An Overview of Semantic Search Evaluation Initiatives

Khadija M. Elbedweihy[a,*], Stuart N. Wrigley[a], Paul Clough[b], Fabio Ciravegna[a]

[a]*Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK*
[b]*Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK*

Abstract

Recent work on searching the Semantic Web has yielded a wide range of approaches with respect to the underlying search mechanisms, results management and presentation, and style of input. Each approach impacts upon the quality of the information retrieved and the user's experience of the search process. However, despite the wealth of experience accumulated from evaluating Information Retrieval (IR) systems, the evaluation of Semantic Web search systems has largely been developed in isolation from mainstream IR evaluation with a far less unified approach to the design of evaluation activities. This has led to slow progress and low interest when compared to other established evaluation series, such as TREC for IR or OAEI for Ontology Matching. In this paper, we review existing approaches to IR evaluation and analyse evaluation activities for Semantic Web search systems. Through a discussion of these, we identify their weaknesses and highlight the future need for a more comprehensive evaluation framework that addresses current limitations.

Keywords: semantic search, usability, evaluation, benchmarking, performance, information retrieval

1. Introduction

The movement from the 'web of documents' towards structured and linked data has made significant progress in recent years. This can be witnessed by the continued increase in the amount of structured data available on the Web, as well as the work carried out by the W3C Semantic Web Education and Outreach (SWEO) Interest Group's community project *Linking Open Data*[1] to link various open datasets. This has provided tremendous opportunities for changing the way search is performed and there have been numerous efforts to exploit these opportunities in finding answers to a vast range of users' queries. These efforts include *Semantic Web search engines*, such as Swoogle (Ding et al., 2004) and Sindice (Tummarello et al., 2007), which act as gateways to locate Semantic Web documents and ontologies in a similar fashion to how Google and Yahoo! are used for conventional Web search. Whilst these systems are intended for Semantic Web experts and applications, another breed of tools has been developed to provide more accessible approaches to querying structured data. This includes *natural language interfaces* operating in a single domain, such as NLP-Reduce (Kaufmann et al., 2007) and Querix (Kaufmann et al., 2006), or in multiple and heterogeneous domains, such as PowerAqua (López et al., 2006) and Freya (Damljanovic et al., 2010); *view-based interfaces* allowing users to explore the search space whilst formulating their queries, such as K-Search (Bhagdev et al., 2008) and Smeagol (Clemmer and Davies, 2011); and *mashups*, integrating data from different sources to provide rich descriptions about Semantic Web objects, such as Sig.ma (Tummarello et al., 2010) and VisiNav (Harth, 2010).

Similar to designing and developing Information Retrieval (or search) systems more generally, evaluation is highly important as it enables the success of an search system to be quantified and measured (Järvelin, 2011). This can involve evaluating characteristics of the IR system itself, such as its retrieval effectiveness, or assessing consumers' acceptance or satisfaction with the system (Taube, 1956). For decades, the primary approach to IR evaluation has been system-oriented (or batch-mode), focusing on assessing how well a system can find documents of interest given a specification of the user's information need. One of the most used methodologies for conducting IR experimentation that can be repeated and conducted in a controlled lab-based setting is test collection-based evaluation (Robertson, 2008; Sanderson, 2010; Järvelin, 2011; Harman, 2011). Commonly known as the Cranfield methodology, this approach has its origins in experiments conducted at Cranfield library in the UK (Cleverdon, 1991). Although proposed in the 1960s, this approach was popularised through the NIST-funded Text REtrieval Conference (TREC) series of large-scale evaluation campaigns, which began in 1992 and has stimulated significant developments in IR over the past 20 years or so (Voorhees and Harman, 2005).

However, despite the many benefits that come from the organisation of evaluation activities like TREC, the semantic search community still lacks a similar initiative on this scale. Indeed, Halpin et al. (2010) note that *"the lack of standardised*

*Corresponding author
Email addresses: kelbedweihy@cu.edu.eg (Khadija M. Elbedweihy), s.wrigley@sheffield.ac.uk (Stuart N. Wrigley), p.d.clough@sheffield.ac.uk (Paul Clough), f.ciravegna@sheffield.ac.uk (Fabio Ciravegna)

[1] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

evaluation has become a serious bottleneck to further progress in this field". In recent years evaluation activities have been organised to address this issue, including the SemSearch Challenge (Halpin et al., 2010); the SEALS semantic search evaluations (Wrigley et al., 2010c, 2011); the QALD open challenge (Unger et al., 2011) and the TREC Entity List Completion task (Balog et al., 2010c, 2011). However, these initiatives are yet to experience the level of participation shown by evaluation exercises in other fields. Furthermore, much of the experience gained from these initiatives is not accessible to the research community in general since they usually focus on reporting objectives and results with little explanation on the specific details of the evaluations, such as the methods and measures adopted. It is important to emphasise the need for more work towards evaluation frameworks/platforms that enable persistent storage of datasets and results that guarantee their reusability and the repeatability of tests. Forming agreement on the approaches and measures to be used within the search community for evaluating similar categories of tools and approaches is a key aspect of developing standardised evaluation frameworks (Zobel et al., 2011). In addition to the resources created, the value of organised evaluation campaigns in bringing together members of the research community to tackle problems collectively is also a stimulus for growth in a research field.

Although the focus of this paper is the evaluation of semantic search, we believe that there is much to learn from the wider IR community more generally. Therefore, this paper summarises IR evaluation activities and considers how this knowledge can be utilised in meeting the specific requirements of semantic search. The overall goal of this paper is to motivate the future development of a more formalised and comprehensive evaluation framework for semantic search. The remainder of this paper is structured as follows. An overview of evaluation in IR is provided in Section 2, followed by a discussion of important aspects of system-oriented evaluation, such as test collections and measures, in Section 3. Next, Section 4 describes approaches for user-oriented evaluation, such as the experimental setup and criteria to be assessed. Section 5 then goes on to summarise existing semantic search evaluation initiatives with potential limitations in existing approaches to evaluating semantic search and future directions discussed in Sections 6 and 7.

2. Approaches to IR Evaluation

Evaluation is the process of assessing the 'worth' of something and evaluating the performance of an IR system is an important part of developing an effective, efficient and usable search engine (Saracevic, 1995; Robertson, 2008). For example, it is necessary to establish to what extent the system being developed meets the needs of its end users, quantify the effects of changing the underlying search system or its functionality, and enable the comparison between different systems and search strategies. How to conduct IR system evaluation has been an active area of research for the past 50 years or so, and the subject of much discussion and debate (Saracevic, 1995; Robertson, 2008; Harman, 2011). This is due, in part, to the need to incorporate users and user interaction into evaluation studies and the relationship between the results of laboratory-based vs. operational tests (Robertson and Hancock-Beaulieu, 1992).

Harman (2011) describes IR evaluation as "*the systematic determination of merit of something using criteria against a set of standards*". This implies the need for a *systematic approach* for conducting an evaluation, the need for suitable *criteria* for evaluating search and the need to evaluate with respect to *standards*, for example using a standard benchmark or comparing against a baseline system or approach. Cleverdon et al. (1966) identified six evaluation criteria that could be used to evaluate IR systems: (1) coverage, (2) time lag, (3) recall, (4) precision, (5) presentation, and (6) user effort. Of these, precision and recall have been the most widely used to evaluate IR systems, However, the success of an IR system, especially from a user's perspective, goes beyond the performance of indexing and retrieval and may include how well the IR system supports users in carrying out their search tasks and whether users are satisfied with the results (Kelly, 2009; Clough and Goodale, 2013).

Evaluation of search systems can be carried out at various levels and may involve multiple methods of evaluation in an iterative manner during development and subsequent deployment. Saracevic (1995) distinguishes six levels of evaluation for information systems (including IR systems) as follows:

1. The *engineering level* deals with aspects of technology, such as computer hardware and networks to assess issues, such as reliability, errors, failures and faults.
2. The *input level* deals with assessing the inputs and contents of the system to evaluate aspects, such as coverage of the document collection.
3. The *processing level* deals with how the inputs are processed to assess aspects, such as the performance of algorithms for indexing and retrieval.
4. The *output level* deals with interactions with the system and output(s) obtained to assess aspects such as search interactions, feedback and outputs. This could include assessing usability.
5. The *use and user level* assesses how well the IR system supports people with their searching tasks in the wider context of information seeking behaviour (e.g., the user's specific seeking and work tasks). This could include assessing the quality of the information returned from the IR system for work tasks.
6. The *social level* deals with issues of impact on the environment (e.g., within an organisation) and could include assessing aspects such as productivity, effects on decision–making and socio–cognitive relevance.

