data-driven hierarchy creation together with an existing hierarchy (the Art and Architecture Thesaurus) to create an organisational hierarchy that is based on the existing hierarchy, but also includes more specific concepts derived from the items' meta-data.

The Digital Museum Map

The Digital Museum Map (DMM) addresses some of the issues raised above, in particular providing an interface that is amenable to the kind of open-ended browsing discussed earlier, that scales to large collections, and that requires only minimal human input into the curation and visualisation process (https://github.com/scmmmh/museum-map, https://museum-map.research.room3b.eu/). The core idea behind the DMM's exploration interface is that the museum floor plan is an established and well-known method for exploring a physical museum and the DMM uses the same visualisation, but this time for a virtual museum, that is automatically generated for a specific collection. Naturally such an interface is more suited for museums' and archives' collections and for a library-shelves-inspired interface see Hall

(2014).

The DMM is a complete redevelopment of the initial algorithms and interface (Hall, 2018), based on the experience of developing and deploying the initial DMM. In particular the new algorithm scales more easily and is less tailored to the collection used in the development process. The browsing interface has also been revised to take into account informal observations of how non-specialist users interacted with the initial DMM. It does, however, retain the main metaphor of exploring a physical museum, with different rooms, floors, and buildings.

Data

The initial version of the DMM was based on a selection of objects from National Museums Liverpool. For the new version presented here, the DMM uses a collection of 14351 objects from the Victoria & Albert (V&A) museum's digital collection (https://collections.vam.ac.uk/). The items were acquired using the V&A's API and then loaded into a relational database for all further processing. The collection is representative of the kind of heterogeneity that characterises most GLAM collections and contains amongst other things pottery, paintings, prints, clothing, jewellery, designs for various types of objects, sculptures, and photographs.

Each item has a number of meta-data fields attached to it. The ones that are relevant to the DMM are the "object" field, which contains each item's primary classification (jug, earring, ...), the "concepts", "subjects", "materials", "techniques", "year_start", and "year_end" fields, which are used in the group generation process, and the "title", "description", "physical_description", and "notes" fields, which are used to determine similarity between items. We only use the free-text fields for the similarity calculation, as these provide the most nuanced description of the items. All of these are also displayed in the interface.

The DMM Generation Process

The DMM generation process is shown in Figure 1. It's aim is to transform the unstructured set of item meta-data into a set of "rooms", distributed over one or more "floors", where each "room" contains a group of similar items. It starts with an initial processing of all the items in the collection, which extends the meta-data with values required to organise the items (Classification augmentation & Similarity Vector Generation). The processed items are then grouped (Basic Group Generation) and the groups arranged into an hierarchical structure (Parent Group Generation & Large Group Splitting). Finally, the groups are placed into the floor layout (Room Layouting), which is what the users will then use to explore the collection.

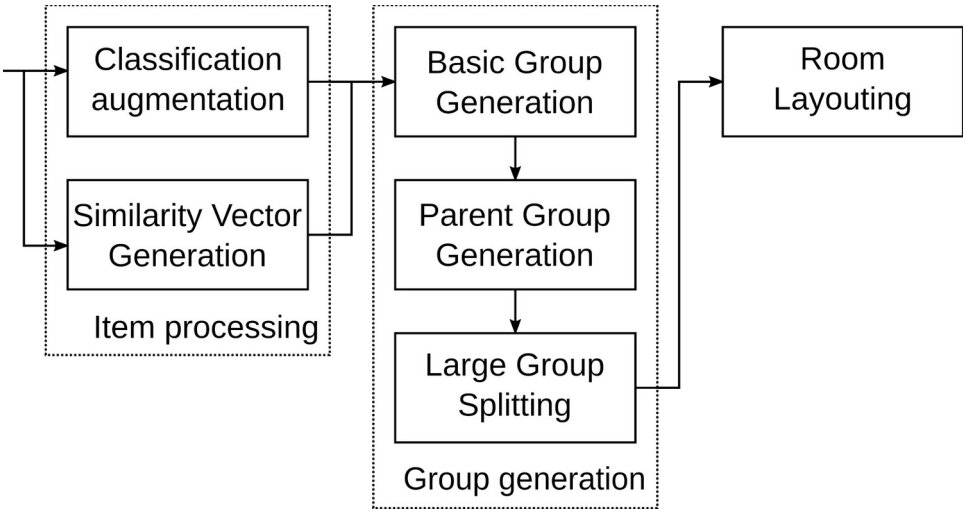

Figure 1: The DMM generation process. The two steps in the item processing stage can be parallelised, but the remainder of the processing workflow is linear.

