# ENRICH 2013
# The First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage

**Séamus Lawless**
Trinity College Dublin, Ireland
*seamus.lawless@scss.tcd.ie*

**Maristella Agosti**
University of Padua, Italy
*agosti@dei.unipd.it*

**Owen Conlan**
Trinity College Dublin, Ireland
*owen.conlan@scss.tcd.ie*

**Paul Clough**
University of Sheffield, U.K.
*p.d.clough@sheffield.ac.uk*

**Abstract**

On August 1st, 2013 the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH 2013) was held as part of the SIGIR 2013 conference in Dublin, Ireland. An invited talk was given by Prof. Jaap Kamps. There were 3 full papers and 3 short papers presented in addition to a poster and demonstration session. Finally, there was a lively discussion on determining the research roadmap for this specialist area of IR research. This contribution discusses the events of the workshop and outlines future directions for the community.

## 1 Introduction

The *First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH 2013)*[1] was held on the 1st of August 2013 in conjunction with the 36th Annual ACM SIGIR Conference in Dublin, Ireland[2]. The aims of ENRICH 2013 were threefold: 1) to discuss the challenges and opportunities for Information Retrieval (IR) research in the area of Cultural Heritage; 2) to encourage collaboration between researchers engaged in work in this specialist area of IR, and to foster the formation of a research community; and 3) to identify a set of actions which the community should undertake to progress the research agenda.

## 2 Background and Motivation

A key challenge facing the curators and providers of digital cultural heritage worldwide is to instigate, increase and enhance engagement with their collections. To achieve this, a fundamental

[1] http://www.cultura-strep.eu/events/enrich-2013/
[2] http://sigir2013.ie/

change in the way these artefacts can be discovered, explored and contributed to by users and communities is required. Cultural heritage artefacts are digital representations of primary resources: manuscript collections, paintings, books, photographs etc. The text-based resources are often innately "noisy", contain non-standard spelling, poor punctuation and obsolete grammar and word forms. The image-based resources often have limited associated metadata describing the resources and their content. In addition, the information needs and tasks of cultural heritage users are often complex and diverse, evolving around information exploration and discovery, synthesis and meaning-making. This presents a specific set of challenges to traditional IR techniques and approaches.

## 3 Goals and Objectives

ENRICH 2013 solicited cooperation between researchers working in many different areas, such as the theory and foundations of IR, cross- and multi-lingual IR, personalised search, recommender systems and human-computer interaction.

The objective of the workshop was to investigate the enhanced retrieval of, and interaction with, cultural heritage collections. Papers were sought which discussed innovative forms of personalised, multi-lingual IR, which included:

- Content-aware retrieval which responds to the entities and relationships contained within the artefacts and across collections.
- Personalised IR, exploration and presentation which responds to models of user and contextual intent.
- Community-aware retrieval which responds to wider community activity, interest, contribution and experience.

These new forms of enhanced IR require rigorous evaluation and validation using appropriate metrics, contrasting digital cultural heritage collections and diverse users and communities. A goal of ENRICH was to investigate such evaluation, taking into account the specific requirements of the domain. ENRICH explored the use of multi-lingual, multi-modal and personalised IR techniques and technologies to enable end-users to search in their native language, but receive results collated from content collections in numerous languages, all tailored for their consumption.

The nature of cultural heritage resources means that content analysis in support of IR was of specific interest. This includes the automated normalisation of historical texts, the use of Natural Language Processing (NLP) for entity extraction and metadata generation.

## 4 Invited Presentation

To set the scene for the workshop, a keynote speaker was invited to discuss the challenges faced by researchers and searchers in the area of Cultural Heritage. This keynote presentation was given by Prof. Jaap Kamps of the University of Amsterdam in The Netherlands and was titled *"When Search becomes Research and Research becomes Search"*. The presentation addressed the particular challenges faced when people searching for information using an IR system are researchers and the searches that they perform are part of their research. The challenges examined can be roughly divided into three areas: Content; Users; and Tools.

With regard to data, great strides have been made. There are huge volumes of cultural heritage information now available online; however, the volume of traffic to these sites tends to be below expectations. While the infrastructure and systems used to access collections have changed in radical ways, the methods of information interaction and search have not changed so radically and tend to be bound by traditional approaches. Appropriate tools must be provided to support the types of tasks researchers need to complete, rather than blindly following a traditional keyword search paradigm.