Traditionally in IR evaluation there has been a strong emphasis on measuring system performance (levels 1-3), especially retrieval efficiency and effectiveness (Robertson, 2008; Harman, 2011). The creation of standardised benchmarks for quantifying retrieval effectiveness (commonly known as *test* or *reference collections*) is highly beneficial when assessing system performance (Robertson, 2008; Sanderson, 2010). However, evaluation at levels 4-6 is also important as it assesses the

performance of the system from the user's perspective and may also take into account the user's interactions with the system, along with broader effects, such as its impact and use in operation (Kelly, 2009; Borland, 2013; Wilson et al., 2010). In the following sections we discuss in more detail the two main approaches referenced in the literature: system-oriented and user-oriented evaluation.

3. System-oriented Evaluation

System-oriented evaluation of IR systems has typically focused on assessing ranked lists of results given a specification of a user's query, although attention has also been given to evaluating IR systems that comprise of multiple finding aids, such as visualisations or facets, and for tasks beyond search, such as exploration and browsing (Kelly et al., 2009). One of the first and most influential proposals for system-oriented evaluation was based upon the Cranfield methodology (Cleverdon, 1960). The Cranfield approach to IR evaluation uses test (or reference) collections: re-useable and standardised resources that can be used to evaluate IR systems with respect to the system (Cleverdon, 1991). Over the years the creation of a standard test environment has proven invaluable for the design and evaluation of practical retrieval systems by enabling researchers to assess, in an objective and systematic way, the ability of retrieval systems to locate documents relevant to a specific user need.

3.1. Evaluation using Test Collections

The main components of a standard IR test collection are the *document collection* (Section 3.2), statements of users' information needs, often called *topics* (Section 3.3), and for each topic an assessment about which documents retrieved are relevant, called *relevance assessments* (Section 3.4). These, together with evaluation measures (Section 3.5), simulate the users of a search system in an operational setting and enable the effectiveness of an IR system to be quantified. Evaluating IR systems in this manner enables the comparison of different search algorithms and the effects of altering algorithm parameters to be systematically observed and quantified. The most common way of using the Cranfield approach is to compare various retrieval strategies or systems (*comparative evaluation*). In this case the focus is on the relative performance between systems, rather than absolute scores of system effectiveness.

Evaluation using the Cranfield approach is typically performed as follows: (1) select different retrieval strategies or systems to compare; (2) use these to produce ranked lists of documents (often called *runs*) for each query; (3) compute the effectiveness of each strategy for every query in the test collection as a function of relevant documents retrieved; (4) average the scores over all queries to compute overall effectiveness of the strategy or system; and (5) use the scores to rank the strategies/systems relative to each other. In addition, statistical tests may be used to determine whether the differences between effectiveness scores for strategies/systems and their rankings are significant. This is necessary if one wants to determine the 'best' approach. This produces a list of relevant documents (often called *qrels*) for each query that is required in computing system effectiveness with relevance-based measures (e.g., precision and recall). In the TREC-style version of the Cranfield approach, there is a further stage required prior to (2) above, whereby the runs for each query are used to create a pool of documents (known as *pooling*). This is discussed further in Section 3.4.

Test collection-based evaluation is highly popular as a method for developing retrieval strategies (Harman, 2011). By modifying the components of a test collection and evaluation measures used, different retrieval problems and domains can be simulated, such as information retrieval, question answering, information filtering, text summarisation, topic detection and tracking, and image and video retrieval. Benchmarks can be used by multiple researchers to evaluate in a standardised manner and with the same experimental set up, thereby enabling the comparison of results. In addition, user-oriented evaluation, although highly beneficial, is costly and complex and often difficult to replicate. It is this stability and standardisation that makes the test collection so attractive.

However, information retrieval researchers have recognised the need to update the original Cranfield approach to allow the evaluation of new information retrieval systems that deal with varying search problems and domains (Voorhees, 2002; Ingwersen and Järvelin, 2005; Kamps et al., 2009). Research is ongoing to tackle a range of issues in information retrieval evaluation using test collections. For example, gathering relevance assessments efficiently (see Section 3.4), comparing system effectiveness and user utility (Hersh et al., 2000a; Sanderson et al., 2010); evaluating information retrieval systems over sessions rather than single queries (Kanoulas et al., 2011), the use of *simulations* (Azzopardi et al., 2010), and the development of new information retrieval evaluation measures (Yilmaz et al., 2010; Smucker and Clarke, 2012). Further information about the practical construction of test collections can be found in (Sanderson, 2010; Clough and Sanderson, 2013).

3.2. Document Collections

IR systems index documents that are retrieved in response to users' queries. A test collection must contain a static set of documents that should reflect the kinds of documents likely to be found in the operational setting or domain. Although similar in principle to traditional IR, in the case of semantic search a knowledge base (e.g., RDF data) is typically the data or document collection (also known as the *dataset*). Datasets can be constructed in different ways. For example, they can be *operationally derived* or *specially created* (Spärck Jones and Van Rijsbergen, 1976); the former refers to datasets created from a real-world setting, for example, via a snapshot of an organisation's existing data and the latter refers to datasets which have been compiled for a particular task. Datasets can also be *closed* or *domain-specific* (e.g., those used in biomedicine) or *open* or *heterogeneous* and spanning multiple domains (e.g., the Web). Datasets may also differ in size and type (e.g., media format) of data.

Examples of document collections include the Cranfield 2 collection (Cleverdon et al., 1966), covering a single domain -

aeronautics - and consisting of 1400 documents (research papers) written in English. An example of a larger collection is the ClueWeb09[2] collection of Web pages used in the TREC Web track. It was specially created (crawled from the Web) in 2009, spans various domains, and consists of more than 1 billion Web pages in 10 languages. In semantic search the geography dataset that forms part of the Mooney NL Learning Data (Tang and Mooney, 2001) has been used in several studies (Tang and Mooney, 2001; Kaufmann, 2007; Damljanovic et al., 2010). It was specially created in 2001 and covers a single domain - geography. It consists of around 5,700 pieces of information in the form of RDF triples published in English. An example of a larger dataset used in semantic search evaluations is DBpedia (Bizer et al., 2009). It is an extract of the structured information found in Wikipedia, that is operationally derived and created in 2009. It covers various domains, such as geography, people and music and consists of around 1.8 billion RDF triples in multiple languages, such as English, German and French.

3.3. Topics

IR systems are evaluated for how well they answer users' search requests. Therefore, the test collection must contain a set of statements that describe typical users' information needs (often referred to as *topics*). These may be expressed as short queries in the form commonly submitted to an IR system, questions, visual exemplars or longer written descriptions. In TREC, the format for most types of search task consists of four fields: a unique identifier (*number*), a query in the form of keywords, phrases or a question (*title*), a short description of the topic (*description*), and a longer description of what constitutes a relevant or non-relevant item for each topic (*narrative*). This template has been re-used in many IR evaluations.

Topics will vary depending on the search context being modelled. For example, topics for an image retrieval system may consist of visual exemplars in addition to a textual description (Grubinger and Clough, 2007; Müller, 2010). An example of a topic from the TREC-9 Web Track (Voorhees and Harman, 2000) is the following:

```
Number: 451
Title: What is a Bengals cat?
Description: Provide information on the Bengal cat breed.
Narrative:
 Item should include any information on the Bengal cat
 breed, including description, origin, characteristics, breeding
 program, names of breeders and catteries carrying bengals.
 References which discuss bengal clubs only are not relevant.
 Discussions of bengal tigers are not relevant.
```

The selection of realistic and representative topics is an important aspect of creating the test collection: the effectiveness of an IR system is measured on the basis of how well the system retrieves relevant items in response to given search requests. Ultimately, the goal for topic creation is to achieve a natural, balanced topic set accurately reflecting real world user statements of information needs (Peters and Braschler, 2001). There are various ways of obtaining typical search requests that may form the basis of topics. For example, analysing query and clickstream logs from operational search systems, utilising the findings of existing user studies, involving domain experts in the topic creation process, and conducting surveys and interviews amongst target user groups. Practically there will be a trade-off between the realism of queries and control over the testing of features for the search system being evaluated (Robertson, 1981). With respect to the number of queries required to effectively test an IR system, research has suggested that 50 is the minimum for TREC-style evaluations (Voorhees, 2009). However, results have also shown that making fewer relevance judgements over greater numbers of queries leads to more reliable evaluation (Carterette et al., 2008).

3.4. Relevance Assessments

For each topic in the test collection, a set of relevance judgements must be created indicating which documents in the collection are relevant to each topic. The notion of relevance used in the Cranfield approach is commonly interpreted as *topical relevance*: whether a document contains information on the same topic as the query. In addition, relevance is assumed to be consistent across assessors and static across judgements. However, this is a narrow view of relevance, which has been shown to be subjective, situational and multi-dimensional (Schamber, 1994; Mizzaro, 1998; Saracevic, 2007). Some have speculated that the variability with which people judge relevance would affect the accuracy with which retrieval effectiveness is measured. However, a series of experiments were conducted to test this hypothesis (Cleverdon, 1970; Voorhees, 1998) with results showing that despite there being marked differences in the documents that different assessors judged as relevant or non-relevant, the differences did not substantially affect the relative ordering of IR systems being measured using the different assessments.

