Grant Agreement No. **ICT-2009-270082**
Project Acronym **PATHS**
Project full title **Personalised Access To Cultural Heritage Spaces**

D1.4 Final State of Art Monitoring Report

Authors: Phil Archer (i-sieve Technologies)
Konstantinos Chandrinos (i-sieve technologies)
Mark Stevenson (University of Sheffield)
Mark M. Hall (University of Sheffield)
Paul Clough (University of Sheffield)
Paula Goodale (University of Sheffield)
Eneko Agirre, German Rigau (EHU/UPV)
Nigel Ford (University of Sheffield)
Jillian Griffiths (MDR Partners)

Contributors: Kate Fernie (MDR Partners)
Timos Kouloumpis (i-sieve technologies)

Project funded under FP7-ICT-2009-6 Challenge 4 – "Digital Libraries and Content"	
Status	Final
Distribution level	Public
Date of delivery	
Type	Report
Project website	http://www.paths-project.eu
Project Coordinator	Dr. Mark Stevenson University of Sheffield

Change Log

Version	Date	Amended by	Changes
0.1	28-06-2012	Phil Archer	Everything but the Executive summary
0.2	29-06-2012	Phil Archer	Executive summary added, proofreading etc.
0.3	30-06-2012	Phil Archer	Addition of references
0.4	18-07-2012	Phil Archer	Finalising

Contents

1. Executive Summary

This document provides an update to an Initial State of the Art Monitoring report that was delivered by the project a year ago (D1.2). An extensive literature review in that document covered the areas of Educational Informatics, Information Retrieval and Semantic Similarity relatedness.

Returning first to the field of Educational Informatics, the partners have considered a number of possible approaches to evaluating PATHS users' Cognitive Styles. Five approaches were considered before selecting Riding's Cognitive Style Analysis as being the most appropriate. This provides methods for distinguishing between people who are 'wholists' and analysts and between those who tend towards verbalising or imaging.

An Information Retrieval paper of particular interest sets out a series of long term research objectives for the IR community that attempts to see search queries and the presentation of results from different perspectives. This is at the heart of PATHS which is focused on stimulating interest on cultural heritage by proposing objects to users in interesting and unusual ways. The partners are also looking forward to the Cultural Heritage in Context exercise that takes place later this year. Organised by the Promise Network of Excellence, the exercise uses Europeana data and evaluates information retrieval in three tasks. Firstly the *ad hoc* retrieval of data based on typical queries; secondly the diversity of objects returned - the more diverse the types of object, video, text and images, the better; and thirdly Semantic Enrichment. This final exercise requires the generation of a set of concepts related to the query from sources such as Wikipedia and the Linked Open Data (LOD) cloud. All three exercises are highly relevant to PATHS and the partners will be submitting an entry to the exercise.

A method used to help identify documents related to a particular topic is to calculate the Semantic Textual Similarity between sentences in two documents. Until recently, this approach suffered from a paucity of reference data on which to train machine-learning algorithms. Recently a corpus of 2000 sentence pairs has been published, based on previously existing paraphrase datasets and machine translation evaluation resources. In addition to the 2000 sentence pairs, further data sets are available for testing purposes. With that data now available, the current best automated systems now achieve better than 80% correlation, well above the previous best of around 31%.

Named Entity Recognition and Classification is now a well established field and Wikipedia is the *de facto* standard named entity catalogue. The TAGME tool achieves 75% accuracy in linking short English texts to relevant Wikipedia articles, and that figure improves significantly for texts describing named entities. An important feature of Wikipedia is the guide to editors that sets out the circumstances in which links should be made between articles and, just as importantly, when links should not be made. It is now possible to use Wikipedia not just as test data but as training data. Taking this approach, recent efforts have achieved better than 74% accuracy in detected related articles, an improvement on the previous best of around 55%.

Two areas covered in this report, that were not covered in D1.2, are: Crowdsourcing and mobile Web.

A prominent platform for crowdsourcing is Amazon's Mechanical Turk. Non-expert workers are requested to carry out tasks that computers cannot do and research shows that the results rival those of domain-specific experts. However, some form of quality control is

necessary. Increasing the money paid to workers generates more results but not better results. It seems that 7 is the optimum number of workers to repeat each task to get the best balance of accuracy and cost. If a particular worker is returning results that are consistently different from the norm then his/her results can be ignored but the individual might also be shown a 'gold standard' example to help them correct their assessments. Traps can also be set and results considered only from workers who correctly avoided them.

In cultural heritage, crowdsourcing is used to carry out a number of tasks such as image cropping (to get the best digital version of a scanned image), proofreading and correction of texts scanned using Optical Character Recognition (OCR), transcribing old hand written texts and adding metadata to objects. The public's willingness to tag items is used extensively and builds up very useful folksonomies that describe objects in ways that have more meaning to end users and therefore can be very useful in search. The experience gained from many institutions and projects has allowed an analysis of 30 considerations for the design of successful crowdsourcing projects. Gaining and sustaining momentum is important, as is the users' motivation. Altruism, subject interest, personal recognition, fun, and being part of a community are often more important than the financial gain sought by Mechanical Turk workers. Simple requirements may lead to tasks becoming addictive but again, as with Mechanical Turk, some form of quality control is necessary for consistently good results. Generally, the use of multiple quality control methods brings more reliable results.

As PATHS prepares to develop applications that access the system other than through a Web site designed for desktop use, the partners considered the state of the art in the Web on mobile devices. Accessing the Web on mobile is now commonplace and, for many users, the primary means of access. The developer's problem of creating content that will work across the many different devices with different capabilities and different screen sizes is largely solved by following two disciplines: Progressive Enhancement and Responsive Web Design. The first assumes a minimal set of capabilities for the device but uses more advanced features to deliver a better experience for the user without depending on those features being available. Responsive Web Design creates a basic layout that works on mobile but then builds in new style features and layout as the available screen size increases.

Finally, in D1.2, the partners presented the results of an analysis of sentiment towards Europeana, as expressed online between its launch on November 2008 and February 2011. A renewed analysis running from June 2011 to June 2012 shows that the negativity surrounding problems at launch has now been forgotten and that negative comments about the service are now negligible. In addition, more of the discussion about Europeana is being held in mainstream media rather then in specialist blogs. Special exhibitions, such as the one focused on the 1914-18 war, are a particular driver for neutral and positive stories about Europeana.

2. Introduction

This report provides an update to the Initial State of the Art Monitoring Report (D1.2), that was completed a year ago. Rather than create an extended version of that document, the current work focuses on developments since then. However, entirely new sections have been added covering the state of the art on crowdsourcing and mobile device capabilities that will become important in the final year of PATHS.

To an end user, the outcomes of the PATHS project should appear very simple to use. This is a key design goal for the system itself. However, that simplicity belies a great deal of complexity beneath the surface. Offering users content that is relevant to their query, and ensuring that it is of a type from which they will benefit most, represents a significant undertaking.

In academia, the problem is broken down and studied in great detail. Such studies, of course, inform everyone's endeavours. Within the cultural heritage space, the aim is to stimulate interest in culture, whether for educational reasons or just for the joy of discovery. PATHS is firmly in the latter category – the aim is to present people with cultural heritage objects and ideas that they are likely to find interesting – but to achieve success requires us to look at a wide range of research results and initiatives, and, as far as possible, the public reaction to them.

This document captures the project partners' combined knowledge of the state of the arts relevant to the project.

Section 3 reviews the most recent academic literature in three areas of interest: Educational Informatics, Information Retrieval, Semantic Similarity and Relatedness.

Section 4 is the first of two entirely new reviews. Crowdsourcing has become a commonly used way of gathering and correcting data in the open data world in general and in cultural heritage in particular. The second new section, 5, reviews the state of the art in mobile Web development as PATHS prepares to create its own mobile application in its final year.

In Section 6 the results of an updated analysis of sentiment towards Europeana is presented that shows interesting and encouraging advances in the public perception of the service.

Throughout the document there are references to the concept of a Path or pathway. These should be taken to mean some sort of directed sequence of steps through objects in a collection or related online resources. The combination of the steps themselves and the commentary provided by the Path creator offer a richer experience than a simple presentation of images or texts.

3. Literature Review

This section of the report provides an update of the three areas covered by the literature review presented in D1.2.

3.1. Educational Informatics

Cognitive styles (CSs) are tendencies displayed by individuals consistently to adopt a particular type of information processing strategy (Entwistle, 1981; Felder and Spurlin, 2005; Ford, 1995; Miller, 1987; Pask, 1976a, 1976b; Riding and Cheema, 1991; Schmeck, 1988; Witkin, Moore, Goodenough and Cox, 1977). Of additional interest to the PATHS project in terms of informing the design of the system is an understanding of cognitive styles, including navigational styles and levels of preferred autonomy.

As a result of the review of the literature and research for D1.2 Initial State of the Art Monitoring, the following approaches to evaluating Cognitive Style have been considered:

- Riding's Cognitive Style Analysis (CSA) test, which measures cognitive style on a verbal–imagery dimension and a wholist–analytic dimension
- Felder and Soloman's *Index of Learning Styles Questionnaire* http://www.engr.ncsu.edu/learningstyles/ilsweb.html
- Ambiguous figures test: Wiseman et al (2011).
- The Alternate Uses Test of divergent thinking: Chamorro-Premuzic and Reichenbacher (2008).
- Convergent Thinking - The Baddeley Reasoning Test (1968)

Of these the PATHS project is has identified Riding's CSA test as a suitable instrument for measuring participant's cognitive style. Riding's CSA measures what the authors refer to as a wholist/analytic dimension (their equivalent to field-dependence/-independence. The analytic/wholist section of this instrument entails two sub-tests. In the first, each screen displays two geometric shapes. The user is asked to determine if the shape on the left is the same as the shape on the right. In the second sub-test, each screen again displays two shapes, but asks the user to decide whether the simple shape on the left is contained within the more complex shape on the right. Speed on the first sub-test is considered to indicate relatively high ability in global, wholist information processing (since this ability is likely to result in global similarities being perceived more quickly). Speed on the second sub-test is considered to indicate high analytic ability, since the simple shape presented at the left must be dis-embeddied from within the complex shape on the right before a decision can be made.

Riding's CSA test also measures the tendency of subjects towards verbalising or imaging. This verbaliser/imager test presents the user with a series of screens, each of which shows a statement about the relationship between two words. The user is asked to decide whether each statement is true or false. Half the statements concern conceptual relationships (asking if x is a type of y); the other half concern visual relationships (asking if x is the same colour as z), for example:

- oak and beech are the same type
- bread and butter are the same colour.

