Explore The Stacks: A System for Exploration in Large Digital Libraries

Mark M Hall
Department of Computing, Edge Hill University
St Helens Road, Ormskirk, L39 4QP
United Kingdom
Mark.Hall@edgehill.ac.uk

ABSTRACT

Providing access to large digital library collections to novice users requires novel interfaces that are not built around the concept of search, as novice users frequently struggle to formulate appropriate queries. This paper presents the "Explore the Stacks" system, which provides a novel, browsing-focused interface for exploring digital library collections that is applicable to Big Data scale digital libraries. The system is demonstrated using a collection of approximately one million book illustrations provided by the British Library.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: User Issues

Keywords

exploration, digital library, user interface

1. INTRODUCTION

In recent years large digital library (DL) collections have become available, for example Europeana[1] holds over 22 million items, while the British Library (BL) bibliographic data[2] contains over 14 million records. However, this vast amount of material can also be difficult to access since users are provided with little or no guidance on the information in these collections. Systems typically offer keyword-based search interfaces, which are well-suited for expert users [1], but which do not support non-expert users, who are often unfamiliar with the collections and struggle to formulate appropriate queries [2, 3, 4].

Alternative systems have been developed that enable discovery and exploration for the novice user, however these are either primarily focused on search [5, 6, 4] or use visualisations that do not scale to millions of records [7]. Thus there remains a need for systems that focus on letting the user browse the collection and that scale to modern DL sizes.

[1]The European Digital Library, `http://www.europeana.eu`
[2]The British Library`http://www.bl.uk`

978-1-4799-5569-5/14/$31.00

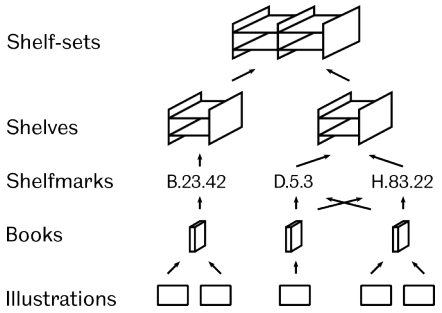

Figure 1: ETS model: Books have one or more Illustrations and are linked to one or more Shelfmarks. The Shelfmarks are grouped together into Shelves, which are hierarchically organised into a Shelf-sets.

2. THE "EXPLORE THE STACKS" SYSTEM

The "Explore the Stacks" (ETS) system was developed around the idea of organising the data into shelves as in a physical library or archive. This allows the user to browse through the collection without having to formulate a query. The system automatically creates this navigation structure from any meta-data available in the collection, such as keywords, user-generated tags, or subject classifications.

The ETS system can be deployed on top of any DL, but in this demo, the data-set is a collection of book illustrations made publicly available by the BL[3]. The data-set consists of meta-data for the books and illustrations extracted from the books, and the illustration image data. The meta-data is very detailed, however in the ETS system only the books' titles and shelfmarks are used together with the actual illustrations. The data-set is loaded in a three-step process:

1. *Data-loading*: First, book and illustration meta-data are loaded into a relational database. Then the shelfmarks are extracted from the book metadata into the database. After loading the collection consists of 31,183 books with 1,021,271 illustrations, linked to a total of 34,240 shelfmarks. Only this step is data-set specific;

2. *Pre-processing*: Next, the shelfmarks are processed to create the shelf structure that the users will explore. The shelfmarks are sorted alphabetically, taking into account any structure in the shelfmarks, and then grouped together into shelves so that a single shelf contains approximately 200 books. Each shelf is given

[3]`https://github.com/BL-Labs/imagedirectory`

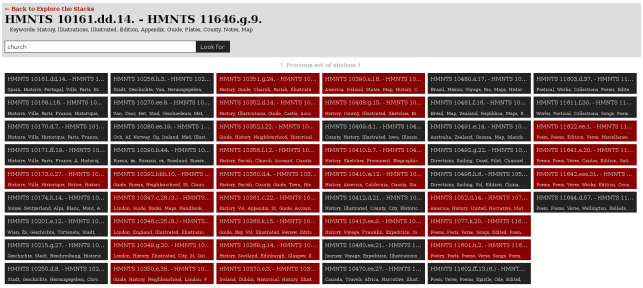

Figure 2: The ETS interface showing a set of shelves. As the user has provided a query term, shelves that contain books matching the query are highlighted in red. The shelf's keywords can scroll to allow all keywords to be displayed within the limited space.

a title based on the first and last shelfmark that belong to the shelf. The shelves are grouped into sets of 50 to create a hierarchical navigation structure. A total of 174 leaf shelves containing actual books and 4 higher-level shelf-sets are created for the data-set.