Item Processing

In order to organise the items into cohesive groups based on their meta-data, the DMM has to generate two pieces of information for each item. The first is the hierarchy of classification values used to create the groups. The second is a similarity vector, which is used in the group creation and item layouting steps.

Classification Augmentation

Each item's primary classification is defined by the value in the "object" field (e.g. "drawing of a wedding dress"). The values are often very specific to the individual item and if only those values were used to group the items, then a significant fraction of the collection would remain ungrouped or the resulting groups would only include a few items. To ensure the generated groups have an appropriate size, from past experience this lies between 15 and 120 items, the DMM initially employs natural language processing (NLP) techniques to extract more generic concepts from the item meta-data and then pulls in additional hierarchy information from the Getty Art and Architecture Thesaurus (AAT) (https://www.getty.edu/research/tools/vocabularies/aat/) (Petersen, 1990).

To extract the more generic concepts the DMM uses a series of heuristics that extract more generic concepts from the existing meta-data (Table 1). These are applied greedily in the order shown in Table 1 and recursively to the extracted concepts. For example, for the primary classification "drawing of a wedding dress", the "A of B" heuristic would be the first one that applies and "drawing" and "wedding dress" would be the two extracted concepts. The heuristics are then applied recursively to both the extracted concepts, resulting in the concept "dress" being extracted from "wedding dress" (using the final "A B" heuristic). The more generic concepts are added to the primary classification value to create a list of classification values ["design for a wedding dress", "wedding dress", "design", "dress"].

Table 1: The NLP heuristics. The "Heuristic" column shows the heuristic pattern, where A and B are one or more words. The "Extracted" column shows the extracted concepts and the order in which they are added to the classification value. The final column shows an example for each heuristic.

Heuristic	Extracted	Example
A for B	B, A	Design for brooch -> brooch, design

A (B)	A, B	Cap (headgear) -> cap, headgear
A with B	A, B	Cup with stand -> cup, stand
A of B	B, A	Drawing of a dress -> dress, drawing
A from B	B, A	Page from a sketchbook -> sketchbook, page
A & B	A, B	Cup & saucher -> cup, saucher
A and B	A, B	Cup and saucer -> cup, saucer
A, B	A, B	Bowl, fragment -> bowl, fragment
A or B	A, B	Screen or balustrade -> screen, balustrade
A B	B	Tea cup -> cup

While the NLP augmentation extracts additional information from the classification value, higher level classification concepts need to be added from an external source and we use the AAT for this purpose. The AAT contains over 71k concepts (with over 400k terms for these concepts), arranged into eight faces (associated concepts, physical attributes, styles and periods, agents, activities, materials, and objects) to support catalogue and retrieve items from art, architecture, and other visual cultural heritage. In addition to a search system it provides a search API (http://www.getty.edu/research/tools/vocabularies/obtain/download.html) and each term in the augmented classification list is sent to the API and the parent hierarchy information extracted from the result. The DMM uses caching to reduce the number of requests sent to the AAT and to improve processing speed.

Where concepts are ambiguous and thus there are multiple hierarchies, the concepts from all hierarchies are added to the classification list. We then apply some post-processing on the concepts from the

hierarchies. Duplicate concepts are only added once to the classification list. Concepts that end in "genre" or "facet" have that suffix stripped and are then added to the classification list, if not already present. Finally, purely organisational sub-division concepts such as "X by Y" (for example "containers by function") are filtered. This is because while they help with organising the AAT, they are not appropriate labels for use in the DMM. The resulting augmented list of classification values is added to the item's meta-data in an additional field.

Similarity Vector Generation

When creating the groups and when arranging individual items in a group for display, the items are ordered so that similar items are placed together. There are a range of similarity measures that could be used, but here we use a very simple approach introduced in Aletras et al. (2013). The similarity measure first creates a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) for all items in the collection and then calculates the topic vector for each item. Item similarity can then be calculated using cosine - similarity between pairs of topic vectors.

The LDA model is created based on the contents of the "title", "description", "physical_description", and "notes" fields. For each item the four fields are concatenated and then tokenised using the open-source NLP library Spacy (https://spacy.io). Punctuation and space tokens are filtered and the remaining tokens stored in the item's meta-data. Using these tokens, we generate a 300 topic LDA model using the Gensim topic-modelling library (Rehurek and Sojka, 2010). We use Gensim's default dictionary extremes filtering settings of removing all tokens that occur less than 5 times or in more than 50% of all items. However, we use all remaining tokens, rather than the default setting of keeping only the 100000 most frequent tokens. This is necessary as some items have very little text and thus very few tokens. Filtering infrequent tokens would lead to these items not having any topics assigned to them. We then calculate the topic vector for each item and store the resulting vector with the item.