It is assumed that correct responses will be given more quickly by verbalisers in the case of conceptual similarities, since these are essentially verbal and cannot be represented visually. Imagers are considered likely to give correct responses more quickly to the statement concerning visual relationships, since retrieval from mental images would be quicker.

Peterson et al (2001) has undertaken work to validate and strengthen Riding's CSA test, but unfortunately the test is incompatible with current systems. As a result, the original Riding test will be adopted during initial experiments with participants of the PATHS project.

3.2. Information Retrieval

Information Retrieval remains a core technology within PATHS. The field has continued to advance since the publication of the initial state of the art monitoring report (D1.2) with work being published in a variety of conferences (including SIGIR 2011, ECIR 2012, CIKM 2011, WSDM 2012 and WWW 2012) and journals (including *Journal of the American Society for Information Science and Technology*, *Information Processing and Management*, *Information Retrieval* and *Foundations and Trends in Information Retrieval*). The PATHS team have monitored these venues and also attended some of the conferences.

Despite these continued advances the state of the art in Information Retrieval remains largely as described in the initial state of the art monitoring report. An interesting paper that has been published in ACM SIGIR Forum is a report entitled *"Frontiers, Challenges, and Opportunities for Information Retrieval*[1]*"* edited by James Allan, Bruce Croft, Alistair Moffat and Mark Sanderson. The report is based on a three-day workshop (SWIRL 2012 – the second Strategic Workshop on Information Retrieval in Lorne) held in Australia in 2012 and involving many of the best known IR academics in the field. From discussions six main themes emerged as future long-term research objectives for the IR research community:

- *Not just a ranked list*. This theme seeks to move IR systems from ad hoc retrieval (query in; ranked list out) to support other forms of information seeking and more interactive forms of use.

- *Help for users*. This theme considers how IR can be extended to assist various users, e.g. those with physical or learning difficulties.

- *Capturing context*. This theme considers how to capture contextual information and then embed this into the search experience. Context could include aspects such as location, time, work task and individual differences (e.g. learning style).

- *Information, not documents*. This theme considers how to push IR beyond document retrieval and instead return information (e.g. answers to questions or visualisations to support analytics).

- *Domains*. This theme addresses aspects such as specialised search and non-textual retrieval. This also considers how IR systems can be embedded within various contexts, such as the workplace.

[1] http://sigir.org/forum/2012J/2012j_sigirforum_A_allanSWIRL2012Report.pdf

- *Evaluation*. This theme considers how to evaluate highly interactive IR systems that involve user-system interactions and how evaluation resources can be developed that go beyond the query-response paradigm.

When considering these themes for future IR research, the PATHS project is seeking to address a number of these. For example, the PATHS system aims to support users as they explore and learn when navigating through a digital information space (Europeana and Alinari collections). This takes us beyond the simple query-response paradigm and use of ranked lists. The PATHS project is also considering how pathways through the collections can be adapted based on contextual factors, for example user's learning style and past items viewed. The PATHS system is working within a specialised-domain (cultural heritage) and with metadata records compared to the more traditional document retrieval from full text. This presents a number of challenges for data processing within the PATHS project that are actively being tackled.

Evaluation is a continual issue for IR research and the PATHS system presents many challenges for project partners as traditional resources, such as test collections, are not suited to evaluating more exploratory forms of search system. An area of particular interest is evaluating the system as a whole, for example by assessing the quality of the pathways that users produce or based on knowledge gained from navigating collections using pathways (i.e. their learning experience). In addition, project partners are considering how re-useable evaluation resources (e.g. test collections) could be created to assess the performance of various browsing functionalities. Simulation is also an area that is gaining interest from IR researchers as a means of bringing users into system-oriented evaluation (see, e.g. the 2011 SimInt[2] – Simulation of Interaction - workshop held at SIGIR).

A further interesting development since the first state of the art monitoring report is the Cultural Heritage in Context (CHiC)[3] evaluation exercise. CHiC is organised by the Promise Network of Excellence as part of the CLEF 2012[4] conference to be held in September 2012. The aim of this exercise is to evaluate the effectiveness of various Information Retrieval strategies for cultural heritage data. The CHiC organisers are using data from Europeana in English, French and German. The exercise involves three tasks: (1) Ad hoc Retrieval, (2) Variability and (3) Semantic Enrichment. The first, *Ad hoc Retrieval*, evaluates Information Retrieval effectiveness for a set of short topics designed to reflect typical queries submitted by users to the Europeana portal. The task can be attempted as a monolingual, bilingual or multilingual retrieval problem. The second task, *Variability*, requires systems to return a small set of objects from Europeana that are as diverse as possible. For example, they should refer to different types of media, content providers and so on. The third task, *Semantic Enrichment*, requires systems to generate a set of related concepts for a query. The concepts can be obtained from a variety of sources, such as Wikipedia, the LOD Cloud and Europeana itself. Each of these tasks has some relevance to PATHS, particularly the ad hoc retrieval and variability tasks. One of the outcomes of the CHiC evaluation exercise will be an insight into the effectiveness of different techniques for Information Retrieval on Europeana data. The PATHS projects will monitor these outputs and also plans to submit an entry to the exercise.

[2] http://www.dcs.gla.ac.uk/access/simint/
[3] http://www.promise-noe.eu/chic-2012/home
[4] http://clef2012.org/

3.3. Semantic Similarity and Relatedness

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two sentences. STS is related to both Textual Entailment (TE) and Paraphrase (PARA). STS is more directly applicable in a number of NLP tasks than TE and PARA such as Machine Translation and evaluation, Summarization, Machine Reading, Deep Question Answering, etc.

STS differs from TE in as much as it assumes symmetric graded equivalence between the pair of textual snippets. In the case of TE the equivalence is directional, e.g. a car is a vehicle, but a vehicle is not necessarily a car. Additionally, STS differs from both TE and PARA in that, rather than being a binary yes/no decision (e.g. a vehicle is not a car), STS incorporates the notion of graded semantic similarity (e.g. a vehicle and a car are more similar than a wave and a car).

STS provides a unified framework that allows for an extrinsic evaluation of multiple semantic components that otherwise have tended to be evaluated independently and without broad characterization of their impact on NLP applications. Such components include word sense disambiguation and induction, lexical substitution, semantic role labelling, multiword expression detection and handling, anaphora and co-reference resolution, time and date resolution, named-entity handling, under specification, hedging, semantic scoping and discourse analysis.

Datasets for STS are scarce. Existing datasets include those produced by Li *et al* (2006) and Lee *et al* (2005). The first dataset includes 65 sentence pairs which correspond to the dictionary definitions for the 65 word pairs in Similarity (Rubenstein & Goodenough 1965). The authors asked human informants to assess the meaning of the sentence pairs on a scale from 0.0 (minimum similarity) to 4.0 (maximum similarity). While the dataset is very relevant to STS, it is too small to train, develop and test typical machine learning based systems. The second dataset comprises 50 documents on news, ranging from 51 to 126 words. Subjects were asked to judge the similarity of document pairs on a five-point scale (with 1.0 indicating ``highly unrelated'' and 5.0 indicating "highly related"). This second dataset comprises a larger number of document pairs, but it goes beyond sentence similarity into textual similarity.

More recently, a SemEval task on STS has released a large number of sentence pairs with annotations. The training data contained 2000 sentence pairs from previously existing paraphrase datasets and machine translation evaluation resources. The test data also comprised 2000 sentences pairs for those datasets, plus two surprise datasets with 400 pairs from a different machine translation evaluation corpus and 750 pairs from a lexical resource mapping exercise. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk, with high Pearson correlation scores, around 90%. The task attracted 35 teams, submitting 88 runs. The best results scored a Pearson correlation > 80%, well above a simple lexical baseline that only scored a 31% correlation.

The three best systems in the task constitute the current state-of-the-art in this field. The top scoring systems tended to comprise a large number of resources and tools (Bär *et al* 2012, Šarić *et al* 2012), with some notable exceptions like Jimenez *et al* (2012) which was based on string similarity.

3.4. Wikification

A key stage in the content enrichment process will be to provide the user with links to background and explanatory content (i.e. Wikipedia, DBpedia, etc.) to assist them in their understanding of items in collections. Furthermore, once identified, the background and explanatory content and its underlying narrative discourse can suggest coherent and plausible new routes to the system. Wikification refers to the augmentation of text with links or mappings to relevant Wikipedia items (either categories or articles).

In Natural Language Processing (NLP), Named-Entity Recognition and Classification (NERC) deals with the detection and identification of specific entities in running text (Nadeau and Sekine 2007). Current state-of-the-art processors achieve high performance in recognition and classification of general categories such as people, places, dates or organisations (e.g. OpenCalais service for English) (Nothman *et al* 2012). Furthermore, once the named entities are recognised they can be identified unambiguously with respect to an existing catalogue. Wikipedia has become the de facto standard as a named entity catalogue. Wikification allows the automatic linking of the named entities occurring in free text to its corresponding Wikipedia articles (Mihalcea & Csomai 2007). Typically regarded as a Word Sense Disambiguation (WSD) problem (Agirre & Edmonds 2007), where Wikipedia provides the dictionary and training examples. Wikipedia and DBPedia play a central role in the development of Linked Data due to the large and growing number of resources linked to it (e.g. YAGO or Freebase), making DBpedia the main interlinking hub of the Web of Data. Since Wikipedia is aligned with many semantic Web resources such links allows the enrichment of text representation with background knowledge retrieved and filtered from the semantic Web resources.

Current research exploits a variety of measures of coherence/relatedness among Wikipedia pages (Mihalcea & Csomai 2007; Cucerzan 2007; Milne & Witten 2008; Medelyan *et al* 2009; Tonelli & Giuliano 2009; Ferragina & Scaiella 2010; *Kulkarni* et al 2012; Bryl *et al* 2010). TAGME (Ferragina & Scaiella 2010) is a particularly interesting approach, relevant to PATHS, since it addresses *Wikifying* short texts (e.g. snippets, short descriptions or news) achieving similar accuracy to that of long document systems. Currently, these systems achieve accuracy rates over 75% on the English corpus ACEToWiki (Bentivogli 2010). The same approach can be extended to languages other than English. Current performance rates can be improved by focusing on the named entities only, avoiding the annotation of the remainder of the text. In a multilingual setting, once in a language-neutral representation, the knowledge captured for a particular item in one language can be ported to another, balancing resources and technological advances across languages. Public demos of approaches which exploit Wikification (only for English) are WikiMiner[5] Spotlight[6], CiceroLite[7] and, Zemanta[8], TAGME[9] or The Wiki Machine[10].

