# Closing the loop: assisting archival appraisal and information retrieval in one sweep

**Yunhyong Kim**
Humanities Advanced Technology and Information Institute (HATII) , University of Glasgow, Glasgow.UK.
yunhyong.kim@glasgow.ac.uk

**Seamus Ross**
Faculty of Information, University of Toronto, Toronto, Canada.
Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, UK.
seamus.ross@utoronto.ca

## ABSTRACT

In this article, we examine the similarities between the concept of *appraisal*, a process that takes place within the archives, and the concept of *relevance judgement*, a process fundamental to information retrieval systems. More specifically, we revisit appraisal/selection criteria proposed as a result of archival and digital curation communities, and, compare them to relevance criteria as discussed within information retrieval's literature based discovery. We illustrate how closely these criteria relate to each other and discuss how understanding the relationships between the these disciplines could form a basis for proposing automated appraisal and selection for archival processes and enabling complex queries within information retrieval.

## Keywords

appraisal, selection, relevance, digital preservation, digital curation, information retrieval, criteria, automation, IIR, archive, research data.

## 1. INTRODUCTION

The notion of *appraisal* plays a central role in archival practices. Simply put, appraisal is the process of evaluating whether a selected resource meets the criteria that warrants its inclusion or continued retention and/or maintenance within the archive (cf. Jenkinson 1922, Schellenberg 1956, and Cook 2005). The criteria for making such decisions have a long history of discussion, but, ultimately, the criteria reflect conclusions archivists reach about the contextual, evidential, informational, operational, societal and technical value of the resource (cf. Oliver et al. 2007).

While the objectives that guide the selection of material to be retained may affect the outcome of the process slightly differently across organisations, groups, and individuals, criteria, for assessing the value of resources to be included in a given collection, display noticeable similarities and are ubiquitous within the digital curation life cycle[1] (e.g. deciding what information to include in the creation of information objects; deciding which parts of a creation, if any, to include in the archive; and, deciding what information will be accessible at the time of dissemination).

The premise in this paper is that the *relevance judgement* process for assigning value to information within information retrieval, and *relevance criteria*, a set of central considerations for assigning value to the resource, can be mapped to the appraisal process and selection criteria used within the archival context. This is only natural as *relevance judgement* is merely appraisal at the use/re-use stage of the digital curation life cycle. An appreciation of the equivalence of these conceptual frameworks has significant theoretical and practical consequences. We will demonstrate, in fact, how we can leverage the equivalence to defray the cost of appraisal and selection in the archives through automation, and to improve the performance of information retrieval algorithms (Section 6).

We will bring together three *selection scenarios*: appraisal and the archive (Section 2), research data management (Section 3), and, relevance judgement and literature based discovery (Section 4). We argue that, despite the different end-objectives driving these scenarios, the fundamental basis for material selection shares commonality (Section 5): we will see that they all consider notions of scope (e.g. topic, genre, geographical region, and time frame); quality (e.g. accuracy, tangibility, quality of source, verifiability), uniqueness (e.g. document, content and source novelty), usability (e.g. clarity, reader background and ability, and technical infrastructure), and resource consumption (e.g. financial, labour).

We conclude by summarising the implications of the study presented here (Section 6) to suggest a roadmap towards a) a semi-automated three-tiered selection and appraisal process within the archives, and b) a reformulation of information retrieval as a multi-objective learning problem drawing on the notion of faceted relevance.

*ASIST 2013*, November 1-6, 2013, Montreal, Quebec, Canada.

[1] See the DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

## 2. SELECTION SCENARIO 1: APPRAISAL AND THE ARCHIVE

The central role that appraisal and selection plays in archiving digital material is discussed extensively within the literature (e.g. Harvey 2011). The need for appraisal and selection can be triggered by such criteria as legal requirements (data protection and limitations on retention), organisational capacity and objectives, and/or content quality[2]. While there is some controversy concerning whether appraisal and selection should take place at all[3] [4], it is generally agreed that, once it is decided that appraisal will take place, the criteria for carrying out this process in the archive should be clearly specified.

For example, Tibbo (2003) suggests criteria dependent on usefulness of the resource in relation to administration, fiscal transactions, meeting legal requirements, intrinsic value, evidential value, and informational value. The InterPARES[5] Appraisal Task Force presents an appraisal model[6] in their final report[7] (Eastwood 2004). They mention "needs of the records creator", "authenticity requirements", "societal needs", "legal requirements", and "archival science" as a source of criteria. However, their work is process driven (i.e. emphasis on the appraisal workflow), based on a model of digital information as a record (i.e. emphasis on "authenticity", "legal requirements", and "record context"), and does not fully explore what criteria drive these needs (e.g. societal needs; scientific value). One of the most comprehensive discussions of appraisal and selection criteria in the broader archival context can be found in Oliver et al. (2007) and the DELOS Preservation cluster[8] report on automated re-appraisal (DELOS 2008).

Figure 2.1 presents the criteria, based on content, technical, contextual, societal, evidential, and operational factors surrounding the material. In Table 2.1, we repeat some of the questions (identified by the DELOS report) that might be asked with respect to these factors. For instance, the first question about spatial region is associated with *Content Coverage*.

In this paper, we adopt the two-tier facets of Figure 2.1 as our reference point for the criteria that might arise within archival appraisal and selection, as these seem to subsume other general criteria such as those suggested by Tibbo and the InterPARES Appraisal Taskforce.