Group Generation

The second step is the creation of groups, where each group has between 15 and 120 items in it. The limits are based on experience, but are fully configurable. In particular, the lower boundary can be increased, if the collection is more homogenous and there are fewer uncommon classification values. As shown in Figure 1, generating the groups consists of a number of steps. First, the basic groups are generated and post-processed, then they are arranged into a hierarchy, which is then cleaned, before splitting any remaining, large groups.

Generating the Basic Groups

Unlike the initial DMM, which took a top-down approach, in the current DMM we take an iterative, greedy, bottom-up approach to grouping the items. The generation is based on the lists of classification values created earlier during the item processing. In each iteration the algorithm first calculates the frequencies for all classification values of those items that have not yet been assigned to a group. We filter all values that occur less than 15 times and from the remaining values the algorithm selects the value with the fewest occurrences. A new group is created for this value and all unassigned items that have that value are assigned to the new group. The algorithm then moves on to the next iteration, until no values remain that occur at least 15 times.

The reason for selecting the classification value with the fewest occurrences is that this is likely to create more cohesive and size-wise more displayable groups of items. At the same time the greedy, bottom up approach can lead to a situation, where some items are not allocated to any group, even though they share a concept with at least 15 other items. This is because the other items may have been allocated to another group, based on another concept, reducing the number of unallocated items with the first concept to below 15. For the current collection, the algorithm fails to allocate 101 (0.7%) items. A manual analysis of these items indicates that the majority fall into three categories: items with very specific classifications that neither the NLP nor the AAt processing can group (e.g. “copy of the hedda”),

items where there are only one or two of that type in the collection (e.g. “gun”), or concepts for which the AAT API does not return a result (e.g. “Tea-urn”).

For future work it may be worth considering whether the similarity vectors could be used to assign the unallocated items to groups. Alternatively, the items may be placed in the “corridors” of the visualisation or simply grouped together in an “Odds & Ends” group.

In the original source data, the classification value is generally in the singular form, while the AAT generally uses the plural form. Because the data-driven generation algorithm does not take this into account, there is the potential for one group to exist with the singular form of a concept and a second one with the plural form. In a post-processing step, we identify all singular-plural pairs, re-assign the items from the singular to the plural form group, and delete the singular form group. We retain the plural form, as this form is more appropriate when labelling rooms.

Adding Parent Groups

The DMM does not use the hierarchical structure for navigation by the user. However, when organising the groups into the 2-d layout, we want related groups to be placed close to each other and to achieve that, we need to organise the groups into a hierarchy. In practice, because the AAT is organised into eight facets that do not have a shared parent, there will not be a single hierarchy, but initially up to eight hierarchies.

To create these hierarchies, we first determine the AAT hierarchy for each group’s concept. Then for each concept in the hierarchy, a group is created, unless that group already exists and the parent-child relationship between the hierarchy concepts is set. If no match is found in the AAT, then the NLP augmentation, as described earlier, is applied to the group’s concept. If for any of the new concepts identified by the NLP augmentation, there is already a group, then the current group is added as a child under that group. If no group exists for any of the concepts identified by the NLP

augmentation, then each concept is looked up in the AAT. The current group is then assigned to the first hierarchy that is found in the AAT.

After the hierarchies have been created, two post-processing steps are applied. First, any groups that have only a single child group and no items are pruned, as they don't add any useful information. Second, for any group that has both child groups and items, the items are added into a new group that has the same label as the original group and the new group is added as a child to the original group. This ensures that items are only placed in the leaf nodes, which makes the layouting algorithm simpler.

Splitting Large Groups

At this point there will be a small set of topics with more than 120 items. For the room layouting we have defined 120 items as the maximum number of items per room. These groups thus need to be split into smaller sub-groups, before they can be placed into rooms. When splitting these we treat groups with between 120 and 300 items separately from those with over 300 items. For groups of the first type, we first attempt to split them by time and if that does not produce a split, then they are split by similarity. For the larger groups, we first attempt to split them by one of four attributes ("concepts", "subjects", "materials", "techniques"). If that does not work, then we attempt to split by time and if that does not work, then by similarity.

When splitting by attribute, the approach is similar to that used when generating the basic groups. First we calculate the frequency of all attribute values, filtering those attribute values that occur less than 15 times or cover more than two-thirds of the items, as neither are appropriate for splitting the group. Next, we check that the remaining attribute values cover at least 90% of the items. We then sort the attribute values by increasing frequency and then iterate over the sorted attribute values, assigning items to the first value that they have. The new groups are all labelled with the label of the original group plus their attribute value. Finally, any items that are not allocated to an attribute value group are

placed into a new group with the same label as the original group.