Wikify! (Mihalcea & Csomai 2007) was the first attempt to address the problem of augmenting a text with links from Wikipedia. This work identified that the two main problems were keyword extraction (finding the most important words in the text), and link disambiguation. The Wikipedia manual of style provides a set of guidelines which although

5 http://wdm.cs.waikato.ac.nz:8080/service?task=wikify
6 http://spotlight.dbpedia.org/demo/index.html
7 http://demo.languagecomputer.com/cicerolite
8 http://www.zemanta.com
9 http://tagme.di.unipi.it
10 http://thewikimachine.fbk.eu

designed for human annotators provided useful insights for designing the keyword extraction module. The main recommendations from the Wikipedia style manual are:

- Authors/annotators should provide links to articles that provide a deeper understanding of the topic or particular terms, such as technical terms, names, places etc.
- Terms unrelated to the main topic and terms that have no article explaining them should not be linked.
- Special care has to be taken in selecting the proper amount of keywords in an article – as too many links obstruct the readers' ability to follow the article by drawing attention away from important links.

The first stage was to form a controlled vocabulary comprising article titles in Wikipedia, plus all surface forms (to capture synonyms). Then, an input document is parsed to give all possible n-grams that are present within the controlled vocabulary (i.e. that link to a valid article). These keyword candidates are then ranked using a keyphrase metric which is assigned based on their existing use as a keyword within Wikipedia:

$$P(keyword|W) \approx \frac{count(D_{key})}{count(D_W)}$$

where $count(D_{key})$ is the number of documents where the term was selected as a keyword, and $count(D_W)$ the number of documents where the term appeared. Disambiguation, selecting the appropriate Wikipedia page for the keyword, is then performed using knowledge based and data-driven approaches.

It should be noted that PATHS also requires the ability to select the words to be linked to the background content data. However, in PATHS only particular instances of general concepts should be linked.

Milne and Witten (2008) use Wikipedia not just as a source of information to link to but also as a training data. Here the first step disambiguates the terms in the document. This is done using a machine learning approach which takes as features the commonness of each sense, the relatedness of the sense to the surrounding context, and the quality of the context (as determined by the number of context terms, their inter-relatedness and their frequency of their use as Wikipedia links). The key difference with the approach of Mihalcea & Csomai (2007) is then in the way links are detected. Once the terms are disambiguated, the Wikipedia articles are then used as training data for a classifier. Positive examples are the articles that were manually linked to, while negative ones were those that were not. Several features are used in the classifier. The first is the same link probability as used in Mihalcea & Csomai (2007). The second is relatedness, since terms most related to the subject are likely to be of most interest. Third is disambiguation confidence, which is obtained from the disambiguation in the previous step. Fourth is generality, defined as the minimum depth at which it is located in Wikipedia's category hierarchy. Finally location and spread; this comprises a set of features including frequency (since the more times a topic is mentioned the more important it is likely to be), the first occurrence (since topics mentioned in the introduction are more likely to be important), last occurrence (similarly for topics in the conclusion), and spread (distance between first and last occurrence, to indicate how consistently the document discusses the topic). Evaluation shows that Milne & Witten (2008) achieve 74.1% F-measure on a sample of 100 Wikipedia articles, compared to an upper bound of 54.6% for the Mihalcea & Csomai (2007) approach.

4. Crowdsourcing

Crowdsourcing makes use of unknown, usually large, populations to carry out online tasks (Howe, 2006). These range from evaluating collections of items, building physical artefacts and social networks, labelling images, rating films, digitising text, spelling correction and assessing recommendations (Doan et al., 2011). Estellés Arolas & González Ladrón-de-Guevara survey over 40 articles that define crowdsourcing to propose a single definition:

> *"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."*

Crowdsourcing has grown in popularity online, with huge success stories such as Wikipedia and Project Gutenberg utilising the 'cognitive surplus' (Shirky, 2010) of many thousands of anonymous web users to develop and edit major sources of digital content for mass consumption. Another interesting development has been the emergence and increasing popularity of Crowdsourcing Systems (e.g. Amazon's Mechanical Turk system, Crowdflower, and others) as a means of accessing very large crowds to complete simple computational tasks. In fact crowdsourcing systems have been identified as a field where rapid growth is expected in coming years (Doan et al., 2011). However, there are several issues facing the development of crowdsourcing systems ranging from ethical concerns to ensuring quality of data collected using unknown populations. Here the focus is on Amazon's Mechanical Turk as the crowdsourcing platform and its use for data collection in Natural Language Processing (NLP) and Information Retrieval (IR) tasks. In contrast, many cultural heritage projects that harness crowdsourcing focus on extending the long-standing tradition of volunteering and altruism, utilising communities of interested users to support large-scale digitisation projects.

4.1. Crowdsourcing in Natural Language Processing and Information Retrieval

Amazon's Mechanical Turk (AMT) is an increasingly popular online crowdsourcing system. It provides a platform for collecting human intelligence for tasks that cannot be done by computers. There are two groups of people: *requesters* and *workers*. Requesters are people who submit tasks, and workers (also known as providers) are those who complete the tasks and receive monetary payments from the requesters. A task is called a HIT (Human Intelligence Task), which is usually in the form of one or more web pages containing text, images, or videos, and input elements like text boxes and radio buttons. Requesters can collect results from MTurk once the tasks are finished. As a cheap and fast service, MTurk is now being used by an increasingly large number of people for a variety of tasks.

Mechanical Turk has been used for various natural language processing and information retrieval tasks. Snow et al. (2008) explored the use of MTurk for collecting annotations for

five tasks: affect recognition, word similarity, recognising textual entailment, event temporal ordering and word sense disambiguation. They showed that across the tasks the use of non-expert labellers rivalled the performance of using an expert annotator. MTurk has also been used for gathering relevance judgements for evaluating search output (Alonso et al., 2008; Alonso & Mizzaro, 2009; Alonso & Baeza-Yates, 2011; Carvalho et al., 2011; Hosseini et al., 2012; Kazai, 2011), generating data to assess attribute extraction and entity resolution (Su et al., 2007), image annotation (Sorokin & Forsyth, 2008), the assessment of the quality of Wikipedia articles (Kittur et al., 2008), the extraction of document facets (Dakka & Ipeirotis, 2008), comparing different length summaries from search results (Kaisser et al., 2008), building datasets for question answering (Kaisser & Lowe, 2008), extracting key phrases from documents (Yang et al., 2009), evaluating machine translation output (Callison-Burch, 2009) and gathering user preferences for evaluating search engine output (Sanderson et al., 2010; Tang & Sanderson, 2010).

4.2. Amazon's Mechanical Turk

Amazon's Mechanical Turk is recognised as a promising platform for crowdsourcing. Although requesters have full control over how much a worker is paid on completing a task, many of them seem to pay between $0.01 and $0.10 for a task taking "a few minutes". It is much less than the MTurk suggested amount of at least the equivalent to the minimum federal wage of $8 per hour or $0.13 per minute (Downs et al., 2010). Interestingly, Mason & Watts (2009) investigated the relationship between financial incentives and the performance of AMT workers by varying the amount of payment. They found that increased financial incentives improved the quantity, but not the quality, of work performed by participants. It was explained that workers who were paid more were no more motivated than workers paid less, because they perceived their work to be more valuable. Feng et al. (2009) also found in their experiment that paying $0.10 per task generated lower quality of data than paying $0.05.

In order to obtain reliable results, quality control is very important when using Amazon Mechanical Turk, as such monetary incentive based crowdsourcing services are the most susceptible to fraud (Kazai et al., 2009; Kazai, 2011). In order to obtain reliable results, quality control is important (Kazai, 2011). Collecting multiple assessments for each task is probably the most popular strategy to control quality and has been used extensively in almost every experiment. Multiple assessments can be simply aggregated as an average (Alonso & Mizzaro, 2009; Tang & Sanderson, 2010), voting scheme or weighted sum as suggested by Alonso & Rose (2008).

Feng et al. (2009) found that seven crowdsourced workers per task seemed to be a reasonable number for their experiment. They utilised a two-phase framework for acquiring high quality data from AMT. In the first phase, *validation*, a small number of tasks were sent to AMT to find malicious workers (via inter-person agreement), the optimal number of workers per task, and the optimal pay rate. The second phase, large-scale submission phase, submitted all tasks using the parameters from the first phase. Sorokin et al. (2008) tried to encourage the participants to follow the task protocol by injecting gold standard into the image annotation process. If the annotations provided deviated significantly from the gold standard, the gold standard would be shown to the participants to warn them after the task was submitted. Tang & Sanderson (2010) attempted to eliminate noise by planting "traps" in each task. Only the work from those who answered all "traps" correctly was accepted. Snow et al. (2008) used the Pearson correlation coefficient score of one annotator compared to the average of all annotators as a measure of inter-annotator agreement. Submissions with a low agreement with the others were filtered out. Sheng et al. (2008) proposed the use of selective repeated-labelling, a supervised learning technique that

models the uncertainty of labels and chooses which examples should get more labels, to improve the quality of data labelling.

According to Down et al.'s (2010) study, young men were the most likely to play the system, while men over 30 and women of any age were more reliable; Professionals, students and non-workers were more likely to take the tasks more seriously than financial workers, hourly workers and other workers. AMT itself provides a facility called *qualification* to assist quality control. Requesters can attach different qualification requirements to their tasks so that workers must meet the requirements before they are permitted to work. A popular and useful qualification type is approval rate, which is the percentage of accepted tasks completed by a worker since joining AMT, reflecting their overall credibility. The qualification test includes a set of questions and can be constructed by requesters to allow only workers who give correct answers to work on the tasks. However, this cannot ensure workers do not cheat after passing the tests. Generally, the use of multiple quality control methods brings more reliable results.

4.3. Crowdsourcing in the Cultural Heritage domain

Crowdsourcing in cultural heritage (CH) is not new, existing long before the digital era and Upshall (2011) cites the Oxford English Dictionary as one of the largest and most successful crowdsourcing projects in history, using volunteers to provide provenance for the use of words in original source materials. In CH today, many of the largest crowdsourcing projects are focused on enhancing digitisation efforts, either by improving the quality of digitised content (e.g. via editing, proof-reading or transcribing), or by adding metadata (e.g. tags) to improve information retrieval capabilities and accessibility. A useful overview of current crowdsourcing activity in CH is provided by Oomen & Aroyo (2011), including a classification of the different types of crowdsourcing initiatives to be found in the sector and a summary of some notable projects, and Holley (2012) provides similar classifications and context for digital libraries, with a summary of the two classifications given in Table 1. In addition, Ridge (2012) provides a more personal commentary and running account of issues arising within the context of her ongoing PhD research in this area.