To assess relevance, different scales have been used. The most popular of these are binary and graded relevance scales. When using a *binary relevance* scale, a document is judged as either relevant or non-relevant; in the case of *graded relevance* a document is judged for relevance on a scale with multiple categories (e.g., *highly relevant*, *partially relevant* or *non-relevant*). Robertson (1981) argues that relevance should be treated as a continuous variable and hence, different levels of relevance should be incorporated in an evaluation model. Therefore, researchers have attempted to experiment with non-dichotomous relevance scales (Cuadra, 1967; Eisenberg, 1988; Janes, 1993; Spink et al., 1998; Tang et al., 1999). The study by Tang et al. (1999) showed that a graded relevance scale with seven points led to the highest levels of confidence by the judges during their assessments. The additional benefit of using graded relevance scales is that a wider range of system effectiveness measures, such as Discounted Cumulative Gain (DCG), can be used (see Section 3.5). In recent years, this use of ordinal relevance scales, together with the appropriate measures, have become more common. For instance, a three-point relevance scale together with DCG as a measure were used in the TREC Entity Track 2009; whilst the SemSearch evaluation (see Section 5.2)

[2] http://lemurproject.org/clueweb09/

similarly used a three-point relevance scale and a normalised version of DCG as the evaluation measure.

In the TREC approach to building test collections *pooling* is used to generate a sample of documents for a given topic that can be assessed for relevance. Ideally, for each topic, all relevant documents in the document collection should be found. However, pooling may only find a subset of all relevant items which will affect the measurement of system effectiveness. Generating complete sets of relevance judgements helps to ensure that, when evaluating future systems, improvements in results can be detected. The effects of incomplete relevance assessments, imperfect judgements, potential biases in the relevance pool and the effects of assessor domain expertise in relation to the topic have been investigated in various studies (Cuadra, 1967; Zobel, 1998; Buckley and Voorhees, 2004; Yilmaz and Aslam, 2006; Büttcher et al., 2007; Bailey et al., 2008; Kinney et al., 2008). Approaches to ensure completeness of relevance assessments include using the results from searches conducted manually to generate the pools and supplementing pools with relevant documents found by manually searching the document collection with an IR system, known as Interactive Search and Judge or ISJ (Cormack et al., 1998)

Generating relevance assessment is often highly time-consuming and labour intensive. This often leads to a bottleneck in the creation of test collections. Various 'low-cost evaluation' techniques have been proposed to make the process of relevance assessment more efficient. These include approaches based on focusing assessor effort on runs from particular systems or topics that are likely to contain more relevant documents (Zobel, 1998), sampling documents from the pool (Aslam et al., 2006), simulating queries and relevance assessments based on users' queries and clicks in search logs (Zhang and Kamps, 2010) and using crowdsourcing (Alonso and Mizzaro, 2009; Kazai, 2011; Carvalho et al., 2011).

3.5. Evaluation Measures

Evaluation measures provide a way of quantifying retrieval performance (Manning et al., 2008; Croft et al., 2009). Although many properties could be assessed when evaluating a search system (Clough and Goodale, 2013), the most common is to measure retrieval effectiveness: the ability of an IR system to discriminate relevant from non-relevant documents. The most well known measures are *precision* and *recall* (Kent et al., 1955). Precision measures the proportion of retrieved documents that are relevant; recall measures the proportion of relevant documents that are retrieved.

However, these measures do not take the ordering of results and the behaviour of end users into account. For example, research has shown that users are more likely to select documents higher up in the ranking (*rank bias*) and often start at the top of a ranked list and work their way down. Therefore, to accommodate for this, *ranked-based* evaluation measures are commonly used, some of which take into account characteristics of the end user (e.g., DCG and ERR).

3.5.1. Binary-Relevance Measures

When used for assessing relevance of the results, the following measures consider a result item to be either relevant or non-relevant.

Average Precision (AP). A common way to evaluate ranked outputs is to compute precision at various levels of recall (e.g., 0.0, 0.1, 0.2, ... 1.0), or at the rank positions of all the relevant documents. The scores are then averaged to produce a single score for each topic (average precision). This can be computed across multiple queries by taking the arithmetic mean of average precision values for individual topics. This single-figure measure of precision across relevant documents and multiple queries is referred to as *Mean Average Precision* (or MAP).

Precision@k (P@k). Measures the proportion of relevant items found in the top k results (also known as the cutoff value) for a specific topic. Precision@k is commonly used in assessing the performance of Web search engines using *k=10* or *k=20*. An issue with Precision@k, however, is that the score is sensitive to the choice of cutoff value and number of relevant documents for each topic. For example, if there are 5 relevant documents in total for a topic and Precision@10 is used, even if all relevant items are returned in the top 10 results the maximum score will only be 0.5.

R-Precision (R-Prec). Because the number of relevant documents can influence the Precision@k score as previouly mentioned, R-precision can be used (R refers to the number of relevant documents for a specific topic). The use of an unfixed/changeable cutoff value guarantees that a precision of 1 can be achieved. When this measure is used, precision and recall are of equal values since both of them are calculated as the number of relevant documents retrieved/overall number of relevant documents (R).

Mean Reciprocal Rank (MRR). Kantor and Voorhees (2000) defined this measure in order to evaluate the performance of IR systems in retrieving a specific relevant document, also known as a *known-item search*. It is calculated as the mean of the reciprocal ranks (RR) for a set of queries as illustrated in Equation 1 in which $|Q|$ is the number of queries and $rank_i$ is the rank of the first relevant result for query number i. Thus, for a specific query, RR is the reciprocal of the rank where the first correct/relevant result is given. Although this measure is mostly used in search tasks when there is only one correct answer (Kantor and Voorhees, 2000), others used it for assessing the performance of query suggestions (Meij et al., 2009; Albakour et al., 2011) as well as ranking algorithms in particular (Damljanovic et al., 2010) and IR systems (Voorhees, 1999, 2003; Magnini et al., 2003) in general.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

3.5.2. Graded-Relevance Measures

Although the above measures are commonly used in both IR and semantic search evaluations, their main limitation is that they must be used with a binary relevance scale. As discussed in Section 3.4, this is insufficient when comparing IR systems based on different levels of relevance (or graded-relevance judgements). The remainder of this section describes measures that are widely used with graded relevance assessments.

Cumulated Gain (CG). Järvelin and Kekäläinen (2000) based this measure on the observation that "*highly relevant documents are more valuable than marginally relevant documents*". Therefore, the more relevant a retrieved document is (with higher relevance grade), the more gain the evaluated IR system achieves. This gain is accumulated for the documents and thus the *CG* is calculated according to Equation 2 in which $G[i]$ is the relevance value of the document at position i.

$$CG[i] = \begin{cases} G[1] & \text{if } i = 1 \\ CG[i-1] + G[i] & \text{otherwise} \end{cases} \quad (2)$$

Discounted Cumulated Gain (DCG). The direct cumulated gain (CG) does not account for ranking: differences in the ranking of documents do not change its value. To account for ranking and based on their second observation: "*the greater the ranked position of a relevant document (of any relevance level) the less valuable it is for the user, because the less likely it is that the user will examine the document*", Järvelin and Kekäläinen (2000) defined DCG with a discounting function to reduce the credit given for lower-ranked results. This function is chosen as the log of a document's rank in the list of results. DCG is calculated according to Equation 3. Again, $G[i]$ is the relevance value of the document at position i. Finally the log base b can be adjusted as required, for instance, to have high discounts for Web search users who are interested in getting the most relevant results as highly ranked as possible.

$$DCG[i] = \begin{cases} G[1] & \text{if } i = 1 \\ DCG[i-1] + \frac{G[i]}{\log_b(i)} & \text{otherwise} \end{cases} \quad (3)$$

Normalised Discounted Cumulated Gain (NDCG). Similar to how precision@k is influenced by the chosen cutoff value (k), the CG and DCG measures are influenced by the number of relevant documents for a query. This limitation prevents the comparison of different (D)CG values for different queries. This is tackled in the NDCG in which the DCG values are normalised with respect to an *ideal result list*. To calculate the NDCG, the DCG values are divided by the equivalent ideal values (those in the same position). As illustrated by Järvelin and Kekäläinen (2002), if the (D)CG vector V of an IR system is $< v_1, v_2, ..., v_k >$, and the ideal (D)CG vector I is $< i_1, i_2, ..., i_k >$, then the n(D)CG vector is given by

$$normVect(V, I) =< v_1/i_1, v_2/i_2, ..., v_k/i_k > . \quad (4)$$

Expected Reciprocal Rank (ERR). The main advantage of the CG measures is that they account for both the rank and the relevance level of retrieved items. However, they do not account for previous items found in the list of results and how they affect the relevance/usefulness of the current document. In contrast to this *simple position* model that assumes independence between relevance of documents found in a ranked result list, a *cascade* model is one in which the relevance of one document is influenced by the relevance of documents ahead in the result list. Craswell et al. (2008) showed that the latter better explains web search users behavior: *"users view results from top to bottom and leave as soon as they see a worthwhile document"*. Therefore, Chapelle et al. (2009) proposed *Expected Reciprocal Rank (ERR)* which is based on the cascade model in an attempt to have a more accurate measure reflecting users' satisfaction. The ERR, the position at which a user is satisfied and examines no more documents, is calculated according to Equation 5[3] in which n is the number of documents in the ranking.

$$ERR = \sum_{r=1}^{n} \frac{1}{r} P(user\_stops\_at\_position\_r) \quad (5)$$

Together, the test collection and evaluation measures provide a simulation of the user of an IR system. For example, in the case of ad-hoc retrieval, the user is modelled as submitting a single query and being presented with a ranked list of results. One assumes that the user then starts at the top of the ranked list and works their way down examining each document in turn for relevance. This, of course, is an estimation of how users behave; in practice they are often far less predictable. The choice of evaluation measure depends on several aspects, such as the scale of relevance adopted, the number of queries available as well as the number of results considered/assessed for each query, and also the purpose/goal of the search task (Buckley and Voorhees, 2000). For example, additional measures have been developed that also take into account other properties of the search results, such as *novelty* and *diversity* (Clarke et al., 2008).