If no attribute can split the group, or if the group has less than 300 items, then an attempt is made to split the group by time. In order to split the group by time, at least 95% of items in the group must have a temporal attribute set. Then the number of items per year is counted and the earliest and latest year determined. If the time span defined by the earliest and latest year is greater than 10 years and less than or equal to 100, the group is split by decade. If the time span is greater than 100, it is split by century. In either case the items are then sorted by the temporal attribute and placed into decade or century bins. Where temporally adjacent bins have less than 100 combined items, the bins are merged. For each unmerged bin, a new sub-group is created, labelled by the name of the parent group and the time period it covers. Any items that do not have a temporal attribute are placed in a new sub-group, with the same label as the original group, as is done when splitting by attribute case.

Finally, if neither attribute nor temporal splitting are possible, then the large groups are split into smaller groups using the item similarity. In this approach the items are first sorted using a greedy algorithm. The first item is copied from the input list to the sorted list and set as the current item. Then the current item's similarity to all unsorted items is calculated, using cosine similarity of the topic vectors calculated earlier. The most similar item is added to the sorted list and set as the current item. This is repeated until all items have been sorted. The sorted list is then split evenly into bins, with the number of bins calculated as the number of items in the group divided by 100. For each bin a new sub group is added, with the same label as the original group.

Overall, for the 14351 items, the algorithm generates a total of 390 groups in 7 hierarchies. Of these 286 are leaf groups, which contain items, and which are used in the next layouting step.

Room Layouting

Unlike the group generation, which is completely automatic, the room layouting requires some manual

input: a 2-d floor layout, a list of rooms with their maximum sizes, and the order in which the rooms should be processed. Each room has a set maximum number of items that it can contain. In the hierarchies created in the previous step, only the leaf nodes contain items, so only those are fed into the room layouting algorithm. To ensure that related leaf nodes are placed as closely as possible, the list of leaf groups to layout is generated by walking the trees in a depth-first manner.

To assign the groups to rooms, the algorithm loops over the list of rooms. It then checks if the next unassigned group has less items than the room can contain. If this is the case, then the group is assigned to the room. Then the algorithm repeats this process for the next unassigned group. If the next unassigned group has more items than the current room can contain, then that room is left “unused”.

If, after all rooms have been assigned one or more groups, there are still unassigned groups, a new “floor” is created with a new list of rooms and the assignment algorithm restarts with the new list of rooms. This process is repeated until all groups have been assigned to rooms. In the case of the example collection, the 286 leaf groups are assigned to 286 rooms, spread over 5 floors (an example of a single floor’s layout is shown in Figure 2). The rooms, spread across the floors, together with the assigned groups are then used by the browsing interface to let the user explore the collection.

Browsing Interface

Figure 2 shows the main floor plan interface the users use to explore the collection. As the screenshot shows, because the floorplan is provided, the resulting layout looks very natural and similar to a physical museum’s layout. Using the arrows next to the floor number, the user can move between the different floors. When moving the mouse over the floorplan, the user is shown a sample image taken from the items in that room. We are currently experimenting with how many samples to show and how to generate a brief textual summary of the items, to give the user a better idea of what the room contains.

When viewing a room, the items are arranged based on the similarity vectors calculated earlier and sorted using the greedy algorithm described earlier. To navigate between the rooms the user can always show the floorplan. Additionally, where the map shows “doors” between the current room and another room, a link is shown in the room view, allowing the user to move from room to room, exploring the museum. By clicking on a single item, the user can show a relatively standard detail-view of the item and its meta-data.