Oomen & Aroyo, 2011	**Holley, 2012**
Co-curation	Opinions
(no equivalent)	Ideas
Contextualisation (contributing knowledge & stories)	Adding knowledge
Classification	Categorising and classifying
Correction & transcription tasks	Skills requiring human eye/hand
Complementing collection (contributing objects)	Creation of content
Crowdfunding	Raising funds

Table 1: Comparison of Classification of Crowdsourcing Opportunities in CH and Libraries

Crowdsourced human effort in major digitisation projects is utilised to overcome some of the issues arising from automation. Scanning of images is relatively quick to do, but post-processing takes more time and some degree of human judgement to complete successfully, so the V&A uses volunteers to crop digitised images to show them at their best in the online Search the Collections catalogue. Like Project Gutenberg, the newspaper

archives in the National Library of Australia's Trove project uses scanning technologies to capture large quantities of text materials, which then requires proof-reading and editing to eradicate errors that are an inherent to OCR (optical character recognition). Earlier manuscripts are much more difficult to scan successfully and may even require full transcription, as is the case of the handwritten documents digitised via the Transcribe Bentham project at UCL (Moyle, et al, 2010; Causer, et al, 2012).

Once items have been digitised there is a need to add metadata so that they can be found via search tools. Again this is more labour-intensive than the initial capture of digital text and/or images, and often requires a degree of interpretation that can only be achieved via human effort. Projects such as What's The Score (Bodleian Library) use crowdsourced volunteers to create initial metadata, although the more prevalent approach is to source additional social metadata in the form of keywords or 'tags'. Social metadata can be a useful aid to information access and discovery for non-expert users (i.e. those without detailed subject and domain knowledge) of online collections, as the resulting 'folksonomies' tend to be more descriptive and use everyday terminology compared with more specialist taxonomies. Leading the way in CH research for social tagging is the Steve Museum project (Trant, 2008), with more recent high profile public projects including Your Paintings, which is a collaboration between the BBC and Public Catalogue Foundation aimed at tagging all artworks held in the UK national collections.

In terms of enhancing the context, content and presentation of collections, projects involving co-curation and the acquisition of user-generated content and objects are also proving to be fruitful areas for crowdsourcing. High profile examples include Brooklyn Museum's 2008 Click! exhibition of photographic images in which many thousands of voters helped to prioritise images for display, and the ongoing Europeana 1914-18 project that is encouraging contributions from the general public of artefacts and stories to Oxford University's Great War Archive.

A comprehensive list of 30 considerations for the design of successful CH crowdsourcing projects is provided by Visser (2011). One important element of success in CH crowdsourcing projects is gaining momentum and sustaining effort over the long-term. Motivation is a key element of engaging users and encouraging continued participation in crowdsourcing and other distributed human computing projects (Quinn & Bederson, 2011). Participants in CH crowdsourcing are motivated variously by altruism, being interested in the content, having fun (particularly in the case of game-based projects), learning, credit and recognition, and being part of a community (Holley, 2010, Visser 2011), rather than by the extrinsic motivation of monetary gain as in the Mechanical Turk examples above (Oomen & Aroyo, 2011). Tasks also often become addictive and are more successful if they are kept relatively simple in their requirements (Holley, 2011). This finding is borne out in an analysis of non-participation by registered users in Steve Museum (Trant 2009), with reasons including users being too busy and being sceptical, but also not understanding the task and feeling unqualified.

The success of crowdsourcing projects is also determined by the quality of the work, output and contributions derived from the crowd. In community-based systems such as Wikipedia, quality control is achieved via verifying and editing by other participants. Other systems (e.g. WAISDA and Click!) rely on volume participation and achieving consensus. In fact, it was found that by using a large enough crowd and by randomising the images presented and allowing only single votes per person per image, selections of images for the Click! exhibition by non-experts had a significant degree of overlap with those of experts, disproving the claims of dumbing-down (Surowiecki, 2008). Establishing behavioural norms and expectations of the quality required further aids quality standards to be maintained (Oomen & Aroyo, 2011).

Project	Organisation	Type	Comments
Wiki Loves...	Wikipedia	Content sourcing	A number of related projects (Art, Libraries, Museums, Monuments) focused on sourcing images to support GLAM content in Wikipedia.
Steve Museum	Steve Museum (US collaborative project)	Tagging	Adding user-generated tags to fine art images.
WAISDA	Netherlands Institute for Sound & Vision	Tagging	A game-based tagging system to aid information retrieval in audio-visual archives.
Your Paintings	BBC / Public Catalogue Foundation	Tagging	Adding user-generated tags to fine art collections.
V&A Catalogue	Victoria & Albert Museum	Editing	Cropping digitised images to enhance their usability.
Transcribe Bentham	UCL	Editing / Transcribing	Transcribing the full text of early manuscripts.
Trove	National Libraries of Australia	Editing	Correcting OCR errors in scanned historic newspapers.
Click!	Brooklyn Museum	Curating	Rating images selected for an exhibition as part of the curation process.
What's The Score	Bodleian Library	Cataloguing	Adding metadata to scanned music scores.
Library Thing	Library Thing	Tagging	Adding user-generated tags and other social metadata to books.
YUMA	Austrian Institute of Technology / EuropeanaConnect	Annotation	Originally for annotation and tagging of historic maps, now applicable to multiple media types.
Europeana 1914-18	Europeana / Oxford University	Collecting	Gathering objects from users relating to World War One to add to official collections.
FlickrCommons	Flickr/various CH institutions	Annotations /tagging	Soliciting user-generated annotations (comments and tags) to images from historic CH collections.

Table 2: Examples of Crowdsourcing Projects in the Cultural Heritage

Benefits of crowdsourcing in CH accrue to both the institutions involved and to the participants. For the institution these might include strong economic gains from access to free or very low-cost volunteer labour, access to content and ideas that would be otherwise unavailable, and improved provision in terms of search and discovery for their online collections. For participants, it has been reported that much satisfaction is gained from exploration of the digitised collections, feeling a sense of purpose, recognition and belonging, and in the case of genealogy and local history researchers, furthering the success of their own projects.

5. Mobile Web and Applications

The marriage of mobile devices and cultural heritage is nothing new. From hand held audio guides to the use of PDAs in projects like the Smart Museum[11] as recently as 2009, mobile technology of one sort or another is a familiar feature at many cultural heritage sites. However, the unveiling of the iPhone in January 2007, and its release on 29 June that year, lead to the general use of mobile technology for accessing online content to become increasingly commonplace. A recent Pew Internet study (Smith 2012) showed that a majority of American adult mobile phone owners go online using their phone. If the age is restricted to just 25-35, 80% of mobile phone users go online using their handheld device. A Europeana-commissioned study in 2011, Culture on the Go[12], made the importance of mobile access to cultural heritage data very clear and predicts that by the end of this year, 18% of all traffic to Europeana will be from mobile devices.

The growth in mobile Web usage is the driver behind a great deal of work in standardisation and implementation of those standards. This presents many opportunities for developers, including within PATHS, as the project begins to consider its own mobile application in the final year.

As well as the new standards-based features and capabilities of modern devices, another driver for change is the huge competition between the device manufacturers. As a result, an ever-growing number of different devices of different sizes with different capabilities is now available. Such is the rapidity of change that developers face a real challenge in knowing which features are supported, or rather, which features have sufficient support that they can be relied upon to be available to the majority of end users.

Two methods are now recommended for handling these issues:

1. Progressive Enhancement;
2. Responsive Web Design.

Progressive Enhancement[13] addresses the issue of feature support. It encourages developers to assume that the target device supports only minimum capabilities that can be relied upon to deliver a functional user experience. More advanced device capabilities are used, but in a way that enhances the baseline experience without depending on them. For example, if an application needs to know a user's location, the default would be to ask the user to enter it manually. Progressive Enhancement would call on the device to replace the input form with a suitable message if it is possible to report the location automatically as modern devices can all do.

Responsive Web Design (Marcotte 2010) handles the issue of different device sizes and depends on one particular technology: CSS Media Queries. This allows designers to set styles depending on the available screen size. Although only recently reaching W3C Recommendation status (Rivoal 2012), it has been around a long time and nearly became a standard back in 2007[14]. Although not implemented on older versions of Internet Explorer, it is almost universally available on mobile[15] and, since it follows the progressive enhancement

[11] http://www.smartmuseum.eu/
[12] http://kwz.me/Hu
[13] http://www.alistapart.com/articles/understandingprogressiveenhancement/
[14] http://www.w3.org/TR/2007/CR-css3-mediaqueries-20070606/
[15] http://caniuse.com/css-mediaqueries

approach, can be used effectively and with confidence. When following Responsive Web Design, the developer begins with a layout optimised for mobile phones, i.e. narrow screens, and then replaces relevant styles to define wider layouts as more screen space becomes available. The W3C Web site provides an example of this at http://www.w3.org/.

Both Progressive Enhancement and Responsive Web Design are sufficiently proven techniques to be included in the World Wide Web Consortium's online training courses[16].

Many powerful 2D image technologies are now widely available on mobile, including Scalable Vector Graphics (SVG) and HTML5 Canvas. Support for CSS rounded corners and box shadows has been implemented in all the major browsers and 3D support is starting to grow. One of the most talked about features of HTML5 is the video playback element. This is widely implemented in modern browsers, however, HTML5 does not specify the video encoding mechanism to use. Rather, even to reach all HTML5 conformant browsers, any video must be made available in multiple formats. The most up to date source of information on this particular issue is the relevant Wikipedia page[17].

One proposed Web technology that has not yet been adopted by the browsers is responsive images. A large image delivered to a small device is a waste of bandwidth, battery and processing power and may cost the user money to download. Conversely, a small image delivered to a desktop browser but scaled up to look large will offer poor quality. Discussion around this issue appeared to have reached consensus around a solution based exactly on the HTML5 video compromise. However, an alternative proposal was made in controversial circumstances[18] and so the matter remains unresolved and the discussion continues[19]. For now, server side detection of desktop or mobile environments is required to ensure that the right image is sent to the right class of device.

One of the key features of HTML5 is its emphasis on offline as well as online usage. To this end, modern browsers support data storage directly without having to send and receive it from the server (although best practice is to replicate the data on the server when connection is available). The most widely supported in-browser data storage mechanism is Web Storage[20], which allows simple name/value pairs to be stored. Standardisation of access to the device's file system is under way but it not yet sufficiently deployed to be useful to PATHS.