Various studies have compared these measures against each other, especially with respect to their stability and discriminative power. Tague-sutcliffe and Blustein (1994); Buckley and Voorhees (2000); Voorhees and Buckley (2002) and Sakai (2006) showed that precision at a fixed level of rank (P@k) was usually found to be the least discriminating with the highest error rates among other measures, such as R-Precision and MAP. This is mainly due to the influence of the choice of the cutoff value on the results. Although Sanderson and Zobel (2005) confirmed this finding, they showed that when taking the assessor effort into consideration, P@10 is much more stable than MAP since it required only around 14% of the assessor effort required to calculate MAP. When comparing R-Precision and MAP, Tague-sutcliffe and Blustein (1994) concluded that MAP had a higher discriminating power. A similar finding was made by Buckley and Voorhees (2000); however, the latter showed

[3] The reader can refer to Chapelle et al. (2009) for a detailed explanation of this equation.

that the two measures had almost equal error rates. Studies have also investigated the stability of graded relevance-based measures and in comparison with those based on binary relevance. Sakai (2007) found that NDCG was as stable and sensitive as MAP while Radlinski and Craswell (2010) found that the first is more stable when a small number of queries is used. When tested with query set sizes from 5 to 30, the authors showed that MAP results could be misleading since the worse ranking was sometimes statistically significantly better.

4. User-oriented Evaluation

Ultimately IR systems are used by people in an interactive way and so require humans to determine success. Evaluation from a user-oriented perspective is important in going beyond retrieval effectiveness to assess retrieval performance, for example assessing user satisfaction with the results, usability of the interface, whether users are engaged with the system, user performance with a task and the effects of changes in the retrieval system on user behaviour (Tague and Schultz, 1989; Saracevic, 1995; Harter and Hert, 1997; Su, 1992; Saracevic, 1995; Voorhees, 2002; Ingwersen and Järvelin, 2005). This requires going beyond the traditional Cranfield-style IR experiment and various studies have been carried out from an Interactive Information Retrieval (IIR) perspective, such as those in TREC in the 1990s (Over, 2001; Kelly and Lin, 2007) along with many others (Su, 1992; Dunlop, 1996; Koenemann and Belkin, 1996; Xie, 2003; Petrelli, 2008; Hearst, 2009).

In a user-oriented approach to evaluation people are involved in some way, although their level of involvement can vary widely (Kelly, 2009). A common approach is to recruit users to participate in retrieval tasks in a controlled lab-based environment. Their interactions with the system are recorded, along with their feedback on the system and information about their individual characteristics (e.g., age and cognitive abilities). Alternative types of study include comparing results from multiple systems in a side-by-side manner (Thomas and Hawking, 2006); A/B testing, where a small proportion of traffic from an operational system is directed to an alternative version of the system and the resulting user interaction behaviour compared (Manning et al., 2008); and use of search engine log data to observe how users click patterns for relevant documents vary. In this case, a relevant document could be determined based on signals from the log data, e.g. dwell time and rank position of the clicked item are often strong indicators of relevance (Baeza-Yates and Ribeiro-Neto, 2011).

The remainder of this section discusses important aspects that should be addressed in conducting a user-oriented evaluation, such as the criteria to be assessed (Sub-section 4.1), the experimental setup (Subsection 4.2) and the choice of data collection methods (Subsection 4.3).

4.1. Criteria and Measures

IR systems are used by people to fulfil their information needs. Simply evaluating in terms of retrieval effectiveness is limiting. Indeed, various studies have showed that users' satisfaction and success at performing a search task does not always correlate with high retrieval effectiveness (Hitchingham, 1979; Su, 1992; Tagliacozzo, 1997; Hersh et al., 2000b; Turpin and Hersh, 2001; Hersh et al., 2002; Huuskonen and Vakkari, 2008). In part, this is because users are able to adapt to poorly performing systems. Additionally, other factors influence a user's satisfaction with the search results, e.g., their domain knowledge and expertise; aspects of retrieved results such as its quality, language or authoritativeness; the presentation of search results; as well as the usability of a search system user interface.

Criteria and measures concerned with how well users achieve their goals, their success and satisfaction with the results have been used to evaluate IR systems and may be assessed by quantifying efficiency, utility, informativeness, usefulness, usability, satisfaction and the user's search success. Many forms of criteria and associated evaluation measures have emerged in the literature of user-oriented IR. For example, Wilson et al. (2010) identify various criteria (and measures) that broadly reflect the levels identified in Section 2, but grouped into three contextual levels: (1) evaluation within the IR context; (2) evaluation within the information seeking context; and (3) evaluation within the work context. Kelly (2009) also identifies four basic measures: (1) *contextual*, that captures characteristics of the subject and tasks undertaken (e.g., age, gender, familiarity with search topics); (2) *interaction*, that captures aspects of the user-system interaction (e.g., number of queries issued, number of documents viewed, etc.); (3) *performance*, that relates to the outcome of user's interactions (e.g., number of relevant documents saved, precision, NDCG, etc.); and (4) *usability*, that captures evaluative feedback from subjects (e.g., satisfaction, suggestions, attitudes, etc.).

The term *relevance* has been vaguely and inconsistently used in IIR literature (similarly to IR). For example, Vickery (1959b,a); Taube (1965) and Soergel (1994a) used it to refer to the degree of match between a document and a question; in contrast, Foskett (1972); Kemp (1974); Goffman and Newill (1966) and Goffman (1967) distinguished between *relevance* as a notion similar to system-relevance where this degree of match or relation between a document and a question is assessed by an external judge/expert and *pertinence* to refer to the user-relevance in which the assessment can only be performed by the real user with the information need represented in the question. Often, measures adopted in user-oriented studies are those which account for non-binary, subjective relevance assessments usually given by real users. These include the cummulated gain measures presented earlier; the *relative relevance* (Borlund and Ingwersen, 1998) and *ranked half-life* (Borlund and Ingwersen, 1998) which are proposed specifically for IIR; as well as the binary-based measures: *expected search length* (Cooper, 1968) and *average search length* (Losee, 1998) proposed earlier in literature.

Relative Relevance (RR). Relevance in IR has centred around two notions: *system-relevance* (also referred to as objective relevance) and *user-relevance* (referred to as subjective relevance). In an attempt to bridge the gap between both notions and evaluate the retrieval performance of an IR system with respect to both of them, Borlund and Ingwersen (1998) proposed the rela-

tive relevance measure. It is based on calculating an association between the two kinds of relevance based on the Jaccard measure. It also helps understanding if, for instance, one IR system outperforms another when evaluated objectively but not when evaluated subjectively from a user's perspective. However, it does not take into account the rank of the retrieved document.

Ranked Half-Life (RHL). In contrast to RR, this measure is more related to the *Expected Search Length (ESL)* and the *Average Search Length (ASL)* (described below) since it evaluates an IR system with respect to its ability to position relevant documents high in a ranked result list. It is defined as the position at which half of the relevant documents are retrieved. Losee (1998) explains that the advantage of having this median ranking is that its increase indicates that more highly-relevant documents were ranked at the top of the result list while its decrease indicates that these relevant documents were ranked in low or scattered positions in the result list. However, Järvelin and Kekäläinen (2002) argue that this is a disadvantage of the measure, similar to ASL, since it is affected by outliers: relevant documents ranked at low positions.

Expected Search Length (ESL). Cooper (1968) criticised most of the IR traditional measures based on precision and recall especially for not being able to report a single measure of a system's performance and for not accounting for the user's need while evaluating the system's performance. Therefore, he proposed the ESL to provide a single measure for assessing an IR system's performance in helping the user find the required amount of relevant documents while reducing the effort wasted in examining irrelevant documents. It is calculated by finding the number of irrelevant documents that appear before the user retrieves the required k relevant documents. Specifying the value of k is often considered a disadvantage, especially since it will likely differ according to the user and the query.