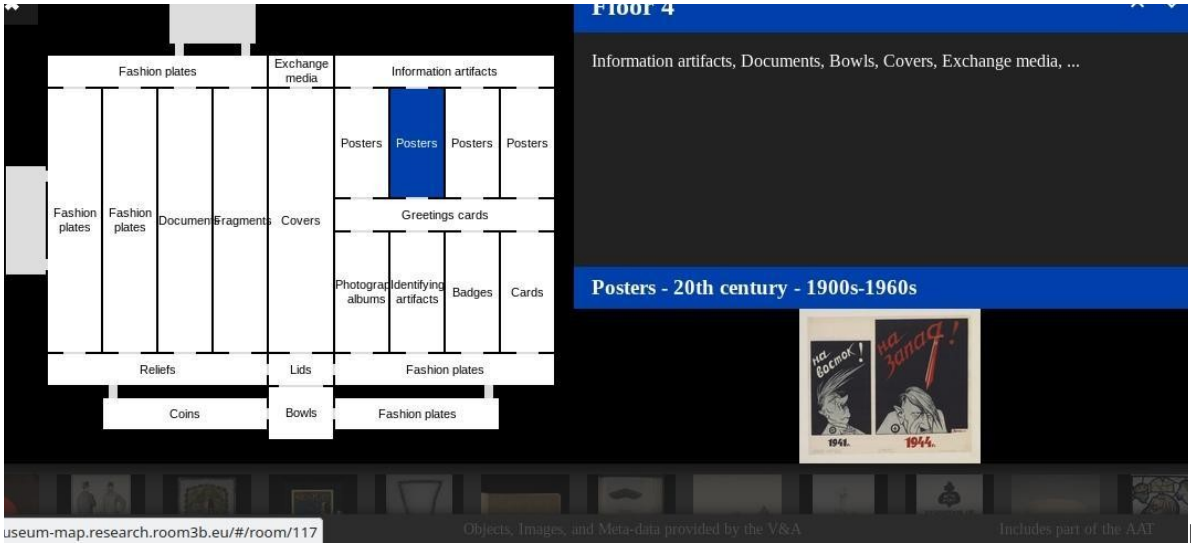

Figure 2: The Museum Map browsing interface showing the floorplan interface for exploring the collection in the foreground and the grid of items for a single room in the background. The currently visited room is highlighted, as is the room the user has moved their mouse over. For the room the user’s mouse is hovering over, a preview is shown in the bottom-left corner.

Conclusion

The digital cultural heritage collections created through the digitisation of GLAMs’ holdings have made available vast numbers of digital items to everybody. However, non-specialist users generally lack the expertise needed to access these successfully. Various approaches have been made to open the digital collections of GLAMs to wider audiences. From improving simple search boxes by adding facets to the search interface, all the way to browsable, visual interfaces under the label of rich prospect browsing or

generous interfaces designed to overcome the inadequacies of the search-only and faceted search interfaces (Vane, 2020). However, none of the browsable, visual interfaces have seen any widespread uptake, in part because they are time-consuming and expensive to design and develop and are usually built for one specific collection (Haskiya, 2019). As a result, they tend not to be applied to collections other than the one they were made for.

The open-source Digital Museum Map (DMM) system presented in this chapter addresses this limitation by providing a generous, browsing-based interface that can be applied to any collection and that generates the interface with minimal manual input (https://github.com/scmmmh/museum-map, https://museum-map.research.room3b.eu/).

This chapter illustrates that while there has been some work looking at moving beyond search as the interface for exploring large DCH collections, the area is still in its infancy and has a large number of open questions that need investigation, some of which are discussed below.

The biggest is how to evaluate the success of an interface designed for open-ended exploration. As such an interface is designed for users with no or at most a very vague information need, how does one judge to what degree the interface has worked for them? Is it a success if the users are engaged with the system? If they spend longer exploring than with standard search systems? If they return at a later point? If they show an increase in knowledge of some kind? Developments in this area are particularly crucial, as they will enable comparisons between solutions, transforming research in this area from the current more exploratory approach into a more formal structure.

Another major direction for future work highlighted by this chapter is how to generate an overview or summary of the items in the collection. Such an overview would always be based on a sample drawn from the collection and the sample would have to be both representative of the whole collection and also enticing enough that it engages users and encourages exploration. This requires developing an

understanding of what makes a “good” sample, what makes an “interesting” item to sample from the collection, and how these items should be tied together and presented to the user. It is also necessary to investigate whether such an overview sample should be static, change with each viewing, or mix static and dynamically selected items.

How to make browsing interfaces scale not just to tens of thousands of items, but to millions of items, is another open research question for browsing systems. The DMM interface enables one possible approach, which is grouping together the “floors” into “wings”, “galleries”, or “museums” to create a navigation hierarchy, allowing the visual metaphor to scale. Scaling to this size also requires the addition of some kind of horizontal browsing support, most likely in the form of recommendations. While there has been much work on recommendation in general, little is known about what type of recommendation users would like to see in a DCH context, in particular how interested and open users are towards recommendations that have the potential to surprise them.

While there has been some work on integrating search and browse functionality into one combined interface (Hall, 2014; Golub, 2018), in general they are often treated as separate interaction modes. How to integrate the two more deeply and allow for the user to seamlessly switch between them remains an open question.

Finally, in addition to evaluating the system as a whole, we are also in the process of setting up evaluations for individual parts of the DMM, including the classification augmentation, similarity calculation, and group generation, in order to develop an in-depth understanding of how to create a high-quality structure for exploration.