It remains the case that 'native apps' - that is, applications developed for specific platforms, notably Apple's iOS and the Android platform - do have greater access to device capabilities than Web applications. However, by their nature they exclude user of other devices and therefore the Progressive Enhancement and Responsive Web Design techniques become irrelevant. A developer must make separate versions of applications for use on iPhone and iPad for instance. As the existence of online resources such as the W3C's Standards for Web Applications on Mobile: current state and roadmap (Hazaël-Massieux, 2012) and the tables provided at caniuse.com show, the landscape continues to evolve rapidly.

[16] http://www.w3devcampus.com/mobile-web-and-application-best-practices-training/
[17] http://en.wikipedia.org/wiki/HTML5_video
[18] http://www.netmagazine.com/news/html5-responsive-images-spat-explodes-121961
[19] http://www.w3.org/community/respimg/
[20] http://www.w3.org/TR/webstorage/

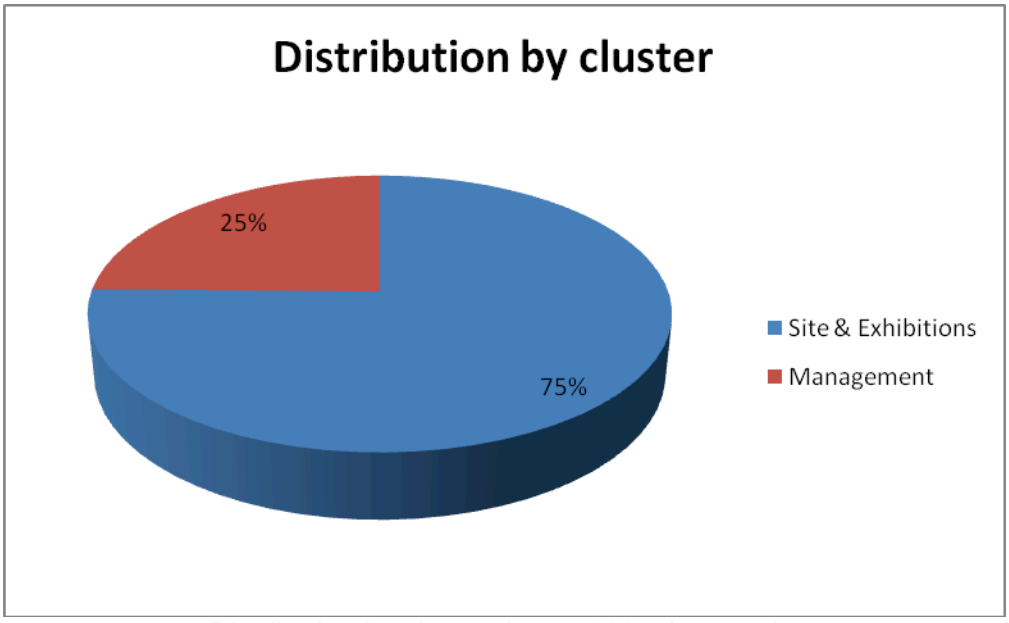
6. Sentiment Analysis

The Initial State of the Art Monitoring Report, D1.2, presented an analysis of sentiment towards Europeana as expressed online from its launch in November 2008 to February 2011. The analysis has been re-run to cover the period June 2011 – June 2012 and from this it is possible to note changes for the better in the perception of Europeana and attitudes towards it.

As before, the data was segmented in two clusters.

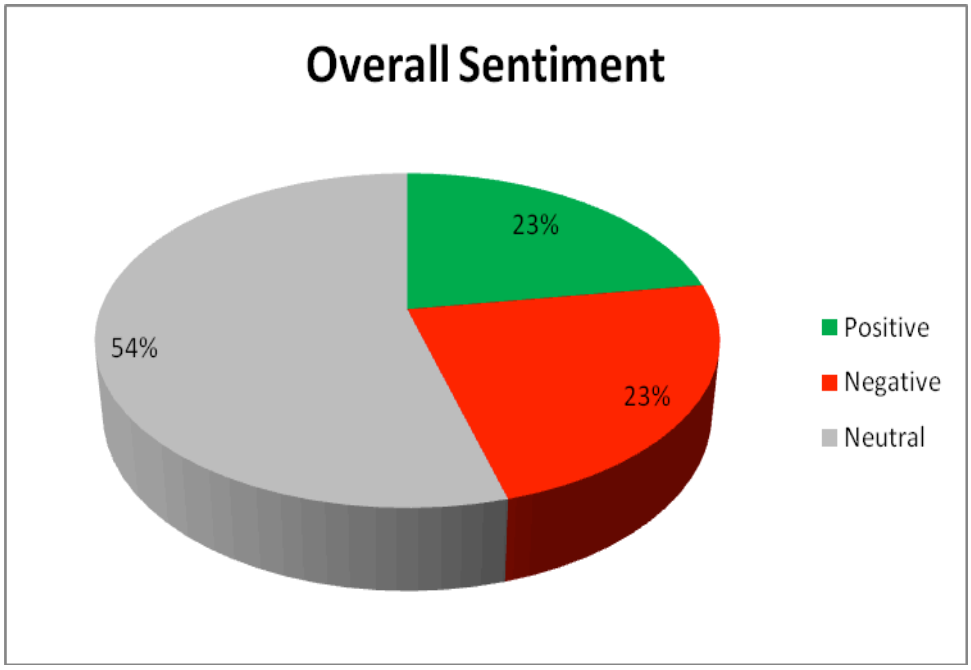
> Cluster 1 refers to the content on the actual site and the collections presented there
>
> Cluster 2 refers to comments and issues related to the management of the endeavour, e.g. Co-operation with other organisations and news on relevant EU policies, etc.

Below is the distribution of each cluster – note that the respective cluster sizes remain stable from the previous research.



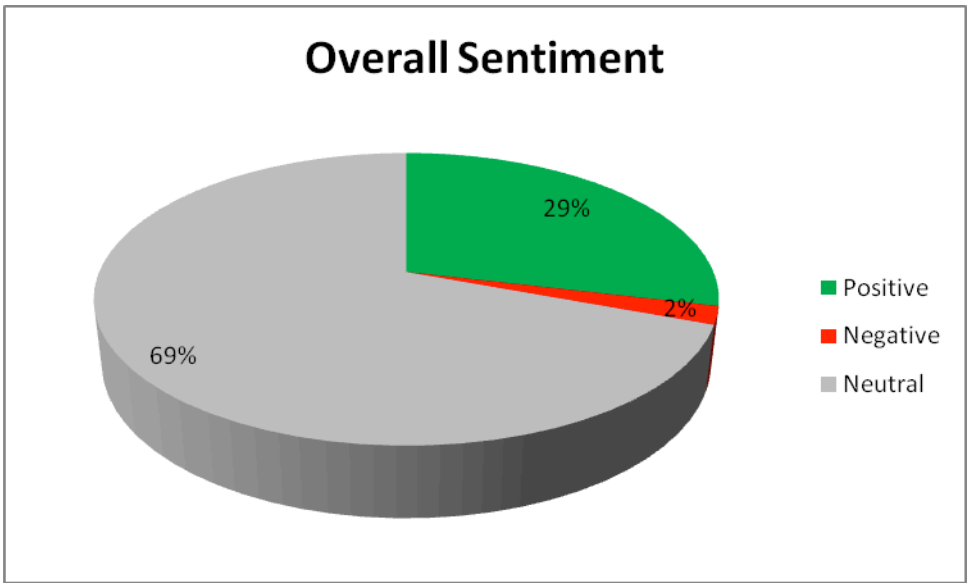
Distribution by cluster June 2011 - June 2012

In the original research, where sentiment was expressed, it was evenly divided between positive and negative.



Overall Sentiment November 2008 - February 2011

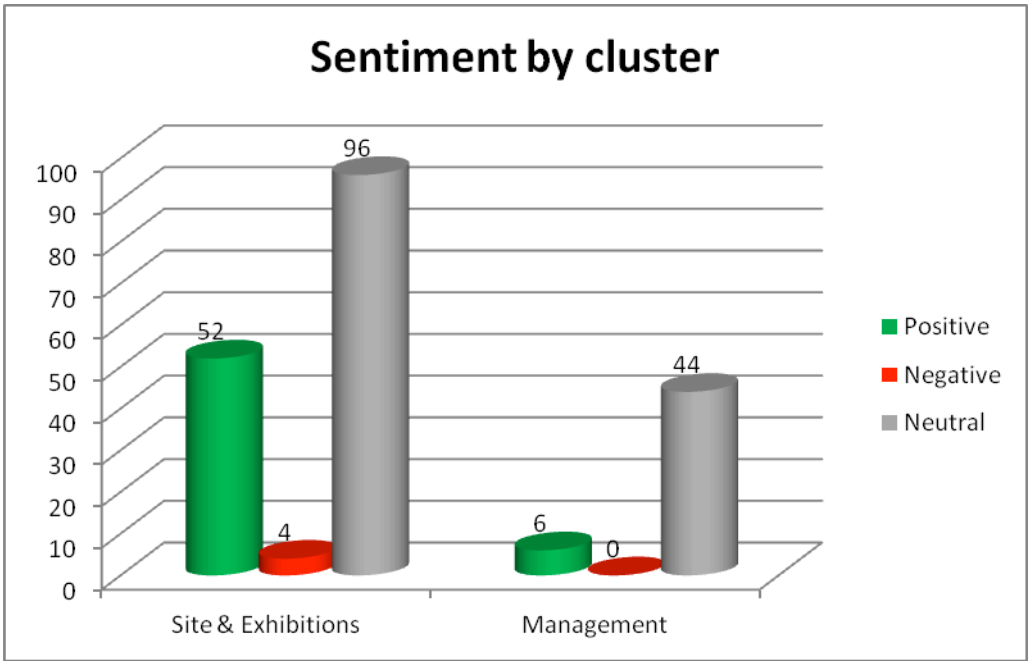
In the updated survey, the majority of online published references to Europeana remain neutral, in the sense that they come from news sites and reproductions of announcements relevant to projects and expansions of the digital library, for example: *Oxford University is providing expert advice to the Europeana 1914-1918 project which runs history roadshows*[21]. Positive references remain about about one third of the total, however, the negative sentiment has practically disappeared. It is noteworthy that in the previous research, the negative comments were mostly on the issues of the early server collapse on the first attempt to go live.