To provide additional exploration support, each shelf is annotated with a set of 20 keywords. The keywords are extracted from the book titles using NLTK[4] to tokenise and stopword them. Then TFIDF is used to select the 20 most distinctive keywords for each shelf;

3. *Indexing*: As the ETS system also includes search functionality the books and shelves are indexed using ElasticSearch[5]. For individual books, all of the meta-data provided in the original data-set is indexed. For the shelves and shelf-sets, the titles of all books allocated to that shelf or its descendant shelves are combined and the resulting data indexed.

When the user first accesses the ETS system, they are shown the top-most set of shelves (fig. 2). At the top of the page the user is provided with the current shelf's title, its keywords, and a search box in case the user wishes to filter specific shelves. Below that, the the shelf-set's shelves are shown in a grid structure. For each shelf its title and keywords are shown. When the user moves their mouse over a shelf, the keywords scroll in order to show all 20 keywords. If the user provides search keywords, then any shelf that matches the keywords is highlighted in red, to sign-post potential areas of interest to the user.

If the user clicks on a shelf, they are taken to the level of individual books shown in figure 3. The books are shown in a dense list in order to accommodate the display of the shelf's approximately 200 books. The dense list allows for quick scanning of large numbers of books and has the potential to promote serendipitous discovery[8]. As with the shelves, any books that match the search terms provided by the user are highlighted in red. If the user clicks on an individual book, then the illustration browser pop-up appears. This lets the user browse through the illustrations from that book. The user can click on the illustration to navigate to a the BL's page for that illustration.

[4] `http://www.nltk.org`
[5] `http://www.elasticsearch.org`

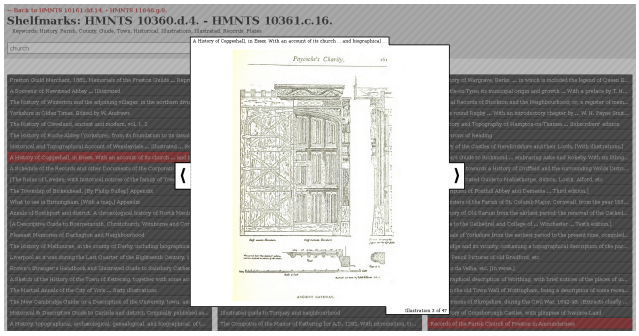

Figure 3: The ETS interface showing a single shelf with all its books. The user has clicked on a book, opening a pop-up that lets the user scroll through the book's illustrations.

3. DEMO DESCRIPTION

The "Explore the Stacks" demo will provide a guided tour to the ETS system, demonstrating how the data is transformed from the source data into the final interface. Participants will then be able to explore the ETS system, available at `http://promise.sheffield.ac.uk/explore-the-stacks`.

4. REFERENCES

[1] A. Sutcliffe and M. Ennis, "Towards a cognitive theory of information retrieval," *Interacting with Computers*, vol. 10, pp. 321–351, 1998.

[2] M. Wilson, K. B, S. MC, and S. B, "From keyword search to exploration: Designing future search interfaces for the web," *Foundations and Trends in Web Science*, vol. 2, no. 1, pp. 1–97, 2010.

[3] G. Geser, "Resource discovery - position paper: Putting the users first," *Resource Discovery Technologies for the Heritage Sector*, vol. 6, pp. 7–12, 2004.

[4] M. M. Hall, O. L. de Lacalle, A. Soroa, P. D. Clough, and E. Agirre, "Enabling the discovery of digital cultural heritage objects through wikipedia," in *Proceedings of the LaTeCH workshop held at EACL 2012*, 2012.

[5] A. Shiri, C. Revie, and G. Chowdhury, "Thesaurus-enhanced search interfaces," *Journal of information science*, vol. 28, no. 2, pp. 111–122, 2002.

[6] M. Hearst, "Clustering versus faceted categories for information exploration," *Communications of the ACM*, vol. 49, no. 4, pp. 59–61, 2006.

[7] P. Cubaud, P. Stokowski, and A. Topol, "Binding browsing and reading activities in a 3d digital library," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2002, pp. 281–282.

[8] S. Makri, A. Blandford, M. Woods, S. Sharples, and D. Maxwell, "Making my own luck: Serendipity strategies and how to support them in digital information environments," *Journal of the Association for Information Science and Technology*, 2014. [Online]. Available: http://dx.doi.org/10.1002/asi.23200