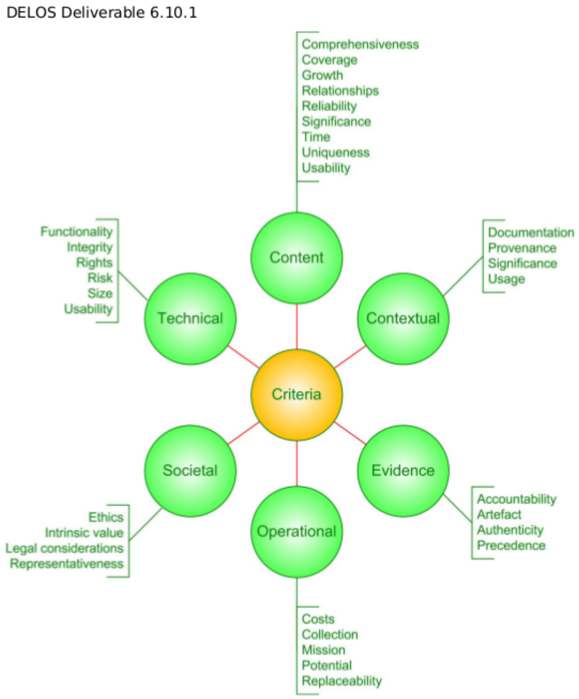

**Figure 2.1 From DELOS Report:"Automated re-Appraisal: Managing Archives in Digital Libraries"**

| **Category** | **Question** |
|---|---|
| **Content** | What spatial region does it cover? |
| **Content** | Does it have relationships to existing items? |
| **Content** | Are there similar objects already ingested into the collection? |
| **Content** | Is the language comprehensible? |
| **Content** | What is it about? |
| **Content** | What genre is it? |
| **Content** | What time frame does it cover? |
| **Contextual** | What business function/organisation was it created for? |
| **Contextual** | How often has it been accessed? |
| **Evidence** | Is it identifiable? |
| **Operational** | Are there special hardware requirements? |
| **Operational** | Are there special software requirements? |

[2] http://www.paradigm.ac.uk/workbook/appraisal/

[3] http://www.paradigm.ac.uk/workbook/appraisal/digital-appraisal.html

[4] http://www.digitalpreservationeurope.eu/publications/appraisal_final.pdf

[5] http://www.interpares.org

[6] http://www.interpares.org/display_file.cfm?doc=ip1_aptf_model.pdf

[7] http://www.interpares.org/display_file.cfm?doc=ip1_aptf_report.pdf

[8] http://www.dpc.delos.info/

| Category | Question |
|---|---|
| | |
| **Societal** | Is the creator significant? |
| **Societal** | Was it created at a significant time? |
| **Technical** | Can it be accessed? |
| **Technical** | Can it continue to be accessed? |
| **Technical** | What format is it in? |

**Table 2.1 Appraisal questions: sample from the DELOS report Deliverable 6.10.1 (2008)**

## 3. SELECTION SCENARIO 2: RESEARCH DATA MANAGEMENT

Another approach to appraisal and selection arises within the context of research data management, and, is represented by the list of criteria[9] recommended by the Digital Curation Centre (DCC)[10]. The list of criteria are re-introduced in Table 3.1.

| Selection Criteria | Description |
|---|---|
| **Relevance to Mission** | The resource content fits the organisation's remit and any priorities stated in the research institution or funding body's current strategy, including any legal requirement to retain the data beyond its immediate use. |
| **Scientific and Historical Value** | Is the data scientifically, socially, or culturally significant?Assessing this involves inferring anticipated future use, from evidence of current research and educational value. |
| **Uniqueness** | The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere. |
| **Potential for Redistribution** | The reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property 13 or human subjects issues are addressed. |
| **Non-Replicability** | It would not be feasible to replicate the data/resource or doing so would not be financially viable. |
| **Economic Case** | Costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate. |
| **Full Documentation** | The information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation and use. |

**Table 3.1 DCC selection criteria for research data (Whyte & Wilson 2010).**

In the context of the DCC guidelines, *Relevance to Mission* operates on the institutional level with respect to funder requirements, organisational priorities, and, legal requirements. However, in the current context, we would like to take this out of the institutional context: even on an individual level there are requirements, priorities, and legal conditions imposed (e.g. in relation to degree requirements; validity of approach; adequacy of evidence; permission to use third-party material). In considering the *Potential for Redistribution*, DCC is considering not only technical and content usability, but also adequate provision of provenance information, evidence of reliability, and legal permissions.

The Non-Replicability criterion is emphasised in such cases as observational datasets (e.g. astronomical data, meteorological data) and experimental data(sets) which are immensely costly to collect (e.g. the Large Hadron Collider). This type of uniqueness undoubtedly plays a role for curators of general digital information where the authentic information could be no longer reproducible. This, however, is expected to be less of a concern than uniqueness in the form of intellectual novelty in the case of information end-users and re-users.

In fact, both the case of the DELOS criteria (Figure 2.1) and the DCC criteria illustrates that current approaches to uniqueness, in information management circles, reflect mostly the nature of the content, although increasingly we are coming to recognise that more attention should be paid to uniqueness of manifestation and or instantiation than has been in the past. We will see that, information users and re-users (Section 4.2), place emphasis on uniqueness with respect to manifestation and also that of provenance, including source (for example, publisher and author). It should be mentioned that, in all the scenarios being discussed in this paper, uniqueness is not a property of the information itself but "uniqueness within context", a relationship between the information and other information or knowledge bases.