Overall Sentiment June 2011 - June 2012

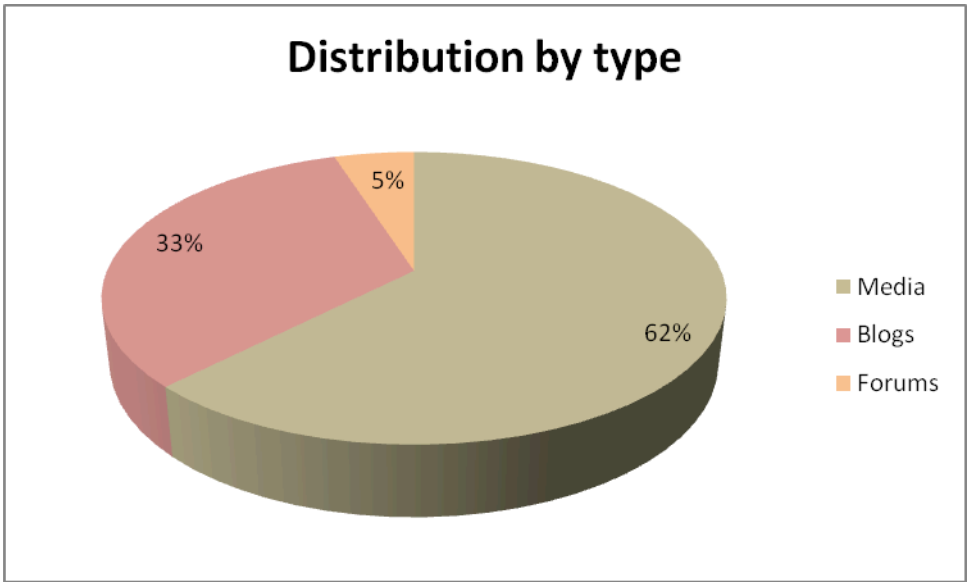
[21] http://www.city-data.com/forum/history/1565163-hitler-postcard-found-world-war-i.html

The following chart shows the breakdown of sentiment distribution by cluster.



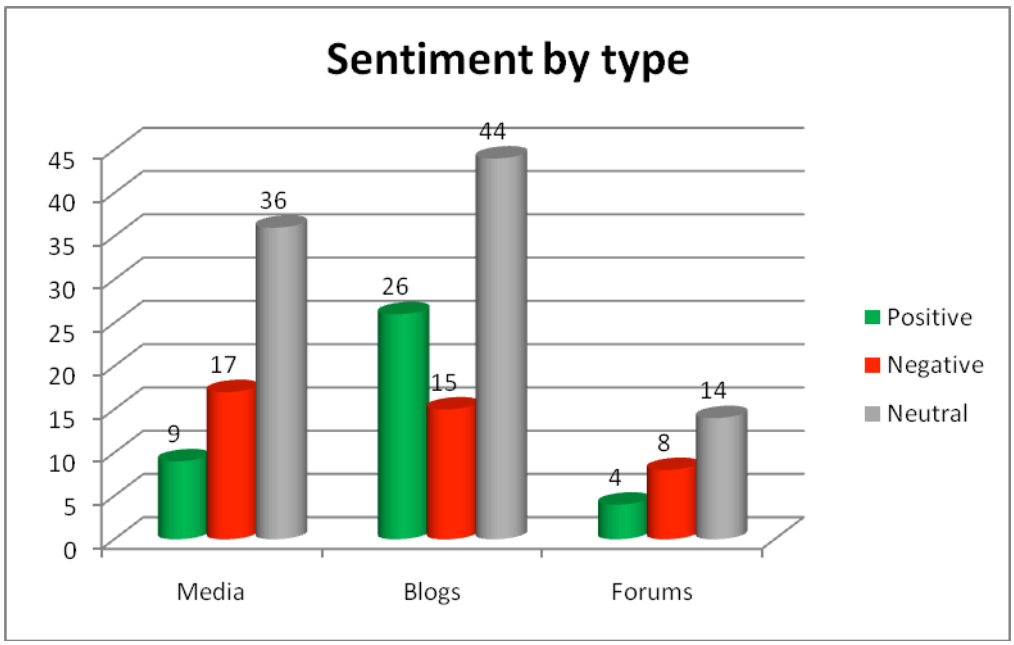
Sentiment by cluster June 2011 - June 2012

More than half the mentions appear to come from media sites, when in previous years blogs contributed 49% and forums 15% to the overall buzz. This is almost certainly to be taken as a sign of a succesful media campaign, however, the “grassroots” buzz on Europeana (blogs and forums) seems to be lagging behind.

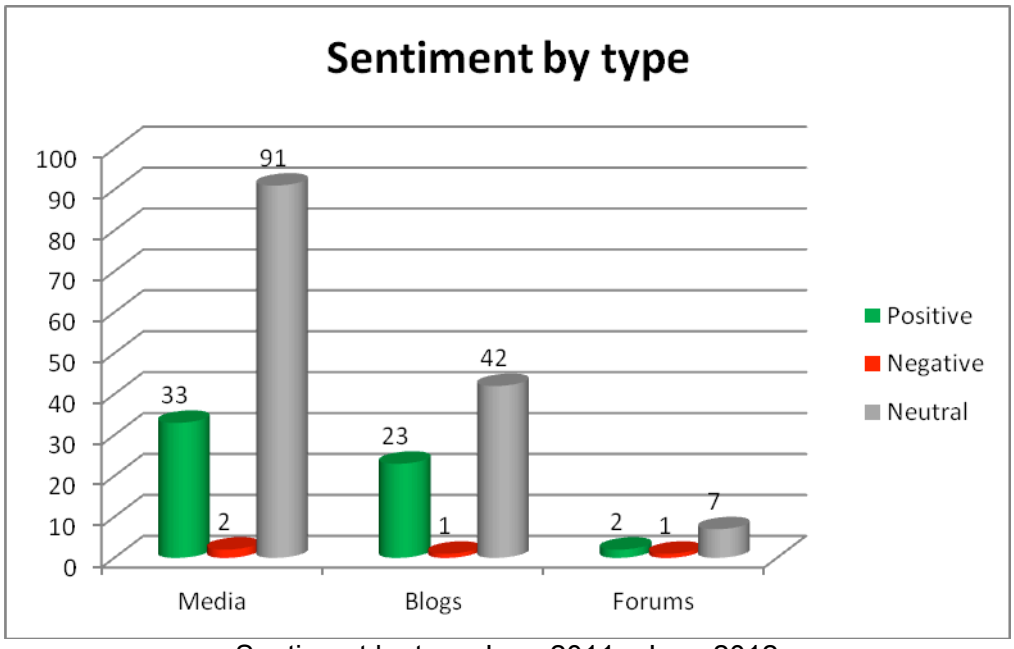


Media sites and blogs account for most positives, for example this from the Spittal Street blog: *The Europeana portal which has now gone far beyond expectations, should become*

the central reference point for Europe's online cultural heritage[22]. Again, the near-disappearance of negative comments since the initial survey is the most pronounced change.



Sentiment by type November 2008 - February 2011



Sentiment by type June 2011 - June 2012

Now that the initial site crash and re-launch are no longer part of the story, the four most important sentiment drives can be assessed.

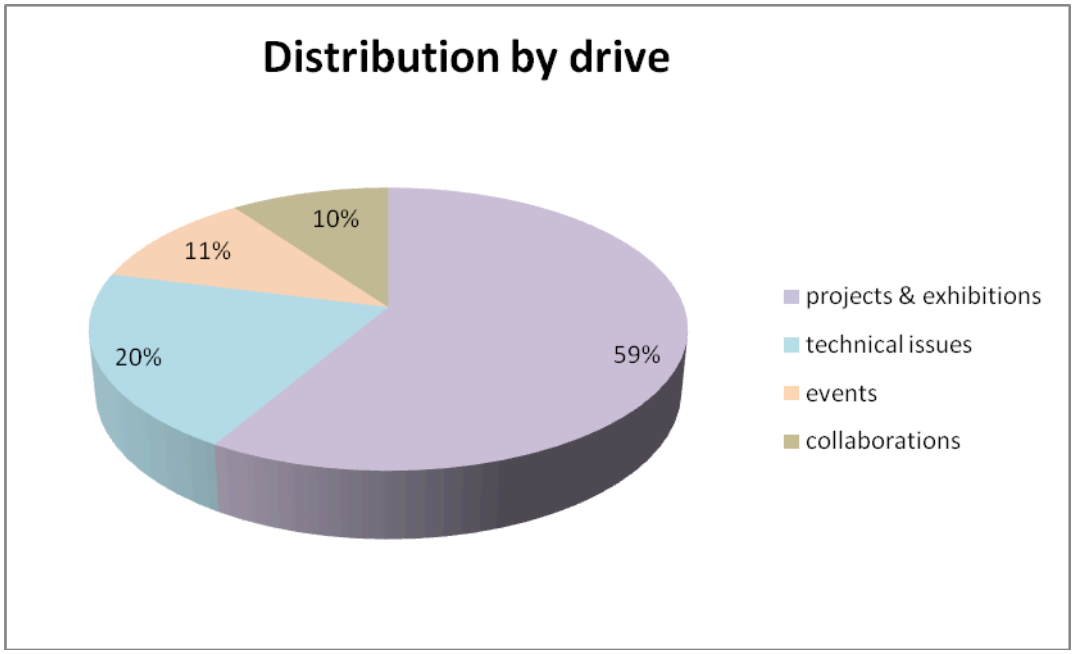
The hottest topic is the actual collections and the various projects in support of Europeana – most notably the one referring to the First World War, which claims 28 out of the 100 references to collections and virtual exhibitions.

[22] http://spittalstreet.com/?p=867

Next comes technology and particularly the Linked Open Data plan of Europeana.

Conferences and workshops for the development of Europeana follow suit – most notably the Europeana Plenary conference.

Finally, various announcements of collaborations and enrichment efforts also drive online discussions and sentiment.