Average Search Length (ASL). The ASL is the expected position of a relevant document in a ranked result list. It is a related measure to the ESL since it similarly takes into account the user's effort wasted in examining irrelevant documents and credits an IR system for positioning relevant documents high in the list to reduce this effort. However, both ESL and ASL are criticised for allowing only binary-relevance assessments and thus they do not account for the degree of relevance of a document, unlike the cumulated gain measures.

4.1.1. Efficiency

The International Standards Organisation (ISO) defines efficiency as the *"resources expended in relation to the accuracy and completeness with which users achieve goals"* (ISO, 1998). IIR focuses on the interaction between users and IR systems and assesses the success of users in achieving their goals through this interaction. Therefore, efficiency of the users in this process has been one of the main evaluation criteria studied in IIR. While some work investigated the degree of correlation between efficiency and the user's overall success or satisfaction, others proposed and examined different measures that can be used to assess efficiency. The most common measures of efficiency are time and those based on effort since both can indicate the user's efficiency in achieving a specific goal with a specific IR system. Dunlop (1986); Su (1992, 1998); Tsakonas et al. (2004) and Johnson et al. (2003) used search time (also referred to as completion time) to measure efficiency. This is usually the time from when a user starts a specific search task till its end. In contrast to measuring user-efficiency, response time has been used to measure system-efficiency (Dong and Su, 1997; Johnson et al., 2003). Additionally, user effort has been measured in different ways: number of search terms used in a specific task (Dunlop, 1986); number of commands used (Dunlop, 1986), and also number of queries issued to complete a specific task (Su, 1998). Tsakonas et al. (2004) also used the number of completed tasks in a time period or a session – which can be seen as the inverse of the search time – to measure efficiency. Su (1992); Dong and Su (1997); Su (1998) and Johnson et al. (2003) found that efficiency of an IR system was an important factor that influenced user's overall success and satisfaction.

4.1.2. Learnability

Learnability, used interchangeably with the term *ease of learning*, is an important aspect of usability that focuses on the ease with which users learn how to use a system or an interface. Shackel (1986) describes learnability as the relation of performance and efficiency to the training and frequency of use. Nielsen (1993) discusses how learnability can be measured in terms of the time required for a user to be able to perform certain tasks successfully or reach a specified level of proficiency. A similar definition is given by Shneiderman (1986) as "*the time it takes members of the user community to learn how to use the commands relevant to a set of tasks*".

Nielsen (1993) argues that learnability could be seen as the most fundamental usability attribute since, in most cases, systems need to be easy to learn. Tullis and Albert (2008) additionally argue that measuring usability in a one-time evaluation might be misleading since the use of some applications/systems requires longer-term use and therefore assessing learnability would be essential. Past studies on learnability have focused on *initial learnability*, referring to the initial performance with the system; whilst others have studied *extended learnability*, referring to the change in performance over time (Grossman et al., 2009). Additionally, Roberts and Moran (1983); Whiteside et al. (1985); Davis et al. (1989); Hearst et al. (2002) showed that learnability and usability are congruent.

4.1.3. Utility

While Foskett (1972); Kemp (1974); Goffman and Newill (1966) and Goffman (1967) tried to differentiate between *relevance* and *pertinence* in an attempt to account for the subjectivity of relevance assessments, Saracevic et al. (1988); Cooper (1973b,c) and Soergel (1994b) argued that utility was a more appropriate measure for evaluating IR systems and their ability to support users in their search tasks. Soergel (1994b) explained that a document has utility if it is pertinent (relevant as perceived by the user) and also contributes to the user knowledge in the context of the query (by providing information that

was previously unknown). Utility is usually adopted as a measure of usefulness and worth of the answers provided by the IR system to its users. The argument for using utility in contrast to relevance or pertinence is that a document that has information about the user query does not have to be 'useful' for the user since this depends on other aspects. For instance, the user might already know this information or it can be in a language or format that is not understood by the user. Additionally, the clarity, reliability and credibility of a document also affect its usefulness (Soergel, 1994b; Cooper, 1973b,c). There have been long debates on the ultimate ways to assess this criterion since it can be difficult to be quantified. Therefore, the most common ways have been through the use of questionnaires gathering users' answers on different questions usually identified by the researcher depending on the goal of the study. For instance, Saracevic et al. (1988) included questions to understand the degree of informativeness and in turn usefulness of a document such as the following: *"On a scale of 1 to 5, what contribution has this information made to the resolution of the problem which motivated your question?"*

4.1.4. User Satisfaction

Both Cooper (1973b) and Saracevic et al. (1988) used the terms *utility* and *satisfaction* equally to refer to the overall value of a search result or a document to the user. Tessier et al. (1977); Su (1992); Crawford et al. (1992); Draper (1996); Su (1998) and Loupy and Bellot (1997) used satisfaction as a multidimensional measure to evaluate IR systems from a user's perspective. Satisfaction based on a single or a group of search results or documents is highly subjective and depends on various factors related to the user (e.g., knowledge or personal preferences), the IR system (e.g., responsiveness, interface or aesthetics) and the search results (e.g., completeness, accuracy or format). Similar to utility, satisfaction is often measured using questionnaires which address the factors mentioned here. In assessing users' satisfaction with libraries, Tessier et al. (1977) included factors such as the output, the library as a whole, its policies as well as the interaction with the library staff. Other factors commonly used in studies include completeness of search results (Su, 1992, 1998); interface style (Crawford et al., 1992; Hildreth, 2001; Manning et al., 2008); interaction with the system (Crawford et al., 1992; Griffiths et al., 2007; Manning et al., 2008); response time (Kelly, 2009); features and functionality of the system (Griffiths et al., 2007); ease of query formulation (Loupy and Bellot, 1997) and ease of use (Hildreth, 2001). Cooper (1973a); Saracevic et al. (1988) and Draper (1996) argued that satisfaction is the ultimate measure to evaluate IR systems. A counter argument by Soergel (1994b) and Belkin and Vickery (1985) explained that utility – in contrast to satisfaction – evaluates IR systems with respect to their main functionality, which is to help users find 'useful' information that can be used to address their information needs.

4.2. Experimental Setup

"An experiment is an examination of the relationship between two or more systems or interfaces (independent variable) and set of outcome measures (dependent variables)" (Hoeber and Yang, 2007). A user-oriented evaluation approach seeks to involve real users in the assessment process of the IR system, take into account real information needs and adopt appropriate user-oriented criteria and measures. In a very broad sense, this is usually an experiment in which real users are recruited to assess a number of IR systems performing specific search tasks with respect to predefined criteria and in which different forms of data are usually collected (e.g., interaction logs, post-task questionnaires, etc.). A common procedure for user studies involving IR systems (also referred to as *study protocol* Kelly (2009) includes the following (Hoeber and Yang, 2007):

1. Assign participants various 'realistic' tasks to perform.
2. Take quantitative measurements of 'performance' (e.g., time taken, number of tasks completed, number of errors made, etc.).
3. Make observations about how the interface/system is being used by the participants.
4. Collect subjective reactions from the participants (e.g., satisfaction, usability)

This experiment process requires careful design choices of several factors that can influence the results and reliability of the evaluation. The rest of this section discusses some of these factors, including the experiment type (laboratory/operational), the experiment design (within/between- subjects), recruitment of subjects (e.g., selection procedure and number of subjects) and search tasks (e.g., number and type of tasks).

4.2.1. Lab-based versus Naturalistic Settings

Experiments can be carried out in a laboratory (also referred to as *controlled or formal experimentation*) or an operational setting (also referred to as *contextual inquiry* or *naturalistic observation*). In the first setting, the experiment takes place in a laboratory where the researcher has (some) control over the experimental variables and environment. In the latter, the experiment takes place in the real operational/natural setting where the IR system is typically used. There are known advantages and disadvantages and important issues acknowledged for each setting and therefore should be considered by the researcher for a careful choice (Tague-sutcliffe, 1992; Robertson, 1981; Borlund and Ingwersen, 1997; Petrelli, 2008). First, the laboratory setting offers more control of independent variables that may affect the outcomes of the experiment. This is difficult if not impossible to achieve in a real setting. This control can be a necessity in some studies, especially those attempting to answer specific questions or examine the effect of one or more variables on another as opposed to open/free studies such as those analysing users behaviour or strategies during search. In these studies, the operational setting might be preferred since it provides the ability to observe users in real scenarios as opposed to simulating them. However, this realism is also the disadvantage of this setting since, beside the lack of control mentioned earlier, it also prevents the repeatability of the experiment, an important aspect especially in large-scale evaluations. In an attempt to have the best of both worlds, experiments can be per-

formed in a combination of both settings (Walker and DeVere, 1990; Walker et al., 1991; Robertson, 1981).

