# Exposing Ourselves: Displaying our Cultural Assets for Public Consumption

Gary Munnelly
Adapt Centre
O'Reilly Building
Trinity College
Dublin, Ireland
munnellg@tcd.ie

Kevin Koidl
Adapt Centre
O'Reilly Building
Trinity College
Dublin, Ireland
Kevin.Koidl@scss.tcd.ie

Séamus Lawless
Adapt Centre
O'Reilly Building
Trinity College
Dublin, Ireland
Seamus.Lawless@scss.tcd.ie

## ABSTRACT

This paper discusses an early stage project to develop a new, enhanced interface for Trinity College Dublin (TCD) Digital Collections website. We describe the current state of the portal and outline some of the unique issues observed when examining user engagement.

A major factor in our development of enhanced search tools will be to leverage the entities present in the documents to establish more reliable connections between items in the collection. Not only do we expect that this will lead to better ranked search results, but we also wish to investigate how these entities may be used to encourage site visitors to explore the site beyond their initial research goal.

The early stage of this project means that plans are still being finalised. Hence we speculate about other methods which may be applied to this corpus.

## Keywords

Entity Search; Digital Libraries; Information Retrieval

## 1. INTRODUCTION

In many ways, the vision of Digital Humanities with regards to cultural heritage is a noble one. It is one in which all people have free, unbridled access to primary sources from which they may learn about their heritage and the rich history of their origins. We are free to lose ourselves in the depths of a historical archive from the comfort of our computer screens and supported in our exploration by a host of intelligent information retrieval systems.

In theory, after the arduous process of digitising the collection, providing such functionality ought to be a simple task. Building and deploying a website has become a trivial process and off the shelf tools such as Solr provide state-of-the-art text retrieval functionality with minimal effort. Given a suitable portal and a search box which returns ranked results, what more could a user want?

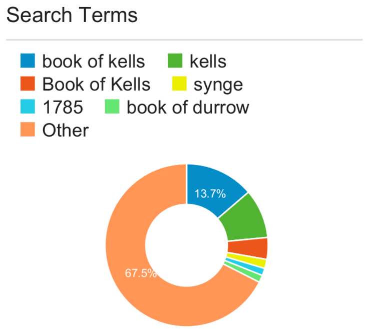


**Figure 1: Graph of most popular search terms on the Digital Collections site**

As it happens, this approach to curating documents has been found wanting in many ways. The most immediate problem with the query-response paradigm is that in order to be able to use the search interface we must know exactly what we are looking for and the manner in which it is represented in the collection. The search engine retrieves documents that it judges to be pertinent to our query and returns them to us without any explanation as to why these might be relevant, nor any encouragement to continue our investigation in a particular direction. It is up to the user to interpret the results, it is up to the user to establish relationships within the collection and it is up to us as the user to identify worthwhile avenues of future research [7]. Given that their knowledge of the collection is probably quite limited to begin with, this is hardly helpful. As was aptly put by Mitchell Whitelaw [8], these interfaces are not “generous”.

This need for a more generous interface is the focus of a project currently being undertaken by Trinity College Dublin (TCD) Digital Collections. At present the website provides the simple search box that we have come to expect which is driven by a default deployment of Solr. After conducting a search, users can narrow their interests along a broad series of facets: genre, media type, Trinity department, date and subject area. This interface results in a limited search experience, particularly with regards to exploration. The effects of this are demonstrable simply by looking at where the majority of traffic flows through the site (Figure 2).

The most famous text on the Digital Collections portal is the Book of Kells [1]. A huge percentage of hits on the site can be attributed to this single page and variants of the query string “Book of Kells” are consistently among the most frequent searches conducted. Indeed, it is worth noting that many visitors to the site land directly on the page for the Book of Kells having been referred there from Google, Facebook, Twitter etc. They never even see the initial search

In Proceedings of 1st International Workshop on Accessing Cultural Heritage at Scale (ACHS'16), June 22, 2016, Newark, NJ, USA. Copyright 2016 for this paper by its authors. Copying permitted for private and academic purposes.

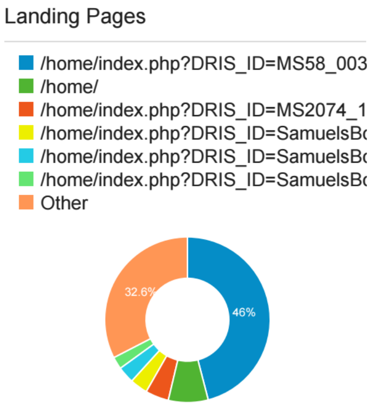


**Figure 2: Graph of pages which site visitors first land on. Note the DRIS_ID for the Book of Kells is MS58_003v which ranks above the home page**

box on the homepage. After viewing the book, most users then simply browse away from the portal, not realising that they have barely touched the tip of the iceberg with regards to the volume of information and material available to them.