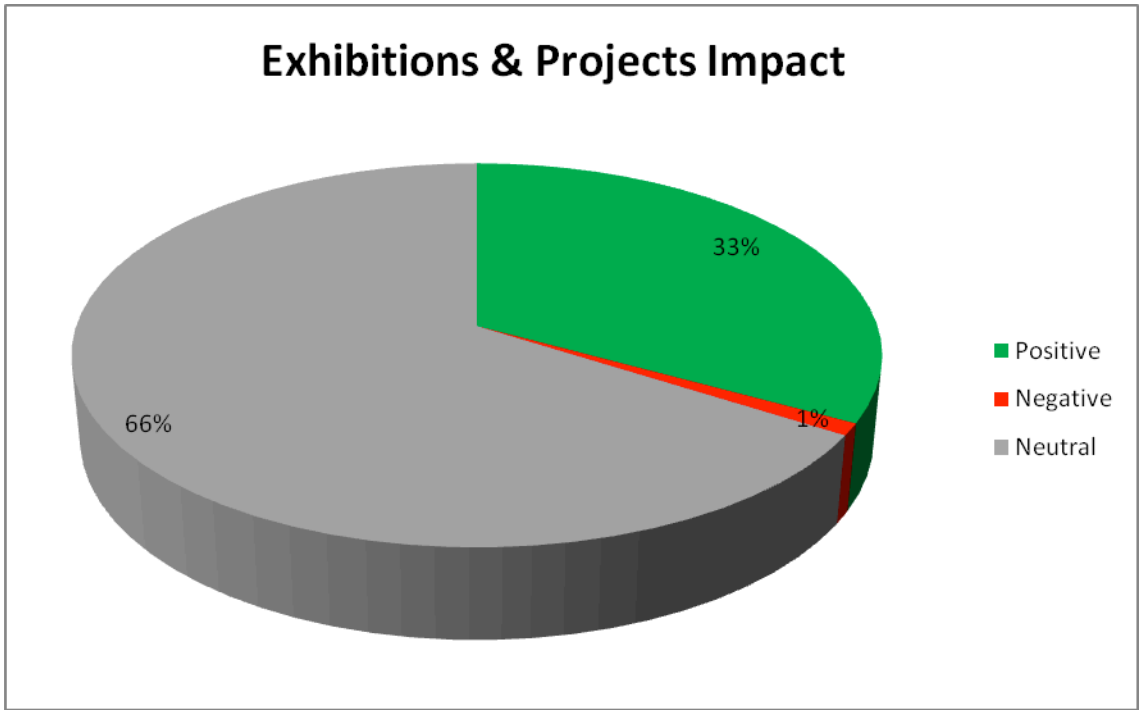
Distribution by drive June 2011 - June 2012

The buzz on the the various projects and virtual exhibitions is depicted below. Buzz is a measure of the impact of a given article, based on factors such as the number of links to it, the number of comments received on a blog post and so on. Among the positive comments there is a clear acknowledgement of the scientific and social merit of Europeana, for example: *Europeana's project is remarkable in that it can bring in the most meaningful story in a person's whole existence, as well as the mundane and mis-spelt jottings of the future dictator*[23], as well as an understanding of what it means for the entire cultural heritage sector across Europe: *And thanks to digitization efforts and Europeana, much of this heritage can also be found online*[24]. Still, in the very few negatives, the issue of open licencing seems to remain a cause for concern. For example, this comments refers to the National Library of Sweden: *The National Library also expressed concern about being unable to deliver bibliographic data to the European digital library project Europeana, as "Europeana aims at only obtaining data that can be made available under open license.*[25]

[23] http://www.culture24.org.uk/history%20%26%20heritage/war%20%26%20conflict/world%20war%20one/art384751

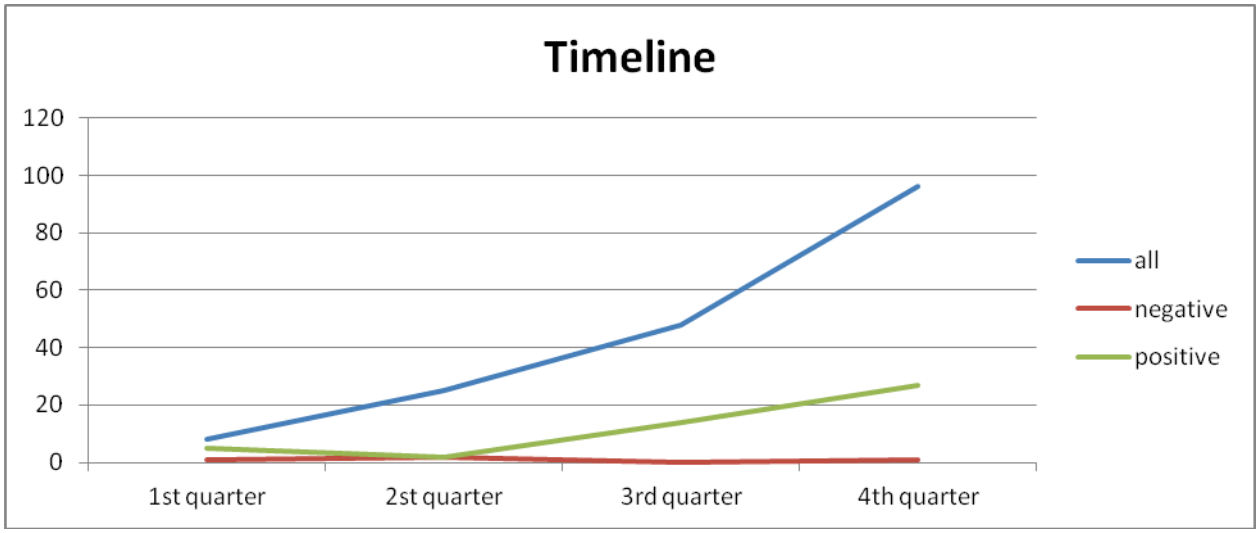
[24] http://www.readwriteweb.com/archives/hacking_europes_cultural_heritage_with_europeanas.php

[25] http://www.libraryjournal.com/lj/home/893131-264/national_library_of_sweden_no.html.csp



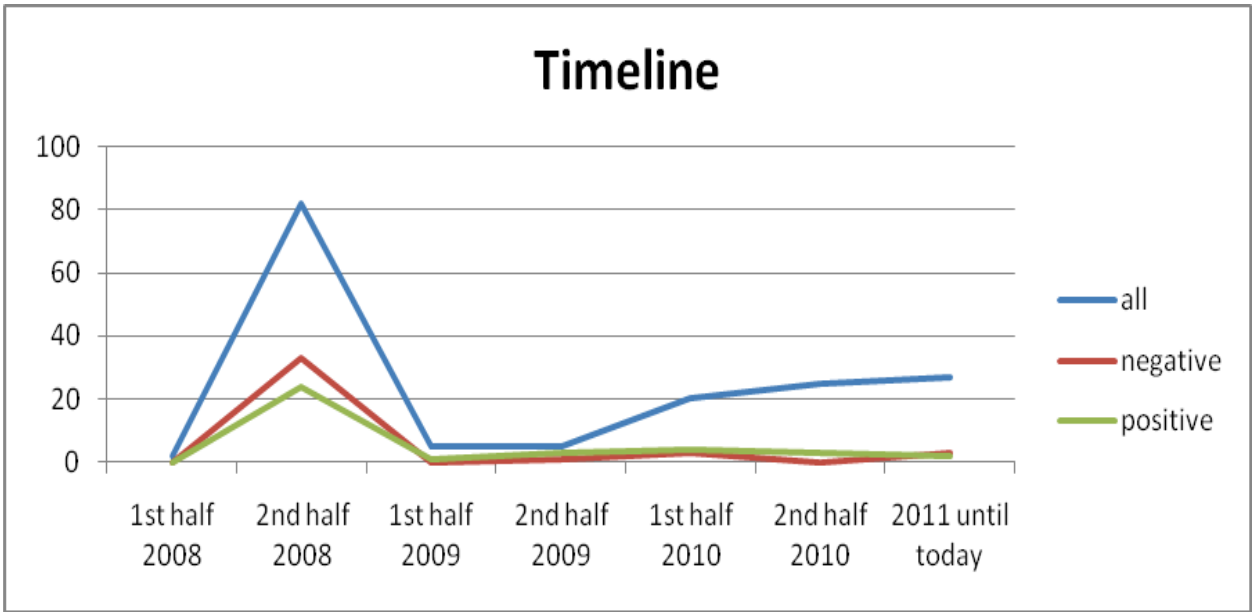
Exhibitions and projects Impact, June 2011 - June 2012

The timeline of references to Europeana below indicates a rise in early 2012. This can be attributed to the discussions on new virtual exhibitions (1914 – 1918 and Newspapers), as well as new technical issues raised (Open Linked Data) and upcoming events, notably the Plenary.



Timeline of all relevant comments detected, June 2011 - June 2012

It is a further sign of Europeana's increasing success and visibility that the number of relevant comments detected is now above that achieved around its launch in 2008 and that the positive comments are now much more noticeable compared with the initial study.



Timeline of all relevant comments detected, November 2008 - February 2011

The raw data is available in Appendix 1.

7. References

Agirre, E., Edmonds, P. (Eds.) (2006). Word Sense Disambiguation: Algorithms and applications. Text, Speech and Language Technology , Vol. 33. 2006

Alonso, O., & Baeza-Yates, R. (2011) Design and implementation of relevance assessments using crowdsourcing. In *Proceedings of the 33rd European conference on Advances in information retrieval* (ECIR'11), Paul Clough, Colum Foley, Cathal Gurrin, Hyowon Lee, and Gareth J. F. Jones (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 153-164.

Alonso, O., & Mizzaro, S. (2009) Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment, In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 15–16.

Alonso, O., Rose, D. E., & Stewart, B. (2008) Crowdsourcing for relevance evaluation. *ACM SIGIR Forum*, Vol. 42(2), pp. 9–15.

Baddeley, A. (1968). A 3 min reasoning test based on grammatical transformation. *Psychometric Science*, *10*, 341-342.

Bentivogli, L. Forner, P. Giuliano, C. Marchetti, A. Pianta, W. Tymoshenko, L. 2010 "Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia". In Proceedings of COLING 2010 Workshop on "The People's Web Meets NLP: Collaboratively Constructed Semantic Resources", Beijing, China, August 28 2010.

Bryl, V. Giuliano, C. Serafini, L. Tymoshenko, K. Using Background Knowledge to Support Coreference Resolution. ECAI. 2010.

Callison-Burch, C. (2009) Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1* (EMNLP '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 286-295.

Carvalho, V. R., Lease, M., & Yilmaz, E. (2011) Crowdsourcing for search evaluation. *ACM SIGIR Forum,* Vol. (44), pp. 17–22.

Causer, T. et al, (2012) Transcription Maximized; Expense Minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham, *Library and Linguistic Computing*, 27(2) 119-137.

Chamorro-Premuzic, T. and Reichenbacher, L. (2008). Effects of Personality and Threat of Evaluation on Divergent and Convergent Thinking. *Journal of Research in Personality*, 42 (4), pp. 1095–1101.

Cucerzan. S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In proceedings of EMNLP-CoNLL. 2007.

Fellbaum, C. WordNet: An electronic lexical database. The MIT press, 1998.

Dakka, W., & Ipeirotis, P.G. (2008) Automatic Extraction of Useful Facet Hierarchies from Text Databases, In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* (ICDE '08). IEEE Computer Society, Washington, DC, USA, pp. 466-475.

Doan, A., Ramakrishnan, R., & Halevy, A.Y. (2011) Crowdsourcing systems on the World-Wide Web, *Communications of the ACM*, Vol. 54, pp.86-96.