[9] http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

[10] http://www.dcc.ac.uk

## 4. SELECTION SCENARIO 3: RELEVANCE JUDGEMENT AND LITERATURE BASED DISCOVERY (LBD)

### 4.1 Common retrieval methodology

The concept of *relevance* is central to the objectives of information retrieval. Information retrieval is typically carried out on a search system, where, in response to a query issued by a user, the system retrieves a set of objects that is deemed relevant to the query. Traditionally the aforementioned *relevance* was expressed as topical nearness between the issued query (e.g. representing *information need* of the user) and the content of the retrieved objects. The Text Retrieval Conference (TREC)[11] has traditionally served as a way of producing bench mark datasets for the evaluation of retrieval systems based on expert judgements. In the basic TREC framework: the competing systems submit the top 1000 objects returned by their systems, these are given back to experts within the information retrieval community, who, through a *relevance judgement process*, label them as relevant or not relevant. The results of the labelling process are then used along with standard performance measures (mostly based on variations of precision and recall, information gain, and/or utility) to evaluate the submitted systems. The datasets resulting from TREC are often re-used in major retrieval conferences (e.g. SIGIR conference[12]).

Initially, the relevance judgements were represented as binary decisions (either relevant or not relevant), but, in recent years, this has been extended to include preferential ratings. Further value has been attributed to diversified search results (e.g. Dou et al. 2011 on coverage of sub-topics; Chander & Carterette 2012 on retrieval of novel documents). However, the lack of insight into reasons for the relevance judgements that these approaches afforded has been questioned because they assume that the information need of a user is static, and, that relevance judgement on one object is not influenced by other objects in the collection. They disregarded a range of contextual and cognitive factors. These limitation are extensively discussed in Borlund (2003). For example, recent study (e.g. Scholer, Turpin & Sanderson 2011) shows that the consistency across relevance judgements vary with respect to topic, and with respect to similar documents for which relevance judgement had already been made.

### 4.2 Relevance criteria

The complexity observed with respect to interactive information retrieval has led to studies that try to identify *reasons* underlying the relevance judgements made by users as part of information search (e.g. see Barry 1994; Barry & Schamber 1998). The criteria have been revisited and adapted to meet different situations. Nevertheless, the lists bear sufficient similarity to each other to warrant further study as a set of factors that influence relevance judgements. Here we focus on one manifestation of the criteria within the context of literature based discovery (LBD)[13] as presented by Cerviño Beresi et al. (2010) – reproduced in Table 4.1.

| **Relevance Criteria** | **Description** |
|---|---|
| **Depth/scope/specificity** | Whether the information is in-depth or focused, has enough detail or is specific to the user's needs. |
| **Currency** | Whether the information is current or is up-to-date. |
| **Accuracy/validity** | Whether the information found is accurate or valid. |
| **Tangibility** | Whether the information relates to tangible issues, and/or hard data/facts are included. |
| **Quality of Source** | Quality of the sources of information: includes authors as well as publishers. |
| **Verification** | Whether other information in the field agrees with the presented information. |
| **Affectiveness** | Whether the user shows an affective or emotional response to the information. |
| **Document Novelty** | The extent to which the document itself is novel to the user. |
| **Source Novelty** | The extent to which a source of the document (i.e., author, journal) is novel to the user. |
| **Content Novelty** | The extent to which the information presented is novel to the user. |
| **Clarity** | Whether the information is presented in a clear fashion. |
| **Ability** | Whether the user judges that he/she will be able to understand information presented. |
| **Background Experience** | Whether the knowledge with which the user approaches information will be sufficient. |
| **Accessibility** | Whether there is some cost or technical difficulty involved in obtaining the information. |
| **Availability** | Whether the information is |

[11] http://trec.nist.gov/

[12] http://sigir2013.ie/

[13] The use of literature, for example, academic publications, to discover new relationships between existing knowledge.

| Relevance Criteria | Description |
|---|---|
|  | available at that point in time. |

**Table 4.1 Relevance criteria for LBD: as presented in Cerviño Beresi et al. 2010.**

The list in Table 4.1 confirms the discussion in Section 3. In the literature based discovery domain, Non-Replicability is not an explicit concern. And, the notion of *Uniqueness* arises in three dimensions; that of the content, the document (i.e. manifestation or instantiation), and source (information producer – say, for example, author, illustrator, presenter, and/or publisher). Unlike the archival appraisal criteria and DCC criteria, *Full Documentation* is not mentioned as an explicit criterion for LBD. It is worthy of note that many of the decisions regarding the criteria in Table 4.1 would only be possible with the help of some form of documentation or evidence to support inference.

In Section 5, we will see that the criteria used within the three scenarios (scenarios of Section 2, 3 and 4) can be mapped to each other.

## 5. MAPPING CRITERIA ACROSS SCENARIOS

In Figure 5.1, we have mapped the criteria that arise within the three selection scenarios, discussed in previous sections, to each other.