4.2.2. Within or Between Subjects Design

Another important choice for the experiment is with respect to the use of subjects. For example, if comparing two IR systems, subjects might be asked to test only one system (known as *between-subjects* or *independent* design) or test both systems (known as *within-subjects* or *repeated measures* design) (Kelly, 2009). The advantages of a within-subjects design are that the subjects test all IR systems and therefore they are able to compare the results of multiple systems. In addition, fewer subjects are required to conduct the experiment because subjects test multiple systems. However, a disadvantage of this experimental design is the *carryover/order* effect. This occurs when the subject 'carries over' certain effects from the experiment with one system to the next in 'order'. These effects include learning effects and pre-conditioning, as well as emotional effects, such as tiredness, boredom and frustration.

The results of these effects can be reduced if the order of the search systems or topics used is varied. One possible solution to achieve this order variation is through counterbalancing (e.g., two subject groups evaluate two search systems in reverse order) (Tague-sutcliffe, 1992). To apply counterbalancing, a *Latin Square* (to control effect of one variable) or *Graeco–Latin Square* design (to control the effects of multiple variables) are the most common approaches used for overcoming any topic or system ordering effects that may influence the obtained results (Kelly, 2009, pp. 44-60).

4.2.3. Recruitment of Subjects

Identifying the 'right' number of subjects to recruit for an IIR or usability study is an open question. Nielsen (1993, 1994); Lewis (1994) and Virzi (1990, 1992) argue that five or fewer subjects are sufficient to identify most of the usability problems found in a system. Virzi (1990, 1992) showed that this could account for around 80% of the problems. However, other studies recommended using more subjects. For example, Turner et al. (2006) suggested that seven subjects may be optimal; Perfetti and Landesman (2001) argued that more than eight subjects are required; Spyridakis (1992) called for using a minimum of 10-12 subjects and Faulkner (2003) explained that 15 subjects are required in order to detect between 90 to 97% of the problems.

Similarly, this has been a difficult choice for IIR evaluations since it usually includes a trade-off between available resources and validity of the evaluations. For example, interactive CLEF (Gonzalo and Oard, 2004; Gonzalo et al., 2006) used a minimum of eight subjects while the Interactive Track at TREC-9 and TREC-6 used 16 and 20 searchers respectively (Hersh and Over, 1999; Over, 1997). The number of subjects directly influences the amount of resources required in terms of cost, time and effort. Additionally, there is the common difficulty of finding volunteers with the required characteristics, such as a specific age group or knowledge in a particular field.

Another important aspect with recruiting subjects is the choice of a representative sample of the target population. For instance, if an IR system is operationally used only by librarians then recruiting subjects from a different domain would bias the results. Also, the type of users is an important factor and there have been two main approaches for categorising users: in terms of search experience and skills or domain/field experience and knowledge (Hsieh-Yee, 2001; Hölscher and Strube, 2000). For instance, Navarro-Prieto et al. (1999); Hölscher and Strube (2000) and Tabatabai and Shore (2005) differentiated between *inexperienced/casual users* and *experienced/expert users* according to their web expertise which was defined by Hölscher and Strube (2000) as "*the knowledge and skills necessary to utilise the World Wide Web and other Internet resources successfully to solve information problems*". Evaluating systems with the different types of users and comparing their results could help in understanding the suitability of certain search approaches to specific type of users or furthermore, the different requirements to cater for when targeting one of these types of users.

4.2.4. Tasks and Topics

Subjects may be provided with a set of tasks to carry out during an experiment, or may be asked to think of their own. Tasks could include: *specific fact-finding* (e.g., finding someone's phone number), *extended fact-finding* (e.g., finding books by the same author), *open-ended browsing* (e.g., identifying any new work on voice recognition from Japan), and *exploration* (e.g., finding out about your family history from an online archive) (Shneiderman, 1986, p. 512). For more browsing-oriented tasks, such as exploration, then evaluation may go beyond simply dealing with ranked lists of results and assessing visualisations (Isenberg et al., 2013).

In an attempt to provide a balance between realism and control, Borlund and Ingwersen (1997) proposed the *simulated work task* which is a short cover story describing a situation that leads to the information need. The authors explain how this helps in simulating real life by allowing individual interpretations of the situation. Typically a simulated work task situation includes: a) the source of the information need; b) the environment of the situation; and c) the problem which has to be solved. Researchers used one or more of these types of tasks depending on the research goal/questions. Search tasks/topics were used at TREC and CLEF interactive tracks (Over, 1997; Gonzalo et al., 2006); while Jose et al. (1998); Borlund (2000); White et al. (2007) and Petrelli (2008) adopted Borlund's simulated work task. In the same context, another approach to achieve realism and motivate and engage the recruited subjects in the evaluation is to let them choose the search tasks from a pool of available tasks, for instance, in different domains (Spink, 2002; Su, 2003; White et al., 2007; Joho et al., 2008).

Additionally, the number of tasks included in a user study is an important aspect to consider as this has an impact on cost, time and effort required to conduct the study. Some studies have imposed time limits on each task to allow for a specific number of tasks within particular period of time (Over, 1997; Kaki and Aula, 2008). For instance, six tasks were included in the TREC-6 Interactive Track with a time limit of 20 minutes for each task (Over, 1997). The total amount of time required

for the experiment was around three hours. This was the same amount of time which the organisers of iCLEF 2004 found to be the longest for subjects to offer in a single day (Gonzalo and Oard, 2004). However, in this study, 16 different tasks were performed by the subjects: too many when compared to the small number of tasks commonly adopted in other studies (Borlund, 2000; Over, 1997; Kaufmann, 2007; Elbedweihy et al., 2012a). Indeed, the total amount of time required depends on many factors, especially the type of tasks as well as steps performed in the experiment (e.g., briefing, questionnaires, interviews, etc).

4.3. Data Collection Methods

Another important design choice within IIR studies concerns the methods employed to collect the data generated from the experiments which is influenced by the evaluation criteria and measures. For instance, to assess systems with respect to users' satisfaction, usually approaches for gathering subjective data are used, such as questionnaires or interviews with the recruited subjects. In contrast, system event logs are often used for collecting objective data, such as different timings required for assessing user-efficiency.

4.3.1. Event Logs

Event logs (also referred to as system usage logs, query logs, clickstream data or transaction logs) are a type of *unobtrusive method*, which typically collect data without direct engagement of the user (McGrath, 1995; Page, 2000; Webb et al., 2000). Although the main definition in literature for logs is a method that captures the interaction between an IR system and its users, including the content, type and time of transactions (Rice and Borgman, 1983; Peters, 1993), it is also used in IIR user-studies to refer to other types of data automatically recorded while users are performing the search tasks. This data is usually gathered for performance- or efficiency- based measures. An example is using software to record different timings, such as search time per task or response time. Logs are widely used as an inexpensive data collection method – in contrast to questionnaires or interviews which require time from the recruited subjects – that allows large amounts of different types of data to be generated. For instance, this includes gathering the number of search reformulations attempted by users for a search task or the time it takes them to formulate their queries in a specific interface.

4.3.2. Think Aloud

Think aloud (Nielsen, 1993) is a well-established method for gathering data in user-studies (Ericsson and Simon, 1993; Over, 2001; Jenkins et al., 2003; Wrigley et al., 2012). In essence, it attempts to gain more insight and understanding of the user's actions and strategies; their interactions with and perception of the system as well as their behaviour and rationale during different scenarios or search tasks. The users are therefore asked to *think-aloud* while performing the tasks, for instance, explaining what they are doing and why, or any problems or difficulties they face. Various researchers showed that this method allows generating valid qualitative data that could be used to study cognitive tasks (Rhenius and Deffner, 1990; Ericsson and Simon, 1993). However, this validity was questioned by other researchers arguing that thinking aloud could influence users performance and behaviour during completing the tasks. For instance, Rhenius and Deffner (1990) and Bainbridge (1979) found that thinking aloud slows down users in carrying out their tasks. Additionally, Ingwersen (1992) argued about the reliability of the data generated from this method since there is no guarantee that it reflects the real behaviour of the users. Finally, since these loud-thoughts are usually audio-recorded, this results in the method being expensive, requiring large amount of resources (time, effort, personnel) to analyse the data.

4.3.3. Questionnaires

This is one of the most commonly used data collection methods in user-studies (Harper, 1996; Kelly, 2009). Pre-search questionnaires are often used to gather information about subjects' knowledge, experience or skills in a specific field. Hence, they are common in studies investigating the effect of specific tasks or systems on subjects and the resulting changes in this knowledge. When included, the demographics questionnaire is used to gather subjects' demographics data. Most of the time, these two questionnaires are merged in one questionnaire presented to the subjects at the beginning of the experiment. The data collected can be used to identify correlations (for instance, between performance and age) or specific behaviour (for instance, different search strategies depending on level of experience). Typically, standardised questionnaires are utilised in post-task or post-experiment questionnaires aimed at capturing feedback about the usability of the system and subjects' satisfaction. System Usability Scale (SUS) (Brooke, 1996), Computer System Usability Questionnaire (CSUQ) (Lewis, 1995), Questionnaire for User Interface Satisfaction (QUIS) (Chin et al., 1987) are some of the widely used satisfaction questionnaires in HCI. For instance, the SUS comprises ten questions which are answered on a 5-point Likert scale identifying subjects' views and opinions of the system. The test incorporates a diversity of usability aspects, such as the need for support, training and complexity.