Hence our goal is twofold; to provide a better, more accurate, more supportive search experience to users who come to explore the TCD Digital Collections site and to foster a sense of curiosity in those who come to see one artifact, but may have an interest in so many more.

## 2. CORPUS

The corpus is comprised of approximately 100,000 high resolution scans of various documents curated by the Digital Collections group. These range from manuscripts to illustrations, etchings, postcards, templates, graphs, musical scores and more, spanning more than 1,000 years of human history. Information extraction techniques such as optical character recognition (OCR) have not been applied to the renderings, but each image has meta-data associated with it describing important attributes of the artifact. This data is listed in a single XML file which has been provided to us and is the foundation upon which we must build a new search interface.

As is typical in collections of this type, many of the XML fields denote information such as page number, document ID, catalogue number etc. However, there has also been some effort made to make the collection semantically inclined, although not fully semantically linked. The names of several fields are designed to reflect the structure of four well established library cataloguing ontologies: The Library of Congress Name Authority File (NAF), The Library of Congress Subject Headings (LCSH), Getty Vocabularies Art and Architecture Thesaurus (AAT) and Getty Vocabularies Union List of Artist Names (ULAN). The choice of ontology for a particular field is dependent on the nature of the content it represents and the availability of information within the ontologies themselves. For example, if an artistâĂŹs name cannot be found in NAF, then ULAN is used instead.

Although the entries in these ontologies are not explicitly referenced by the meta-data (i.e. there are no URIs used in the XML file), the names of various fields have been selected so that they may be related back to their ontological equivalents. For example, the field denoting the subject of a document is named `subjectlcsh` indicating that the data stored here is relevant to the LCSH ontology. While this is not ideal, it does mean that semantically linking the collection is possible and has be made easier by this method of annotating the data.

In addition to these rigidly defined attribute fields, there are also a number of free text fields, `abstract` and `description` being the two most verbose. These free text fields contain additional information about the artifact, much of which is not actually described in the more semantic attributes. These are human readable sections which describe the artifact in moderate detail, giving an explanation of its origins, who commissioned it, where was it commissioned, how it came to be in the library or any other information which was available to the transcriber. Often these fields reference entities which are not mentioned in any of the other document attributes, meaning there is much information hidden in these fields which could be extracted and harnessed to power a more meaningful search experience.

## 3. METHOD

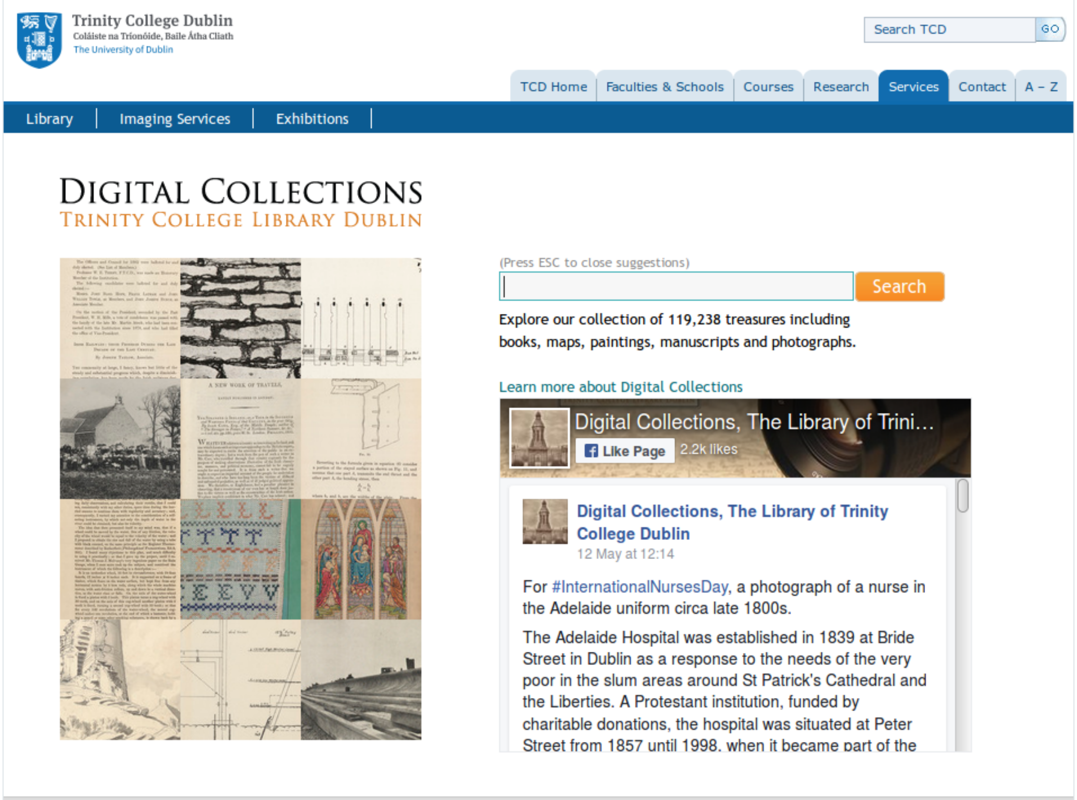
Fostering engagement and encouraging exploration means discerning what interests a user and presenting them with content which relates to that interest. It may also mean determining what is of interest to a community of people at large and using this group perspective to assist an individual whose exploration has stalled.