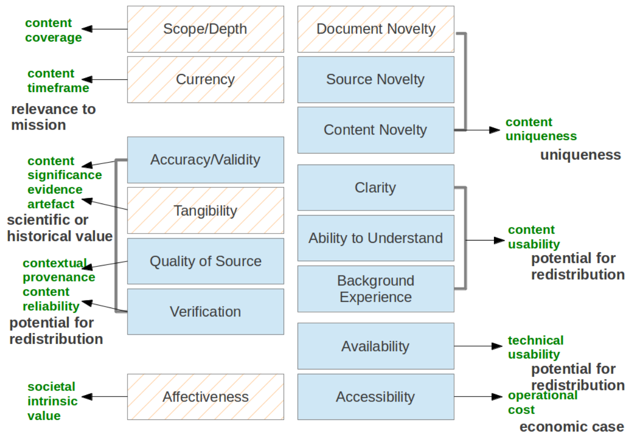

**Figure 5.1 Mapping relevance criteria (displayed in the boxes) to DCC criteria (annotated in smaller boldface black fonts) and DELOS (smallest boldface green fonts) appraisal criteria. The lighter shaded boxes are the five top criteria observed in LBD.**

The criteria in the boxes represent relevance criteria used in literature based discovery. These have been annotated with DCC selection and appraisal criteria (indicated in slightly smaller black fonts) and DELOS selection and appraisal criteria (in even smaller green fonts). For example, the relevance criterion Scope/Depth (labelled *Depth/scope/specificity* in Table 4.1) is mapped to the DCC criterion *Relevance to Mission*, which, in turn, is mapped to the DELOS criterion *Content Coverage*. The mapping is not meant to be definitive. It is meant to illustrate the common themes arising in these scenarios.

Some sets of relevance criteria serve as evidence for scientific value as well as potential for redistribution. For example, while the *Quality of Source* is evidence for scientific value, in order for the evaluator to be able to assess the quality of source, it is a prerequisite condition that the source be identifiable. Likewise if an item has been selected on the basis of quality of source then the likelihood is that source information has been made available for that item. This is, therefore, directly related to meeting the DELOS criteria *Contextual Provenance* and *Content Reliability*, and, in the same vein, the DCC criterion for *Potential for Redistribution*, as well as, criterion for *Full Documentatio*n. The criterion *Full Documentation* was not expressed in Figure 5.1 because the relevance criterion relates to the quality of the source, not the identification of the source itself.

In the case of the archival scenario, the key lies in understood controlled processes of evidence gathering that establishes accountability, authenticity, integrity and reliability. In the research data scenario, trustworthiness is driven by peer review, reputation (author and publisher), citation, impact, supporting evidence, and expert opinion. Both cases come from a management perspective.

On the other hand the third scenario involving literature based discovery (LBD) comes from the opposite direction to investigate information use. The objective for LBD is to develop an approach to *operationalise* potentially helpful criteria from Table 4.1. so as to assist users of the information to navigate, browse, explore and synthesise the available information efficiently and effectively, and/or to use the criteria as additional information to improve retrieval processes (e.g. search engine processes) .

All three scenarios share an interest in assessing scope, quality, uniqueness, usability, and costs related to information. In Section 6, we will discuss research that already exists within information retrieval and related disciplines, to determine scope, measure quality, detect uniqueness, assess usability, and defray costs. We will also show how this can be combined to propose an approach to automating or semi-automating the appraisal process for information management. At the same time, we will see how information retrieval approaches (e.g. link analysis, content analysis, and usage analysis) cut across tasks relevant to assessing different criteria. The latter observation suggests that a multi-objective multi-task approach (a machine learning approach - for example, like that proposed in Bagherjeiran 2007) optimised across weighted criteria might serve end-users better in information use as well, leading to a parallel way forward for information management and retrieval that could be of mutual benefit to all parties involved.

## 6. CONCLUSION AND TAKING IT FORWARD

A monumental amount of research has gone into information retrieval (e.g. see Manning, Raghaven & Schütze 2010). While most of us recognise it as what the search engine does for us everyday when we submit a set of keywords into the web browser search box, in recent years the field has blossomed to address a multitude of problems such as those associated with faceted queries, complex queries, cross-lingual queries, natural language queries, document novelty detection, and legal information extraction. This makes the field now ripe for development in considering its role in assisting appraisal.

An early breakthrough in information retrieval was inspired by studies in term weighting introduced by Spӓrck-Jones (1972). Later selected concepts were adopted from text categorisation (e.g. Sebastiani 2002), Latent Semantic Analysis (e.g. Deerwester et al. 1990), Language Models (e.g. Ponte & Croft), Latent Dirichlet Allocation (e.g. see Blei, Ng & Jordan), Support Vector Machines (Vapnik 1995), link analysis (cf. Bianchini, Gori, Scarselli 2005), and collaborative filtering (e.g. see Cacheda et al. 2011). In the next sections we will see how these methods have evolved into applications assessing criteria associated with scope, quality, trustworthiness, uniqueness, usability and resource consumption (Section 6.1), and how these approaches developed in information retrieval can be leveraged to help the appraisal and relevance judgement process (Section 6.2).

### 6.1 Making decisions: scope, quality, uniqueness, usability, costs

*Scope: relevance to mission*

Across the three selection scenarios of this paper, scope is most frequently related to geographical region, temporal region, topic and/or subject, and genre. While some may include language in this section, here we reserve language as a factor in *Ability* and highlight geographical region as the more prominent concern: that is, there are conditions on the language if the information is to be understood, the geographical region, however, might be a matter of jurisdiction. As an English speaking researcher, one might have no objection to exploring a paper written in French (which happens to be written by a colleague at the same institute), but one might not be able to make sense of it if one does not know French.