Means of constructing complex interrogation strategies and novel methods of result exploration need to be investigated. Jaap's suggestion is that we should not be solely building tools, but "building the toolmakers tools".

## 5 Papers Presented at the Workshop

The main issues discussed during the workshop were introduced by the full and short papers which were accepted and presented at the workshop. The full papers, delivered in the first session of the workshop are briefly described, below:

- Trieschnigg et al. [1] report on a transaction log analysis of the Dutch Folktale Database, an online repository of extensively annotated folktales ranging from old fairy tales to recent urban legends, written in (old) Dutch, Frisian and a variety of Dutch dialects. The authors observed that users have a preference for subgenres within folktales such as traditional legends and urban legends and prefer stories in standard Dutch over stories in Frisian. Searches are typically short and aimed at large groups of stories (from the same subgenre or collector), or specific stories with the same main character. Search sessions are relatively long (median of around 2 minutes) and many result pages are viewed (average: 3.4 pages, median: 2 pages). Based on the observations discussed, the authors proposed a number of improvements to the current search and browsing interface. The findings presented offer insight into the search behaviour of folktale searchers, but are also of interest to researchers and developers working on other e-humanities collections.
- Tran et al. [2] present "Gute Arbeit", a research project which aims to provide intelligent access to qualitative social science data on the subject of "good work", specifically, data collected by the Sociological Research Institute (SOFI) in Göttingen. This data is in form of worker interviews from the German Automobile and Shipyard industries and was collected over the last 40 years, starting from early 60's. Topic modelling is applied to the corpora, however, this leads to poor quality topic models due to the limited number of sociological surveys in the dataset. As a result, topic cropping is used, a fully automated process for selecting and incorporating additional domain-specific documents with similar topical content. This expands the dataset in the hope of significantly improving the quality of the inferred topic models. The approach was tested on thematically close English and German document corpora. The results produced for the German corpora slightly outperformed those of the English dataset.
- Sweetnam et al. [3] present work carried out on developing a methodology for identifying and characterising a spectrum of users for a corpus-agnostic environment designed to facilitate research into cultural heritage collections. It outlines the approach taken to the design of the CULTURA environment, and addresses the challenges and opportunities presented by this broadly-based user engagement. It also outlines how adaptive strategies have been used within the CULTURA environment to ensure that it effectively addresses the needs of its different user communities.

The short papers presented in the second paper session of the workshop are described briefly, below:

- Yoshimura et al. [4] propose a method of personal name extraction from ancient Japanese texts based on Support Vector Machine (SVM) using features of character appearance and probabilistic word segmentation. Experimental results showed that the proposed method was able to extract personal names from ancient Japanese texts with a 4% (approx.) improved F-measure.
- Oard et al. [5] describe the design process for the construction and evaluation of a system to

automatically construct links between spoken conversations that address different aspects of the same event. This is seen as one step, among many, towards building richer interconnections both between part of the same collection, and between collections. The corpus used to demonstrate the approach is a collection of several thousand hours of recordings made in the Mission Control Centre during the Apollo space missions of the 1960's and 1970's.

- Munnelly et al. [6] discuss the CULTURA project and its attempts to personalise a user's experience of exploring cultural heritage collections regardless of their level of expertise. The services provided within the environment were introduced with respect to the four phase personalisation model used by CULTURA. In addition to these services CULTURA's comprehensive user model was detailed, in order to explain how a user's actions can affect the environment in which they work.

## 6 Posters and Demonstrations Presented at the Workshop

A number of posters and demonstrations were also presented which helped to stimulate much discussion around the challenges in this area. The presenters had the opportunity to discuss and demonstrate their work to experienced researchers and receive feedback on improvements that could be made or other approaches that could be applied. The presentations were:

- The CULTURA Evaluation Model: An Approach Responding to Evaluation Needs in an Innovative Research Environment. C. M. Steiner, E. C. Hillemann, A. Nussbaumer, D. Albert, M. Sweetnam, C. Hampson & O. Conlan.
- A Social Network Study of Evliyâ Çelebi's Book of Travels-Seyahatnâme. C. Karbeyaz, E. F. Can, F. Can & M. Kalpakli.
- Publishing Social Sciences Datasets as Linked Data: a Political Violence Case Study. R. Brennan, K. Feeney & O. Gavin.
- GALATEAS D2W: A Multi-lingual Disambiguation to Wikipedia Web Service. D. Lungley, M. Poesio, M. Trevisan, M. Althobaiti & V. Nguyen.
- Generating Automatic Keywords for Conversational Speech ASR Transcripts. H. Ryu & M. Lease.
- The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. G. Munnelly, C. Hampson, S. Lawless, M. Agosti & O. Conlan.
- Personalized Access to Cultural Heritage Spaces (PATHS). P. Clough, P. Goodale, M. Stevenson & M. Hall.