Downs, J.S., Holbrook, M.B., Sheng, S., & Cranor, L.F. (2010) Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, NY, USA, pp. 2399-2402.

Estellés Arolas, E., & González Ladrón-de-Guevara, F. (2012) Towards an integrated crowdsourcing definition, *Journal of Information Science*, Vol. 38(2), pp. 189-200.

Felder and Soloman, *Index of Learning Styles Questionnaire* http://www.engr.ncsu.edu/learningstyles/ilsweb.html

Feng, D., Besana, S., & Zajac, R. (2009) Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (People's Web '09). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 51-56.

Ferragina, P and Scaiella, I. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). ACM, New York, NY, USA, 1625-1628.

Hazaël-Massieux, D. (2012) Standards for Web Applications on Mobile: current state and roadmap http://www.w3.org/2012/05/mobile-web-app-state/

Holley, R. (2012) Harnessing the Cognitive Surplus of the Nation: New opportunities for libraries in a time of change, *Jean Arnot Memorial Fellowship Essay 2012*, State Library of New South Wales, Australia, Available: http://www.sl.nsw.gov.au/about/awards/docs/Arnot_Memorial_Fellowship_Winner%202012.pdf [Accessed: 22/06/2012]

Holley, R. (2010) Crowdsourcing: How and Why Should Libraries Do it? D-Lib Magazine 16(3/4) [online], Available: http://www.dlib.org/dlib/march10/holley/03holley.html, [Accessed: 20/06/2012].

Holley, R. (2011) *Crowdsourcing and Social Engagement in Libraries: the state of play*, ALIA Sydney blog post 29th June 2011 [online], Available: http://aliasydney.blogspot.co.uk/2011/06/crowdsourcing-and-social-engagement-in.html, [Accessed: 22/06/2012].

Hosseini, M., Cox, I. J., Milic-Frayling, N., Kazai, G., & Vinay, V. (2012) On Aggregating labels from Multiple Crowd Workers to Infer Relevance of Documents, In *Proceedings of 34th European Conference on Information Retrieval,* Springer, LNCS 7224, pp. 182-194.

Howe, J. (2006) The Rise of Crowdsourcing, Wired magazine.

Jimenez, S. Becerra, C. Gelbukh, A. Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In Proceedings of the 6th International Workshop on Semantic Evaluation, in conjunction with the 1st Joint Conference on Lexical and Computational Semantics. Montreal, Canada. 2012

Kaisser, M., & Lowe, J.B. (2008) A Research Collection of Question Answer Sentence Pairs, In *Proceedings of Conference on Language Resources and Evaluation* (LREC'08).

Kaisser, M., Hearst, M., & Lowe, J.B. (2008) Evidence for Varying Search Results Summary Lengths, In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics* (ACL'08).

Kazai, G. (2011) In search of quality in crowdsourcing for search engine evaluation In *Proceedings of the 33rd European conference on Advances in information retrieval* (ECIR'11), Paul Clough, Colum Foley, Cathal Gurrin, Hyowon Lee, and Gareth J. F. Jones (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 165–176.

Kazai, G., Milic-Frayling, N., & Costello, J. (2009) Towards methods for the collective gathering and quality control of relevance assessments, In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*(SIGIR '09). ACM, New York, NY, USA, pp. 452-459.

Kittur, A., Chi, E.H., & Suh, B. (2008) Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI '08). ACM, New York, NY, USA, pp. 453-456.

Kulkarni, S. Singh, A. Ramakrishnan, G. Chakrabarti, S. Collective annotation of Wikipedia entities in web text. KDD 2009: 457-466. 2009.

Li, Y. McLean, D. Bandar, ZA, O'Shea JD Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, August.

Marcotte, E. (2010) Responsive Web Design http://www.alistapart.com/articles/responsive-web-design/

Mason, W., & Watts, D.J. (2009) Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (HCOMP '09), Paul Bennett, Raman Chandrasekar, Max Chickering, Panos Ipeirotis, Edith Law, Anton Mityagin, Foster Provost, and Luis von Ahn (Eds.). ACM, New York, NY, USA, pp. 77-85.

Medelyan, O. Milne, D. Legg C. and Witten, IH. Mining meaning from Wikipedia. International Journal of Human-Computer Studies. Volume 67, Issue 9, pp. 716-754. 2009.

Mihalcea, R., Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In CIKM, volume 7, pages 233–242, 2007.

D. Milne and I.H. Witten. Learning to link with wikipedia. In Proceeding of the 17th ACM conference on Information and knowledge management, pages 509–518. ACM, 2008.

Moyle, M. et al, (2010) Manuscript Transcription by Crowdsourcing: Transcribe Bentham, *LIBER Quarterly* , 20(3/4).

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

Oomen, J. & Aroyo, L. (2011) Crowdsoucing in the Cultural Heritage Domain: Opportunities and challenges, In: *Proceedings C&T '11, 29 June-2 July 2011, QUT, Brisbane, Australia*, New York: ACM Digital Library.

Peterson, E.R., Deary , I.J. and Austin, E J. (2002) The reliability of Riding's Cognitive Style Analysis test. *Personality and Individual Differences* 34 (2003) pp. 881–891.

Quinn, A. J. & Bederson, B. B. (2009) *A Taxonomy of Distributed Human Computation*, Tech. Report HCIL-2009-23, University of Maryland, Available: http://hcil2.cs.umd.edu/trs/2009-23/2009-23.pdf, [Accessed: 20/06/2012].

Ridge, M. (2012) *Frequently* Asked *Questions about crowdsourcing in cultural heritage*, Open Objects blog post 3rd June 2012 [online], Available http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html. [Accessed: 20/06/2012].

Riding, R. *Cognitive Styles Analysis* (Learning and Training Technology, Birmingham, 1991).

Riding, R. and Sadler-Smith,E. (1992) Type of instructional material, cognitive style and learning performance, Educational Studies 18(3) pp. 323-340.

Riding, R. (1003) *A Trainers' Guide to Learning Design* (Department of Employment, Sheffield).

Riding, R. and Douglas, G. (1993) The effect of cognitive style and mode of presentation on learning performance, *British Journal of Educational Psychology* 63(2) (1993), pp.297-307.

Rivoal, F. (2012) Media Queries W3C Recommendation http://www.w3.org/TR/css3-mediaqueries/

Sanderson, M., Paramita, M., Clough, P., & Kanoulas, E. (2010) Do user preferences and evaluation measures line up?, In *Proceedings of the 33rd Annual ACM SIGIR Conference* (SIGIR'10), Geneva, Switzerland, pp. 555-562.

Sheng, V.S., Provost, F., & Ipeirotis, P.G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '08), ACM, New York, NY, USA, pp. 614-622.

Shirky, C. (2010) *Cognitive Surplus: Creativity and Generosity in a Connected Age*, London: Allen Lane.

Smith, A (2012) Cell Internet Use 2012, Pew Internet http://pewinternet.org/Reports/2012/Cell-Internet-Use-2012/Main-Findings/Cell-Internet-Use.aspx

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A.Y. (2008) Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 254-263.

Sorokin, A., & Forsyth, D. (2008) Utility data annotation with Amazon Mechanical Turk, In *Proceedings of Computer Vision and Pattern Recognition Workshops, 2008* (CVPRW '08), pp.1-8.

Su, Q., Pavlov, D., Chow, J-H., & Baker, W.C. (2007) Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07), ACM, New York, NY, USA, pp. 231-240.

Surowiecki,J. (2008) *Reflection on Click!,* Brooklyn Museum blog post 8th August 2008[online], Available: http://www.brooklynmuseum.org/community/blogosphere/2008/08/08/reflections-on-click-by-james-surowiecki/ [Accessed: 20/06/2012]

Tang, J., & Sanderson, M. (2010) Evaluation and User Preference Study on Spatial Diversity, In *Proceedings of 32nd European Conference on Information Retrieval,* Springer (ECIR'10), pp. 179-190.

Tonelli, S and Giuliano, C. Wikipedia as Frame Information Repository. In Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing (EMNLP 2009), Singapore, August 6-7, 2009.

Trant, J. (2008) Studying Social Tagging and Folksonomy: A review and framework*, Journal of Digital Information* 10(1), 1-44.

Trant, J. (2009) Tagging, Folksonomy and Art Museums: Review of steve.museum's research, Museums & the Web 2009 [online] Available: http://www.museumsandtheweb.com/jtrants/stevemuseum_research_report_available, [Accessed: 22/06/2012].

Upshall, M. (2011) Crowd-sourcing for education, Paper presented at *Online Information 2011, 29 November-1 December, Olympia, London*, Available http://www.online-information.co.uk, [Accessed: 22/06/2012]

Visser, J.(2011) *30 Do's for Designing Successful Participatory and Crowdsourcing Projects*, The Museum of the Future blog post 8th December 2011 [online], Available: http://themuseumofthefuture.com/2011/12/08/30-do%E2%80%99s-for-designing-successful-participatory-and-crowdsourcing-projects/, [Accessed: 20/06/2012].

Wiseman, R., Watt, C., Gilhooly, K. and Georgiou, G. (2011). Creativity and ease of ambiguous figural reversal. *British Journal of Psychology*, 102, 615–622.

Yang, Y., Bansal, N., Dakka, W., Ipeirotis, P., Koudas, N., & Papadias, D. (2009) Query by document, In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (WSDM '09), Ricardo Baeza-Yates, Paolo Boldi, Berthier Ribeiro-Neto, and B. Barla Cambazoglu (Eds.). ACM, New York, NY, USA, pp. 34-43.

7.1. Web sites

Amazon Mechanical Turk	https://www.mturk.com/mturk/welcome
Brooklyn Museum 'Click!'	http://www.brooklynmuseum.org/exhibitions/click/
Crowdflower	http://crowdflower.com/
Project Gutenberg	http://www.gutenberg.org/
Steve Museum	http://www.steve.museum/
Transcribe Bentham	http://www.ucl.ac.uk/transcribe-bentham/
Trove	http://trove.nla.gov.au
V&A	http://collections.vam.ac.uk/crowdsourcing/
WAISDA	http://woordentikkertje.manbijthond.nl/
Wiki Loves Monuments	http://www.wikilovesmonuments.eu/
Wikipedia	http://en.wikipedia.org/wiki/Main_Page
Your Paintings	http://www.bbc.co.uk/arts/yourpaintings/
What's The Score	http://www.bodleian.ox.ac.uk/bodley/library/special/projects/whats-the-score
When Can I Use…	http://caniuse.com/
YUMA	http://yuma-js.github.com/

8. Appendix 1

Appendix 1 includes the raw data from analysis of sentiment towards Europeana and is available on request.