There are papers on automated retrieval of information with respect to each of these dimensions. For example, effective methods have been developed for temporal information retrieval that can detect time associated with queries even when time requirements are not specified within the query (e.g. Mathews & Kanmani 2012); retrieval approaches that identify associated geographic information (e.g. Goldberg, Wilson & Knoblock 2009; Jones & Purves 2008; Jones, Alani & Tudhope 2001); approaches to music genre classification (e.g. Tzanetakis 2002; Scaringella & Zoia 2005); approaches in text genre classification (e.g. Karlgren & Cutting 1994; Stamatatos, Fakotakis & Kokkinakis 2000; Petrenz and Webber 2011); identification of audio-visual genres (e.g. Glasberg 2008; Ianeva 2003); and, genre model to identify anomalous patterns in data (e.g. Xiong, Poczos & Schneider).

*Quality and Trustworthiness: scientific or historical value*

Measure of information quality is strongly bound to a measure of trustworthiness we perceive in relation to the authenticity, integrity and reliability of the material. To gauge this we seek means of evaluating and validating statements (for instance, by providing measures of performance, case studies, experiments, and, examples), confirming quality (for instance, by including citations, expert reviews and opinions, and, peer assessment), presenting evidence that support trustworthiness (by carrying out statistical analysis, by using accepted standards, by employing concrete formulas), and accepting information from recognised sources (by checking for names of high impact journals).

Automated measurements of reliability is especially of interest within social media networks (e.g. see Caverlee, Liu & Webb 2010; Kim & Ahmad 2013) where basically anybody can publish anything. As a consequence, it is also the domain within which much progress has been made to try to understand the notion of *trust* and *quality:* Agichtein et al. (2008) provides an excellent summary of approaches to finding high quality content within social media.. These studies revolve around automatic detection of reliability based on statistical analyses of a wealth of features including:

- Link analysis (e.g. relationships between users and content objects);
- Reputation propagation analysis (e.g. transitivity of trust);
- Expert search (e.g. developing reputation measures[14]);
- Content topic analysis (e.g. n-grams)
- Content quality analysis (e.g. punctuation and typos; syntactic and semantic complexity; grammaticality);
- Content usage statistics (e.g. clicks and dwell time); and,
- Implicit feedback analysis (e.g. position of clicks).

This kind of research is only possible through the availability of rich data that surround social media content

[14] For example, see http://www.limosine-project.eu/events/replab2013

(e.g. hyperlinks, comments, questions & answers, voting, rating, history of users). This proposes social media as a ripe field for studying the question of trust. The research area is especially compelling given recent studies (e.g. Su et al. 2007) on Question Answering systems[15] which show that, while the percentage of good answers to a given question is, approximately, only 45% of the total set of answers, the proportion of questions with at least one good answer is approximately 95% of all questions. Also, the research into trust within the social media environment is further encouraged by its potential use in emergency response initiatives (cf. Hagar 2013).

Another dimension of quality relates to authenticity, more specifically, on authorship attribution and plagiarism detection (e.g. Stamatatos 2011; Savoy 2012), a subject area related to genre classification (see Section on "Scope"). The concept of authenticity emerges also within the study of *spam*, misleading information posing to be something it is not (e.g. Schneider 2003; Castillo & Davison 2011).

*Uniqueness*

Some research in document novelty was already mentioned, in Section 4.2, as an area of research trying to improve on ad hoc retrieval that does not put into consideration the context of search. This area of research (e.g. Dou et al 2011; Chander & Caterette 2012; Zhang, Callan & Minka 2002) has also expanded into research into diversification of search results to widen coverage (e.g. Raman, Shivaswamy & Joachims 2012), and into link analysis to construct a hierarchy in web page similarity (e.g. Schiffman 2006). On the content level, there have been research in detecting near duplicates (e.g. Manku, Jain & Sarma 2007).

*Usability: Potential for Redistribution*

On the content level, usability can be defined as being related to readability, being in an accessible language, correct punctuation and reduced number of typos, and clear organisation. Measurements for text readability is a well studied subject[16]. Understandability does not follow from readability, but it is a start in defining the intended audience of the information. Automatic language identification of text documents was addressed early on as an information processing task (Beesley 1988; Singh 2006). Further approaches to analysing web page language is also available (Selamat 2011). How to assess quality on the level of punctuation and misspelling and approaches to developing automated grading systems are discussed in Agichtein et al. (2008). A study of argumentation structure (Teufel, Siddharthan & Batchelor 2009) can also help to determine which texts are likely to be incoherent.

On the technical level, there are approaches developed to test system usability[17]. It is conceivable that usability testing be carried out for data re-usability. This may initially sound costly but if carried out in conjunction with an understanding of implicit user feedback (Joachims et al. 2007; Shivaswamy & Joachims 2012), system design that allows explicit user feedback (e.g. mechanisms to allow, comments, ratings, votes, suggestions) and the initiatives outlined below for defraying costs, it is not that far fetched.

*Cost: Economic Case*

Recently there have been many initiatives to defray costs by pulling resources together. This includes approaches such as Cloud computing and/or other distributed computing and storage (e.g DuraCloud[18]), and community fundraising initiatives such as Kickstarter to raise capital to carry out a project[19]. It also includes crowd sourcing for collecting metadata (e.g. using games to label and describe objects[20]), discovering new knowledge (e.g. exploration of protein folding with FoldIt[21]; crowd sourcing to explore the ocean floor or galaxies with Zooniverse[22]), and improving information retrieval[23]. The same kind of initiative could be deployed to harvest social benefit from *citizen appraisal.*

### 6.2 Taking it forward

In this paper, we have discussed how appraisal and selection criteria for archives and research data are similar to relevance criteria identified within information retrieval and literature based discovery. In Section 6.1, we have shown that there is a plethora of research supporting tasks relevant to operationalising the criteria common across the three selection scenarios (appraisal in the archives, research data management, and, literature based discovery in information retrieval). We propose that approaches to these tasks be absorbed by archival and curation communities to alleviate the intense labour involved in appraisal and selection.