In addition to the above methods, most researchers also use observations to gain insight into the subjects' behaviour while performing the required tasks as well as any problems or difficulties facing them. Furthermore, interviews are also used as an alternative to, or together with, questionnaires to gather information about subjects' satisfaction.

5. Evaluation Initiatives

In 1992, the National Institute of Science and Technology (NIST) announced the first Text REtrieval Conference (TREC) to encourage IR community researchers to conduct experimental evaluation using a 'standard' test collection. TREC continues to run an annual cycle of releasing re-useable benchmarks and meetings to discuss results – a process which has proven highly influential in the development of IR techniques (Spärck Jones, 2000). While 25 groups participated in TREC-1, the current number of participants is significantly higher (as too are the sizes of the datasets). Driven (in part) by the success

and achievements of TREC, a number of other large-scale IR evaluations have been run. These include the Cross-Language Evaluation Forum (CLEF; started in 2000) for evaluating multimodal and multilingual information access methods and systems, as well as Interactive Information Retrieval (IIR), such as those embodied within TREC – *Interactive Track* (Hersh and Over, 2000) and *Complex Interactive Question-Answering* (ciQA) (Kelly and Lin, 2007) – which involved real users to create topics or evaluate documents. Another well-established evaluation series is the one organised by the INitiative for the Evaluation of XML retrieval (INEX) for evaluating XML retrieval methods and systems with the first run in 2002 (Fuhr et al., 2002). In the database community, the Wisconsin Benchmark (Bitton et al., 1983) was the first attempt to compare database systems followed by others developed for evaluating specific features (Bitton et al., 1985; Stonebraker et al., 1993; Cattell, 1994).

Ontologies were at the forefront of early Semantic Web research and as their number increased, the need for ontology matching evaluations became apparent. The *Ontology Alignment Evaluation Initiative (OAEI)* was founded in 2005 after merging two separate evaluation events[4] and has been driving progress in the area ever since. The evaluation of RDF stores also started around the same time (e.g., Lehigh University Benchmark (Guo et al., 2005), Berlin SPARQL Benchmark (Bizer and Schultz, 2009), SP2Bench (Schmidt et al., 2008)).

Until recently, attempts to evaluate Semantic Search technologies were limited to isolated evaluations of newly developed approaches (Bhagdev et al., 2008; López et al., 2005) or comparing the usability of different interfaces (Kaufmann, 2007). Due to the lack of standardised evaluation approaches and measures, researchers applied a diverse set of both datasets and tasks and usually refrained from comparing their tools with other similar ones in the domain. Fortunately, the community recognised this lack of, and subsequent need for, a comprehensive evaluation to foster research and development. Recently, several evaluation series were initiated, namely the SEALS semantic search evaluations (Wrigley et al., 2010c, 2011), the SemSearch challenge (Halpin et al., 2010), the QALD open challenge (Unger et al., 2011), the TREC Entity List Completion task (Balog et al., 2010c, 2011) and finally, the INEX Linked Data Track. The remainder of this section describes each of these evaluation initiatives.

5.1. Semantic Evaluation at Large Scale (SEALS) - Search Theme

According to the organisers of the SEALS semantic search evaluations, the group of tools considered are user-centred tools (i.e., are intended to interact with people not computers) for retrieving information and knowledge (Wrigley et al., 2010c, 2011). This excludes tools which require structured queries as input as well as document retrieval systems which return results as Semantic Web documents relevant to the given query.

The methodology adopted in running the two evaluation campaigns – which took place in 2010 and 2012 – consisted of two phases: an *Automated Phase* and a *User-in-the-loop Phase* (uitl). These phases allowed tools to be evaluated in terms of both performance as well as usability and user satisfaction. Beside the performance, another criterion assessed in the automated phase was scalability: the ability of tools to scale over large datasets. In order to assess scalability, one or more datasets of different sizes were required. Therefore, the EvoOnt[5] software-engineering dataset was chosen by the organisers for this phase.

Additionally, queries used in this phase were based on templates created after conducting experiments with professional programmers for the identification of standard and useful software engineering questions that programmers tend to ask when evolving a code base (de Alwis and Murphy, 2008; Sillito et al., 2006, 2008). The 50 queries were generated to include ones with varying levels of complexity: simple ones such as *'Which **methods** have the **declared return class** x?'* and more complex ones such as *'Give me all the **issues** that were reported by the **user** x and have the **state** fixed.'*[6]. The groundtruth for the final set of queries were generated by running the SPARQL query equivalent to the NL one on the dataset. Similar to IR evaluations, precision, recall and F-measure were computed, as well as other performance measures, such as the execution time (speed), CPU load and amount of memory required.

In the user-in-the-loop phase, the geography dataset from the Mooney NL Learning Data[7] was selected since the domain is sufficiently simple and understandable for non-expert end-users. NL questions for this dataset were already available and therefore used as templates to generate queries for the evaluation for which the groundtruth was also available with the question-set. Again, questions were chosen to range from simple to complex ones as well as to test tools' ability in supporting specific features such as comparison or negation. Simple questions included ones, such as: *'Give me all the **capitals** of the* USA*?'*; more complex questions included ones such as *'What are the **cities** in **states** through which the* Mississippi *runs?'*; and finally a question such as *'Which **lakes** are in the **state** with the highest point?'*.

Two usability experiments were conducted in a laboratory setting in which the recruited subjects were given a number of questions to solve with one or more tools. The first evaluation campaign (2010) used a between-subjects design in which each participating tool was evaluated with 10 different users. Although this experiment yielded a useful set of findings and recommendations for the community, it did not allow direct comparison of the different tools and their employed query approaches. This was addressed in the second evaluation campaign (2012), where tools were evaluated using a within-

[4] The Information Interpretation and Integration Conference (I^3CON) (`http://www.atl.external.lmco.com/projects/ontology/i3con.html`) and the EON Ontology Alignment Contest (`http://oaei.ontologymatching.org/2004/Contest/`)

[5] `https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/evoont/index.html`

[6] Concepts are shown in bold

[7] `http://www.cs.utexas.edu/users/ml/nldata.html`

subjects design setting with 10 casual users and 10 expert users. Here, the subjects evaluated the five tools in randomised order to avoid any learning, tiredness or frustration effects that could influence the experiment results.

For each tool, subjects were given a short training session explaining how to use the tool to formulate queries and become familiar with the evaluation control software (used for issuing new questions and collecting feedback). Following the training period, subjects were asked to formulate each question using the tool's interface. The order of the questions was randomised to avoid any learning effects. In both evaluations, the objective data collected included: 1) input time required by users to formulate their queries, 2) number of attempts, and 3) answer found rate capturing the distinction between finding the appropriate answer and the user 'giving up' after a number of attempts. Additionally, subjective data was collected using the 'think-aloud strategy' (Ericsson and Simon, 1993) and two post-search questionnaires to capture users experience and satisfaction. In the second evaluation, to allow direct comparison, users were asked to explicitly rank the tools according to certain criteria such as how much they liked the tools or how much they found the results to be informative and sufficient.

5.2. SemSearch

Pound et al. (2010) explain how object-retrieval can be seen on the Web of Data as the counterpart of document retrieval on the Web of Documents, since the first is about resources that represent objects and the latter is about resources that represent documents. They define *object retrieval* as 'the retrieval of objects in response to user formulated keyword queries' and present a classification for these queries according to the primary intent of each query. The most common type is the *entity query* which requires information about a specific instance on the Web of Data (e.g., 'IBM'), followed by the *type query* which asks for instances of a given type (e.g., 'films for Julia Roberts'). The least common types are the *attribute query* which requires information about an attribute of a type or an entity; and the *relation query* in which information about a specific relation between types or entities is requested. If the query does not fall in any of the previous types, it is then classified as *other keyword query*.

Recognising object retrieval as an integral part of semantic search and following the methodology defined by Pound et al. (2010), Halpin et al. (2010) organised a challenge which ran twice within the SemSearch workshops (in 2010[8] and 2011[9]) with a focus on *ad-hoc object retrieval*. The input query is given as a set of keywords and the expected output is a ranked list of object URIs retrieved from an RDF dataset. The requirements for the dataset were: 1) to contain and thus represent real data found on the Semantic Web, 2) to be of large yet manageable size and 3) not biased towards one particular semantic search system. Therefore, the 'Billion Triples Challenge 2009 (BTC-2009)' dataset was chosen. It contained 1.4B triples with data about 114 million objects, crawled by multiple semantic search engines: FalconS (Cheng et al., 2008), Sindice (Tummarello et al., 2007), Swoogle (Ding et al., 2004), SWSE (Harth et al., 2007), and Watson (d'Aquin et al., 2007), during February/March 2009.