All of the publications presented at ENRICH 2013 are available for download from: http://www.cultura-strep.eu/events/enrich-2013/programme

## 7 Research Challenges – Future Directions

The workshop discussion sessions debated various challenges faced by research in the area of exploration, navigation and retrieval in cultural heritage. The sessions also discussed possible ways of progressing this research and constructing an appropriate roadmap for the research community.

The challenges discussed included many of the points raised in the keynote address. What interaction methods should be supported by tools in this area? Is keyword search how people want to explore the collection? The attendees were in agreement that more complex expressions of information need must be supported and novel means of exploring relevant material across collections is essential. Moving beyond the ranked list was seen as key to the success of research

platforms in cultural heritage.

User-driven design was one of the reoccurring themes throughout the workshop and it was clear that the most successful cultural heritage platforms have involved the end-user right from the beginning of the design process. There are numerous approaches to user-centered design that can be adopted, including co-operative inquiry, living labs etc. Attendees from all disciplines felt that such approaches are essential to producing systems that accurately address the needs of users.

As highlighted in Section 3, above, these new forms of enhanced IR require rigorous evaluation and validation using appropriate metrics, contrasting digital cultural heritage collections and diverse users and communities. Traditional IR Metrics may not be appropriate so more research is required to identify what metrics are most suitable, or if in fact new metrics are needed. The notion of test collections for humanities research, such as CHIC was also discussed.

A related discussion focused on the collections and the artefact themselves. The question was raised "Where does an artefact begin and end"? Applications that use Natural Language Processing (NLP) techniques are generating metadata about artefacts, users that annotate resources are adding to that metadata. This produces layers of annotations that are beneficial to an IR system and which may be considered part of the resource itself.

Finally there was a debate around authoritativeness in user-generated content and user-moderated annotation. When decisions are made regarding metadata, such as the merging of entity occurrences in entity extraction, who has the authority to make those decisions, and should they be applied across the entire collection? The accuracy and quality of user-generated annotations is another example where questions of authority must be considered.

In order to encourage the formation of a community of researchers working on these challenges, a number of future directions were discussed at the workshop. Initially, a mailing list will be used to continue discussions that began at the workshop. In addition, many of the attendees are planning to make collaborative submissions to the ACM Journal on Computing and Cultural Heritage. Finally, a second edition of the workshop will be proposed in order to continue community development and the advancement of the ENRICH research agenda.

## 8 Acknowledgements

The work of Séamus Lawless, Maristella Agosti and Owen Conlan has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area Digital Libraries and Digital Preservation (ICT-2009.4.1), grant agreement no. 269973.

The work of Séamus Lawless and Owen Conlan has been partially supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL.

The work of Paul Clough has been partially supported by the PATHS project, as part of the Seventh Framework Programme of the European Commission, Area Digital Libraries and Digital Preservation (ICT-2009.4.1), grant agreement no. 270082.

## 9 References

[1]. D. Trieschnigg, D. Nguyen & T. Meder. *In Search of Cinderella: A Transaction Log Analysis of Folktale Searchers*. In the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.

[2]. N.K. Tran, S. Zerr, K. Bischoff, C. Niederée & R. Krestel. *"Gute Arbeit": Topic Exploration and Analysis Challenges for Corpora of German Qualitative Studies*. In the Proceedings of the

First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.

[3]. M. Sweetnam, M. Ó Siochrú, M. Agosti, M. Manfioletti, N. Orio & C. Ponchia. *Stereotype or Spectrum: Designing for a User Continuum*. In the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.

[4]. M. Yoshimura, F. Kimura & A. Maeda. *Personal Name Extraction from Ancient Japanese Texts.* In the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.

[5]. D. Oard, A. Sangwan & J.H.L. Hansen. *Reconstruction of Apollo Mission Control Center Activity.* In the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.

[6]. G. Munnelly, C. Hampson, O. Conlan, E. Bailey & S. Lawless. *Personalising the Cultural Heritage Experience with CULTURA*. In the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH 2013, held as part of the 36th Annual ACM SIGIR Conference, SIGIR 2013, in Dublin, Ireland. August 1st, 2013.