The review of information processing presented here shows that the tasks scattered across many information retrieval extraction and processing domains operate on similar types of information analysis (link analysis; content analysis, user

[15] For example http://uk.answers.yahoo.com/ and http://stackoverflow.com/

[16] For example, see http://www.standards-schmandards.com/2005/measuring-text-readability/

[17] http://en.wikipedia.org/wiki/Usability_testing

[18] http://www.duracloud.org/

[19] http://www.kickstarter.com/

[20] http://chronicle.com/blogs/wiredcampus/gaming-the-archives/31435

[21] http://fold.it/portal/

[22] https://www.zooniverse.org/

[23] For example, see http://link.springer.com/journal/10791/16/2/page/1

interaction and usage analysis). The approaches that are currently developed in retrieval scenarios such as those described in Section 6.1 tend to be based on single objective and/or task models (i.e. get the high-quality content given the features). We propose that approaches to information retrieval could benefit from a multi-objective multi-task model (e.g. see Bagherjeiran 2007) cognizant of the closely bound criteria developed in appraisal scenarios and exploiting feedback from human appraisal processes.

Most of the current research areas discussed in Section 6.1 depend on availability of user generated data (e.g. links, voting, rating, comments) and statistical comparison across big datasets. In this light, we recommend that:

> *Archives and repositories adapt information management systems to support infrastructure that, allow gathering of user generated data and foster developing repository collections on a large scale, while allowing interaction between automated information selection processes and human information processes.*

A large amount of this can be achieved by archives adopting a range of social media network tools. In fact, reputable preprint servers[24] are already using the trackback[25] functionality, originally implemented within blogs to allow authors of posts to cite other posts. The advantage of the trackback function is that the author of the cited article can immediately see that their article is being cited. In Figure 6.1, we illustrate what an appraisal engine might look like. The social media infrastructure (represented by the component on the top right hand side of Figure 6.1) would enable a three tier appraisal and selection process:

- Tier 1: everyday crowd-sourcing of data (Tier 1);
- Tier 2: followed by an automated preliminary appraisal and selection process (based on a multi-objective multi-task framework drawing on selected work such as those presented in Section 6.1);
- Tier 3: subsequently managed by a second phase of quality assurance by humans. The result of this phase of assessment can be fed back to improve the automated phase in Tier 2.

By re-vamping the archival system to incorporate the benefits of the social media platform and coupling it with a multi-objective retrieval process for selecting high-quality content that learns from archival practices, we are sure to harness the best of both worlds.

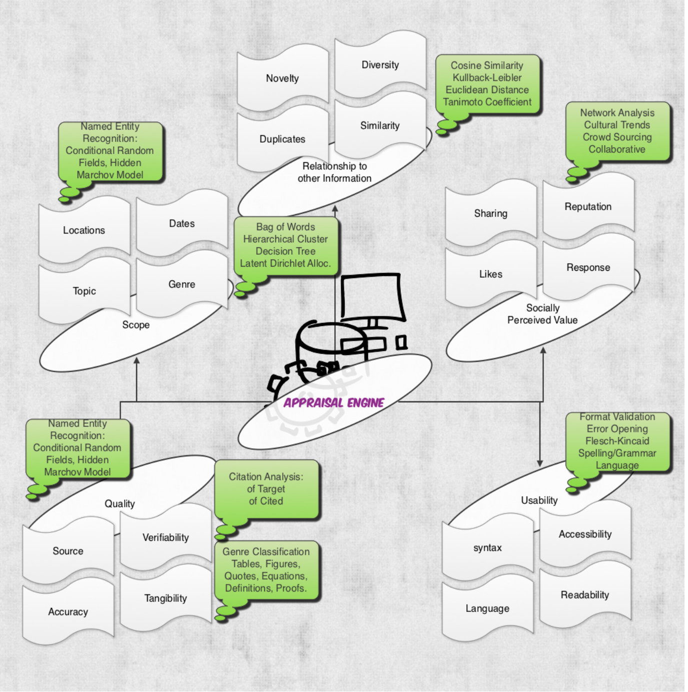
Novelty
Diversity
Duplicates
Similarity
Relationship to other Information
Cosine Similarity Kullback-Leibler Euclidean Distance Tanimoto Coefficient
Network Analysis Cultural Trends Crowd Sourcing Collaborative
Named Entity Recognition: Conditional Random Fields, Hidden Marchov Model
Locations
Dates
Topic
Genre
Scope
Bag of Words Hierarchical Cluster Decision Tree Latent Dirichlet Alloc.
Sharing
Reputation
Likes
Response
Socially Perceived Value
APPRAISAL ENGINE
Named Entity Recognition: Conditional Random Fields, Hidden Marchov Model
Quality
Citation Analysis: of Target of Cited
Source
Verifiability
Accuracy
Tangibility
Genre Classification Tables, Figures, Quotes, Equations, Definitions, Proofs.
Format Validation Error Opening Flesch-Kincaid Spelling/Grammar Language
Usability
syntax
Accessibility
Language
Readability

**Figure 6.1 Appraisal engine incorporating automated process to capture five aspects of information (scope, quality, relationship to other information, socially perceived value, usability). Green bubbles highlight related automated information processing method.**

## 7. ACKNOWLEDGMENTS

The work described in this paper was carried out as part of the BlogForever project co-funded by the European Commission (FP7-ICT-2009-6, Grant agreement no.: 269963) and builds on work of the authors as part of the DELOS Network of Excellence, an initiative co-funded by the European Commission, (FP6-IST-2004, Grant agreement no.: G038-507618)

## 8. REFERENCES

Agichtein, E., Castillo, C., Donato, D., Gionis, A. & Mishne, G. (2008) Finding high-quality content in social media. *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*, Palo Alto, California, USA, 183-194. DOI 10.1145/1341531.1341557, New York: ACM, New York, NY, USA.