References

Agirre, Eneko, Nikolaos Aletras, Paul Clough, Samuel Fernando, Paula Goodale, Mark Hall, Aitor Soroa,

and Mark Stevenson. 2013. "Paths: A system for accessing cultural heritage collections." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 151-156. Association for Computational Linguistics.

Aletras, Nikolaos, Mark Stevenson, and Paul Clough. 2013. "Computing similarity between items in a digital library of cultural heritage." *Journal on Computing and Cultural Heritage (JOCCH)* 5, no. 4: 1-19.

Belkin, Nicholas J., Robert N. Oddy, and Helen M. Brooks. 1982. "ASK for information retrieval: Part I. Background and theory." *Journal of documentation*.

Biedermann, Bernadette. 2017. "'Virtual museums' as digital collection complexes. A museological perspective using the example of Hans-Gross-Kriminalmuseum." *Museum Management and Curatorship* 32, no. 3: 281-297.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3: 993-1022.

Borgman, Christine L. 2010. "The digital future is now: A call to action for the humanities." *Digital humanities quarterly* 3, no. 4.

Bates, Marcia J. 1989. "The design of browsing and berrypicking techniques for the online search interface." *Online review*.

Bates, Marcia J. 2007. "What is browsing-really? A model drawing from behavioural science research.". *Information Research* 12, no. 4. [Available at http://InformationR.net/ir/12-4/paper330.html]

Charles, Valentim, Hugo Manganinhas, Antoine Isaac, Nuno Freire, and Sergiu Gordea. 2018. "Designing a multilingual knowledge graph as a service for cultural heritage–some challenges and solutions." In *International Conference on Dublin Core and Metadata Applications*, pp. 29-40.

Coburn, John. 2016. "I don't know what I'm looking for: Better understanding public usage and behaviours with tyne & wear archives & museums online collections." *MW2016: Museums and the Web* 2.

Coudyzer, Eva, and Anna van den Broek. 2015. "Europeana Exhibitions: A Virtual Trip through Europe's Cultural Heritage, Interview with Anna van den Broek." *Uncommon Culture*: 47-54.

Denbo, Seth, Heather Haskins, and David Robey. 2008. "Sustainability of Digital Outputs from AHRC Resource Enhancement Projects." *Arts and Humanities Research Council, available at: www. ahrcict. rdg. ac. uk/activities/review/sustainability08. pdf (accessed 26/11/2020)*.

Descy, Don E. 2009. "Grooker, KartOO, Addict-o-Matic and More: Really Different Search Engines." *TechTrends* 53, no. 3.

Düring, Marten, Lars Wieneke, and Vincenzo Croce. 2015. "Interactive networks for digital cultural heritage collections-scoping the future of histograph." In *International Conference on Web Engineering*, pp. 613-616. Springer, Cham.

Eklund, Janice L. 2011. "Cultural objects digitization planning: Metadata overview." *Visual*

Resources Association Bulletin 38, no. 1.

Falk, John H. 2016. *Identity and the museum visitor experience*. Routledge.

Feinberg, Jonathan. 2010. "Wordle". *Beautiful Visualization Looking at Data through the Eyes of Experts*. O'Reilly Media, 37-58.

Fernando, Samuel, Mark Hall, Eneko Agirre, Aitor Soroa, Paul Clough, and Mark Stevenson. 2012. "Comparing taxonomies for organising collections of documents." *Proceedings of COLING 2012*: 879-894.

Foo, B. 2016. "NYPL-publicdomain/pd-visualization". *GitHub repository*, GitHub. https://github.com/nypl-publicdomain/pd-visualization

Giacometti, Alejandro. 2009. *The Texttiles browser: an experiment in rich-prospect browsing for text collections (Master's thesis)*. Edmonton, AB: University of Alberta.

Glinka, Katrin, Christopher Pietsch, and Marian Dörk. 2017. "Past Visions and Reconciling Views: Visualizing Time, Texture and Themes in Cultural Collections." *DHQ: Digital Humanities Quarterly* 11, no. 2.

Greene, Stephan, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. 2000. "Previews and overviews in digital libraries: Designing surrogates to support visual information seeking." *Journal of the American Society for Information Science* 51, no. 4: 380-393.

Blei, David M., Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2004. "Hierarchical topic models and the nested Chinese restaurant process." *Advances in neural information processing systems* 16, no. 16: 17-24.

Golub, Koraljka. 2018. "Subject access in Swedish discovery services." *KO KNOWLEDGE ORGANIZATION* 45, no. 4: 297-309.