While we could use traditional language modelling or probabilistic methods to determine which documents may be discussing the same subject and then make recommendations based on that, it is much better if we can establish what real world, tangible objects are influencing the user's search and then trace these figures through the collection. In order to do this, we must know what entities are present in the corpus to begin with. We are fortunate that many potentially useful entities have been manually extracted and stored in the XML file for us. However, much information is also hidden in the free text fields spread throughout the meta-data. This presents some interesting opportunities to perform automatic information extraction and analysis on the collection.

Named Entity Recognition (NER) is a well established field in Natural Language Processing (NLP) for locating references to known entities in a body of text [6]. In general we search for specific patterns, parts of speech or words which appear in a gazetteer of terms. Much like anything involving natural language and computers, the results can be noisy. However, after the results of NER have been sanitised, they may then be disambiguated to a suitable knowledge source [5, 2].

Within the Digital Collections corpus, identifying mentions of entities in the free text fields and disambiguating them to a common knowledge base will allow us to establish which documents are related to which entities and, by extension, which documents are related to each other.

Disambiguation involves more than just co-referencing these entities within the collection. It links the collection's entities to a higher knowledge base which may connect them by proxy to external knowledge sources such as Wikipedia. These external sources may assist the user in understanding the primary source material making the content more accessible for those who are inexperienced with the collection.



Explore our collection of 119,238 treasures including books, maps, paintings, manuscripts and photographs.

Digital Collections, The Library of Trinity College Dublin
12 May at 12:14

For #InternationalNursesDay, a photograph of a nurse in the Adelaide uniform circa late 1800s.

The Adelaide Hospital was established in 1839 at Bride Street in Dublin as a response to the needs of the very poor in the slum areas around St Patrick's Cathedral and the Liberties. A Protestant institution, funded by charitable donations, the hospital was situated at Peter Street from 1857 until 1998, when it became part of the

Figure 3: A screenshot of the current home page of the digital collections website

The challenge is to determine which entity in the knowledge base is being referred to by the mention found in the text.

While this focus on entities may be useful, it may also be of benefit to attempt to establish the larger context in which a user's search is taking place. While the corpus is large in size (the abstracts alone totalling almost 21,000,000 words) the vocabulary is highly constrained (a little over 10,000 unique terms) suggesting that topic modelling may also be a viable option for structuring the corpus and influencing search.

Accurate topic modelling is difficult to achieve. Determining exactly how much content is required in order for a topic model to stabilise can be hard [4] and even after the model has stabilised there is no guarantee that the topics will be of use. Nevertheless, it may still be a worthwhile investigation to perform topic analysis such as Latent Dirichlet Allocation [3] on the collection to see if new, useful patterns beyond the broad facets already in use may be found.

## 4. CONCLUSIONS

As can been seen, there are several options for what can be done when given a collection such as TCD's Digital Collections corpus. The quality with which we can automatically extract information and relationships from the collection are greatly dependent on the quality of the data itself. Quantity of data also plays a role in the accuracy of automatic methods. However with the data extracted from the collection, we have more information at our disposal for assisting and engaging with the user as they search the collection.

Of course, even the best search interface can be felled by poor user interface design. This too will be a factor in the final development of the new Digital Collections portal.

## 5. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) in the ADAPT Centre (adaptcentre.ie) at Trinity College, Dublin.

## 6. REFERENCES

[1] Book of Kells. http://digitalcollections.tcd.ie/home/index.php?DRIS_ID=MS58_003v. [Online; accessed 30-May-2016].

[2] A. Alhelbawy and R. J. Gaizauskas. Graph ranking for collective named entity disambiguation.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] D. Greene, D. OâĂŹCallaghan, and P. Cunningham. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer, 2014.

[5] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 499–508. ACM, 2014.

[6] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[7] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[8] M. Whitelaw. Generous interfaces for digital cultural collections. *Digital Humanities Quarterly*, 9(1), 2015.