Bagherjeiran, A. (2007) *Multi-Objective Multi-Task Learning*. Ph.D. Dissertation. University of Houston, Houston, TX, USA. AAI3263404.

Barry, C. L. (1994) User-defined relevance criteria: an exploratory study. *J. Am. Soc. Inf. Sci. 45*, 3, 149-159. New York: John Wiley & Sons, Inc.

Barry, C. L. & Schamber, L. (1998) Users' criteria for relevance evaluation: a cross-situational comparison. *Inf. Process. Manage. 34*, 2-3, 219-236. DOI 10.1016/S0306-4573(97)00078-2. Terrytown:Pergamon Press, Inc., NY, USA.

Beesley, K. (1988) Language identifier: A computer program for

[24] http://uk.arxiv.org/help/trackback

[25] http://en.wikipedia.org/wiki/Trackback

automatic natural-language identification on on-line text.

Bianchini, M., Gori, M. & Scarselli, F. (2005) Inside PageRank. *ACM Trans. Internet Technol.* 5, 1, 92-128. http://doi.acm.org/10.1145/1052934.1052938

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) Latent dirichlet allocation. *J. Mach. Learn. Res. 3 (March 2003)*, 993-1022.

Borlund, P. (2003) The concept of relevance in IR. *J. Am. Soc. Inform. Sci. Technol. 54*, 10, 913-925.

Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2 (2): 121-167.

Cacheda, F., Carneiro, V., Fernández, D. & Formoso, V. (2011) Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans. Web 5, 1*, Article 2 (February 2011), http://doi.acm.org/10.1145/1921591.1921593

Castillo, C. & Davison, B. D. (2011) Adversarial Web Search. *Found. Trends Inf. Retr. 4, 5 (May 2011)*, 377-486. http://dx.doi.org/10.1561/1500000021

Caverlee, J., Liu, L. & Webb, S. (2010) The SocialTrust framework for trusted social information management: Architecture and algorithms. *Inf. Sci. 180*, 95-112. DOI 10.1016/j.ins.2009.06.027. New York: Elsevier Science Inc. NY, USA.

Cerviño Beresi, U., Kim, Y., Song, D., Ruthven, I. (2010) Why did you pick that? Visualising relevance criteria in exploratory search. *Int. J. Digit. Libr. 11*, 2, 59-74. DOI 10.1007/s00799-011-0067-7. Berlin: Springer-Verlag.

Chandar, P. and Carterette, B. (2012) Using preference judgments for novel document retrieval. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, Portland, Oregon, USA, 861—870. DOI 10.1145/2348283.2348398.

Cook, T. (2005) Macroappraisal in Theory and Practice: Origins, Characteristics, and Implementation in Canada, 1950–2000, *Archival Science* 5:2-4, 101-61.

Deerwester, S. Dumais, S. T., Furnas, G. W. Landauer, T. K. & Harshman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science (1986-1998) 41, 6*, 391.

DELOS (2008) Deliverable 6.10.1: Report on Automated re-Appraisal: Managing Archives in Digital Libraries" Preservation Cluster (6) of DELOS Network of Excellence.

Dou, Z., Hu, S., Chen, K., Song, R. & Wen, J-R. (2011) Multi-dimensional search result diversification. P*roceedings of the fourth ACM international conference on Web search and data mining (WSDM '11)*, 978-1-4503-0493-1, 475-484. DOI 10.1145/1935826.1935897.

Eastwood, T. (2004) Appraising Digital Records for Long-term Preservation. *Data Science Journal 3 (2004),* 202-208.

Glasberg, R., Schmiedeke, S., Kelm, P. & Sikora, T. (2008) An automatic system for real-time video-genres detection using high-level-descriptors and a set of classifiers. *IEEE International Symposium on Consumer Electronics, (ISCE 2008),* 1-4. DOI 10.1109/ISCE.2008.4559449

Goldberg, D. W., Wilson, J. P. & Knoblock, C. A. (2009) Extracting geographic features from the internet to automatically build detailed regional gazetteers. *International Journal of Geographical Information Science 23 (1)*, 93-128.

Hagar, C. (2003) Crisis informatics: Perspectives of trust – is social media a mixed blessing? *Student Research Journal 2(2)*, http://scholarworks.sjsu.edu/slissrj/vol2/iss2/2.

Harvey, R. (2011) *Preservation of Digital Materials (2nd Ed.)}*, Berlin: De Gruyter, Berlin, Germany.

Ianeva, T. (2003) Detecting cartoons: a case study in automatic video-genre classification. *University of Valencia technical report*. http://www.uv.es/~tzveta/invwork.pdf

Jenkinson, H. (1922) *A Manual of Archive Administration.* London: Percy Lund, Humphries and Co., London, UK.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. Radlinski, F. & Gay, G. (2007) Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Trans. Inf. Syst. 25, 2.* http://doi.acm.org/10.1145/1229179.1229181

Jones, C. B., Alani, H. & Tudhope, D. (2001) Geographical information retrieval with ontologies of place. *Spatial Information Theory*, 322-335.