Hall, Mark, Paul Clough, and Mark Stevenson. 2012. "Evaluating the use of clustering for automatically organising digital library collections." In *International Conference on Theory and Practice of Digital Libraries*, pp. 323-334. Springer, Berlin, Heidelberg.

Hall, Mark Michael, Oier Lopez de Lacalle, Aitor Soroa, Paul Clough, and Eneko Agirre. 2012. "Enabling the discovery of digital cultural heritage objects through wikipedia." In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 94-100. Association for Computational Linguistics.

Hall, Mark, and Paul Clough. 2013. "Exploring large digital library collections using a map-based visualisation." In *International Conference on Theory and Practice of Digital Libraries*, pp. 216-227. Springer, Berlin, Heidelberg.

Hall, Mark M. 2014. "Explore the stacks: A system for exploration in large digital libraries." In *IEEE/ACM Joint Conference on Digital Libraries*, pp. 417-418. IEEE.

Hall, Mark, Paula Goodale, Paul Clough, and Mark Stevenson. 2014. "The paths system for exploring digital cultural heritage." In *Proceedings of the Digital Humanities Congress 2012*. Studies in the Digital

Humanities. Sheffield: The Digital Humanities Institute.

Hall, Mark Michael. 2018. "Digital museum map." In *International Conference on Theory and Practice of Digital Libraries*, pp. 304-307. Springer, Cham.

Haskiya, David. 2019. An Evaluation of Generous Interfaces. EuropeanaTech Insight, 11. *Europeana.Eu*. https://pro.europeana.eu/page/issue-11-generous-interfaces#an-evaluation-of-generousinterfaces.

Hearst, Marti A. 2006. "Clustering versus faceted categories for information exploration." *Communications of the ACM* 49, no. 4: 59-61.

Hibberd, Georgina. 2014. "Metaphors for discovery: how interfaces shape our relationship with library collections." *The Search Is Over*.

Hinrichs, Uta, Holly Schmidt, and Sheelagh Carpendale. 2008. "EMDialog: Bringing information visualization into the museum." *IEEE transactions on visualization and computer graphics* 14, no. 6: 1181-1188.

Koch, Traugott, Koraljka Golub, and Anders Ardö. 2006. "Users browsing behaviour in a DDC-based web service: A log analysis." *Cataloging & classification quarterly* 42, no. 3-4: 163-186.

Lardera, Marco, Claudio Gnoli, Clara Rolandi, and Marcin Trzmielewski. "Developing SciGator, a DDC-based library browsing tool." *KO KNOWLEDGE ORGANIZATION* 44, no. 8 (2018): 638-643.

Lang, Bo, Xianglong Liu, and Wei Li. 2013. "The Next-Generation Search Engine: Challenges and Key Technologies." In *Recent Progress in Data Engineering and Internet Technology*, pp. 239-248. Springer, Berlin, Heidelberg.

Lohmann, Steffen, Jürgen Ziegler, and Lena Tetzlaff. 2009. "Comparison of tag cloud layouts: Task-related performance and visual exploration." In *IFIP Conference on Human-Computer Interaction*, pp. 392-404. Springer, Berlin, Heidelber.

Lin, Xia, Michael Khoo, Jae-Wook Ahn, Doug Tudhope, Ceri Binding, Diana Massam, and Hilary Jones. "Mapping metadata to DDC classification structures for searching and browsing." *International Journal on Digital Libraries* 18, no. 1 (2017): 25-39.

Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. "Automatic taxonomy construction from keywords." In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1433-144.

Luu, Anh Tuan, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. "Learning term embeddings for taxonomic relation identification using dynamic weighting neural network." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 403-413.

Marchionini, Gary, Catherine Plaisant, and Anita Komlodi. 2003. "The people in digital libraries: Multifaceted approaches to assessing needs and impact." *Digital library use: Social practice in design and evaluation*: 119-160.

Marchionini, Gary. 2006. Exploratory search: from finding to understanding. *Communications of the ACM* 49, no. 4: 41-46.

Mauri, Michele, Azzurra Pini, Daniele Ciminieri, and Paolo Ciuccarelli. 2013. "Weaving data, slicing views: a design approach to creating visual access for digital archival collections." In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pp. 1-8.

Mayr, Eva, Paolo Federico, Silvia Miksch, Günther Schreder, Michael Smuc, and Florian Windhager. 2016. "Visualization of cultural heritage data for casual users." In *IEEE VIS Workshop on Visualization for the Digital Humanities*, vol. 1.

Meehan, Nicole. 2020. "Digital Museum Objects and Memory: Postdigital Materiality, Aura and Value." *Curator: The Museum Journal*.