Jones, C. B. and Purves, R. S. (2008) Geographical information retrieval. *International Journal of Geographical Information Science 22 (3)*, 219-228. http://www.tandfonline.com/doi/abs/10.1080/13658810701626343

Karlgren, J. & Cutting, D. R. (1994) Recognizing Text Genres With Simple Metrics Using Discriminant Analysis. COLING 1994:1071-1075

Kim, Y. A. & Ahmad, M. A. (2013) Trust, distrust and lack of confidence of users in online social media-sharing communities. *Know.-Based Syst. 37*, 438—450. DOI 10.1016/j.knosys.2012.09.002. Amsterdam: Elsevier Science Publishers B. V., Amsterdam, The Netherlands.

Manku, G. S., Jain, A., and Sarma, A. D., Detecting Near-Duplicaes for Web Crawling. *Proceeding of the International World Wide Web Confernce* (*WWW2007).* http://www.wwwconference.org/www2007/papers/paper215.pdf

Manning, C. D., Raghavan, P. & Schütze, H. (2010) Introduction to information retrieval. *Inf. Retr.* 13, 2, 192-195. http://dx.doi.org/10.1007/s10791-009-9115-y

Mathews, L. K. & Kanmani, D. S. (2012) A Survey on Temporal Information Retrieval Systems. *International Journal of Computer Applications 58(4)*, 24-28. New York: Foundation of Computer Science, NY, USA.

Oliver, G., Ross, S., Guercio, M. & Pala, C. (2007) Report on automated re-appraisal: managing archives in digital libraries, *Archivi e Computer 17(2007)2-3*, 199-253.

Petrenz, P. & Webber, B. (2011) Stable classification of text genres. *Comput. Linguist. 37, 2 (June 2011)*, 385-393. http://dx.doi.org/10.1162/COLI_a_00052

Ponte, J. M. & Croft, W. B. (1998) A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275-281. ACM Press.

Sandholm T. & Ung, H. (2011) Real-time, location-aware collaborative filtering of web content. *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation* (CaRR '11). ACM, New York, NY, USA, 14-18. http://doi.acm.org/10.1145/1961634.1961638

Savoy, J. (2012) Authorship Attribution Based on Specific Vocabulary. *ACM Trans. Inf. Syst. 30 (2).* http://doi.acm.org/10.1145/2180868.2180874

Scaringella, N. & Zoia, G. (2005) On the Modeling of Time Information for Automatic Genre Recognition Systems in Audio Signals. *Procceedings of ISMIR 2005,* 666-671.

Schellenberg, T. R. (1956) *Modern Archives: Principles and Techniques.* Chicago: University of Chicago Press, Chicago, IL, USA.

Schiffman, A. (2006) Hierarchy in Web Page Similarity Link Analysis. Technical Report. http://commerce.net/wp-content/uploads/2012/04/CN-TR-06-02.pdf

Schneider, K-M. (2003) A comparison of event models for Naive Bayes anti-spam e-mail filtering. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1* (EACL '03), Stroudsburg: Association for Computational Linguistics. http://dx.doi.org/10.3115/1067807.1067848

Scholer, F., Turpin, A. & Sanderson, M. (2011) Quantifying test collection quality based on the consistency of relevance judgements. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*, Beijing, China, 1063-1072. DOI 10.1145/2009916.2010057.

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Comput. Surv 34*, 1, 1-47. http://doi.acm.org/10.1145/505282.505283

Selamat. A. (2011) Improved N-grams approach for web page language identification. *Transactions on computational collective intelligence V*, Ngoc Thanh Nguyen (Ed.). Springer-Verlag, Berlin, Heidelberg 1-26.

Shivaswamy, P. & Joachims, T. Online Structured Prediction via Coactive Learning. *International Conference on Machine Learning (ICML)*, 2012.

Singh, A. K. (2006) Study of some distance measures for language and encoding identification. *Proceedings of the Workshop on Linguistic Distances* (LD '06), 63-72.

Spärck-Jones, K. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1): 11-21.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, *26(4)*, 461-485, MIT Press.

Stamatatos, E. (2011) Plagiarism detection based on structural information. *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*, 1221-1230.

Su, Q., Pavlov, D., Chow, J-H. & Baker, W. C. (2007) Internet-scale collection of human-reviewed data. *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, Banff, Alberta, Canada, 231-240. DOI 10.1145/1242572.1242604. New York: ACM, New York, NY, USA.

Teufel, S., Siddharthan, A. & Batchelor, C. (2009) Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09), *Vol. 3,* 1493-1502.

Tibbo, H. (2003) On the nature and importance of archiving in the digital age. *Advances in Computers 57,* 1-67.

Tzanetakis, G. & Cook, P. (2002) Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5.*

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995

Wei, X. & Croft, W. B. (2006) LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '06), 178-185. http://doi.acm.org/10.1145/1148170.1148204

Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides - See more at: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data#sthash.poBHEZqi.dpuf

Xiong, L., Poczos, B., Schneider, J. (2011) Group Anomaly Detection using Flexible Genre Models, *Neural Information Processing Systems (NIPS).* http://www.cs.cmu.edu/~schneide/fgm.pdf

Zhang, Y., Callan, J. & Minka, T. (2002) Novelty and redundancy detection in adaptive filtering. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02),* 81-88. http://doi.acm.org/10.1145/564376.564393

**The columns on the last page should be of approximately equal length.**