Milne, David N., Ian H. Witten, and David M. Nichols. 2007. "A knowledge-based search engine powered by wikipedia." In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 445-454.

Navigli, Roberto, Paola Velardi, and Aldo Gangemi. 2003. "Ontology learning and its application to automated terminology translation." *IEEE Intelligent systems* 18, no. 1: 22-31.

Petersen, T. 1990. *Art & architecture thesaurus*.

Rao, Ramana, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and George C. Robertson. 1995. "Rich interaction in the digital library." *Communications of the ACM* 38, no. 4: 29-39.

Rehm, Georg, Martin Lee, Julián Moreno-Schneider, and Peter Bourgonje. 2019. "Curation Technologies for Cultural Heritage Archives: Analysing and transforming a heterogeneous data set into an interactive curation workbench." In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pp. 117-122.

Rehurek, Radim, and Petr Sojka. 2010. "Software framework for topic modelling with large corpora." In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.

Rice, Ronald E., Maureen McCreadie, and Shan-Ju L. Chang. 2001. *Accessing and browsing information and communication*. Mit Press.

Ruecker, Stan, Milena Radzikowska, and Stéfan Sinclair. 2011. *Visual interface design for digital cultural heritage: A guide to rich-prospect browsing*. Ashgate Publishing, Ltd.

Russell-Rose, Tony, and Tyler Tate. 2012. *Designing the search experience: The information architecture of discovery*. Newnes.

Sanderson, Mark, and Bruce Croft. 1999. "Deriving concept hierarchies from text." In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206-213.

Shneiderman, Ben. 2003. "The eyes have it: A task by data type taxonomy for information visualizations." In *The craft of information visualization*, pp. 364-371. Morgan Kaufmann.

Simon, Rainer, Leif Isaksen, Elton TE Barker, and Pau de Soto Cañamares. 2016. "Peripleo: a tool for exploring heterogenous data through the dimensions of space and time." *Code4Lib Journal*.

Sinclair, James, and Michael Cardew-Hall. 2008. "The folksonomy tag cloud: when is it useful?." *Journal of Information Science* 34, no. 1: 15-29.

Skov, Mette, and Peter Ingwersen. 2008. "Exploring information seeking behaviour in a digital museum context." In *Proceedings of the second international symposium on Information interaction in context*, pp. 110-115.

Speakman, Robert, Mark Michael Hall, and David Walsh. 2018. "User engagement with generous interfaces for digital cultural heritage." In *International Conference on Theory and Practice of Digital Libraries*, pp. 186-191. Springer, Cham.

Stoica, Emilia, Marti A. Hearst, and Megan Richardson. 2007. "Automating creation of hierarchical faceted metadata structures." In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 244-251.

Sun, Yuyin, Adish Singla, Dieter Fox, and Andreas Krause. 2015. "Building hierarchies of concepts via crowdsourcing." *arXiv preprint arXiv:1504.07302*.

Vane, Olivia. 2020. Timeline design for visualising cultural heritage data. PhD diss., Royal College of Art.

Vizine-Goetz, Diane. "Dewey browser." *Cataloging & classification quarterly* 42, no. 3-4 (2006): 213-220.

Walsh, David, Mark M. Hall, Paul Clough, and Jonathan Foster. 2020. "Characterising online museum users: a study of the National Museums Liverpool museum website." *International Journal on Digital Libraries* 21, no. 1: 75-87.

White, Ryen W., and Resa A. Roth. 2009. "Exploratory search: Beyond the query-response paradigm." *Synthesis lectures on information concepts, retrieval, and services* 1, no. 1: 1-98.

Whitelaw, Mitchell. 2015. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly* 9, no. 1: 1-16.

Wilson, Max L., and Elsweiler, David. 2010. Casual-leisure Searching: the Exploratory Search scenarios that break our current models. *HCIR 2010: 28.*

Wilson, Mathew, Jonathan Hurlock, and Max Wilson. 2012. "Keyword clouds: having very little effect on sensemaking in web search engines." In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pp. 2069-2074.

Windhager, Florian, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. 2018. "Visualization of cultural heritage collection data: State of the art and future challenges." *IEEE transactions on visualization and computer graphics* 25, no. 6: 2311-2330.

Mauricio Yagui, Marcela Mayumi, Luís Fernando Monsores Passos Maia, Jonice Oliveira, and Adriana Vivacqua. 2019. "A Crowdsourcing Platform for Curating Cultural and Empirical Knowledge. A Study Applied to Botanical Collections." In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, pp. 322-326.