According to the previous query classification, the 2010 evaluation focused on *entity queries*, while *type queries* were added in the 2011 evaluation. As with the dataset selection, the query requirements were: 1) to represent real-world queries given by actual users, and 2) be unbiased towards one specific semantic search system. To conform with these requirements, the organisers decided to use queries from logs of traditional search engines as opposed to ones from semantic search engines. They argued that the latter – largely centred around *testing and research* – did not represent real information needs of 'casual users' (at least not at the time of the evaluations). In both evaluations, the *entity queries* used were selected from the logs of Yahoo! Search and Microsoft Live Search and included ones such as 'Scott County', *'american embassy nairobi'*, *'ben franklin'* and *'carolina'*. This was, however, different for the *type queries* which were 'hand-written' by the organising committee and included ones such as *'Apollo astronauts who walked on the Moon'*, *'movies starring rafael rosell'*, and *'wonders of the ancient world'*.

Only the first 10 results per query were considered, each of which was assessed on a 4-point scale of relevance (0 being not relevant to 3 being a perfect match). The assessment was carried out by human judges using Amazon's Mechanical Turk crowd-sourcing platform[10]. The measures used to evaluate and compare the performance of the participating tools were the normalised Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), and Precision at rank k (P@k). After the 2010 evaluation, it was observed that most of the participating systems used IR-based techniques and made little use of the semantic data for more advanced reasoning and retrieval. Additionally, the systems did not try to use semantic-based techniques for understanding or expanding the queries.

5.3. Question Answering Over Linked Data (QALD)

In contrast to keyword-based search interfaces, question answering systems allow users to express more complex information needs. The Semantic Web community has significant research related to developing natural language based interfaces for closed domain RDF data. However, there has been little progress in scaling these approaches to deal with linked data with its heterogeneous, noisy, distributed and open nature. With the aim of advancing this topic and facilitating the evaluation of such approaches, the QALD open challenge focused on evaluating question-answering systems that help users find answers for their information needs in semantically annotated data using a natural language interface.

The challenge has run three times (in 2011[11], 2012[12] and

[8] http://km.aifb.kit.edu/ws/semsearch10/

[9] http://km.aifb.kit.edu/ws/semsearch11

[10] https://www.mturk.com/mturk/welcome

[11] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=challenge&q=1

[12] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=challenge&q=2

2013[13]) with two different tasks: an open/cross-domain task and a closed-domain one. It is interesting to note the need raised by the organisers for facilitating multilingual access to semantic data by changing the third challenge (2013) to be on multilingual question answering over linked data.

DBpedia was chosen for the first task since it allowed the participating systems to be tested over large datasets commonly found in the Semantic Web[14]. Additionally, since DBpedia is an extraction of structured data from Wikipedia, it allows testing the systems' ability to deal with noisy data featuring various levels of quality. Moreover, it contains data spanning multiple domains and thus conforms with the heterogeneity requirement. DBpedia 3.6, 3.7 and 3.8 were used in the 2011, 2012 and 2013 evaluations, respectively. The first contained around 670 million triple, the second consisted of 1 billion triples, while the third consisted of 1.89 billion triples. Additionally, in the third evaluation (2013) and to facilitate the multilingual task, DBpedia 3.8 was provided with multilingual labels, in addition to the Spanish DBpedia[15].

In the closed-domain task, an RDF export of MusicBrainz[16] – an open music encyclopedia that collects music metadata and makes it available to the public – was used. Its RDF export contains a small ontology describing the music domain and comprises only a few classes and relations as opposed to DBpedia whose ontology contains more than 320 classes. There are approximately 25 million triples describing artists, albums, and tracks, as well as a subset of important relations between them. Both datasets were selected to represent data found in the Semantic Web since they have been widely used in the research community and largely interlinked with other datasets in the LOD cloud[17].

In the three evaluations, a training set of natural language questions together with their equivalent SPARQL queries and correct answers were provided for each task prior to the challenge. Another set of questions were used to evaluate and compare the participating systems with respect to precision and recall. For the multilingual task in the third evaluation (2013), all questions were provided in six languages: English, Spanish, German, Italian, French, and Dutch.

The training and test questions were written by students who explored the dataset in an attempt to simulate real users' information needs. Also, the questions were not biased towards one specific approach. For instance, questions used in the open-domain task included ones such as *'which river does the Brooklyn Bridge cross?'* and *'give me all cities in New Jersey with more than 100,000 inhabitants'*. The closed-domain task included questions such as *'give me all soundtracks composed by John Williams'* and *'which bands released more than 100 singles?'*. Finally, it is not clear how the groundtruth for the questions were generated but it is highly possible that they were manually produced by the evaluation organisers.

5.4. TREC Entity List Completion (ELC) Task

Similar to SemSearch, the importance of entity-oriented search to address different information needs concerning entities on the Web was recognised by the TREC community. Balog et al. (2010b) categorises three different tasks of entity-oriented search as follows:

1. *entity ranking*: given a query and target category, return a ranked list of relevant entities.
2. *list completion*: given a query and example entities, return similar entities.
3. *related entity finding*: given a source entity, a relation and a target type, identify target entities that exhibit the specified relation with the source entity and that satisfy the target type constraint.

The second task (*list completion*) is the focus of the *Entity List Completion (ELC)* task found in TREC Entity Track. It is similar to the SemSearch list-search task (using *type queries*) in that both limit their queries to ones that require instances of a specific type. However, the ELC task is more specific since each query requests instances of a specific type that are related to a given entity with a given relation.

Again, the BTC-2009 dataset used in SemSearch was used in the first year of running the ELC task. In the second year the organisers used Sindice-2011, a more entity-oriented dataset which is *especially designed for supporting research in the domain of web entity retrieval* (Campinas et al., 2011). Queries were selected from the REF-2009 topics[18] according to their suitability to the task and the dataset: for example, having information about the query entities in the dataset. Additionally, each query included a considerable amount of information for the participating groups, such as the URI of the given entity on the Web of Data, a DBpedia class representing the target entity type as well as URIs for examples of the expected instances. A query example is given below:

```
<query>
<num>7</num>
<entity_name>Boeing 747</entity_name>
<entity_URL>clueweb09-en0005-75-02292</entity_URL>
<target_entity>organisation</target_entity>
<narrative>Airlines that currently use Boeing 747 planes.
</narrative>
<entity_URIs>
<URI>http://dbpedia.org/resource/Boeing_747</URI>
</entity_URIs>
<target_type_dbpedia>dbpedia-owl:Airline</target_type_dbpedia>
<examples>
<entity>
<URI>http://dbpedia.org/resource/Northwest_Airlines</URI>
<URI>http://www.daml.org/2002/08/nasdaq/nasdaq#NWAC</URI>
</entity>
<entity>
<URI>http://dbpedia.org/resource/British_Airways</URI>
```

[13] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home&q=3

[14] For the purposes of this paper, we refer to DBpedia as being open-domain but note that since not all subjects are represented in Wikipedia, it could instead be termed multi-domain.

[15] http://es.dbpedia.org/

[16] http://musicbrainz.org/

[17] http://lod-cloud.net/

[18] http://ilps.science.uva.nl/trec-entity/guidelines/guidelines-2009/

Table 1: Semantic Search Evaluations

Evaluation Name	Dataset	# of Queries	Query Type	# of Participants
SEALS-1 automated	EvoOnt	50	Artificial	4
SEALS-1 uitl	Mooney-Geography	20	Artificial	4
SEALS-2 automated	EvoOnt	50	Artificial	5
SEALS-2 uitl	Mooney-Geography	5	Artificial	5
SemSearch-1	BTC 2009	92	Real-world	6
SemSearch-2	BTC 2009	50 (entity), 50 (list)	Real-world	5
QALD-1 closed-domain	MusicBrainz	50	Artificial	2
QALD-1 open-domain	DBpedia 3.6	50	Artificial	2
QALD-2 closed-domain	MusicBrainz	55	Artificial	0
QALD-2 open-domain	DBpedia 3.7	100	Artificial	4
QALD-3 closed-domain	MusicBrainz	100	Artificial	1
QALD-3 open-domain	DBpedia 3.8	100	Artificial	6
TREC ELC-1	BTC 2009	8	Artificial	3
TREC ELC-2	Sindice-2011	50	Artificial	7
INEX ad-hoc 2012	Wikipedia, DBpedia, YAGO2	140	Artificial	7
INEX faceted 2012	Wikipedia, DBpedia, YAGO2	23	Artificial	7
INEX jeopardy 2012	Wikipedia, DBpedia, YAGO2	50	Artificial	7
INEX ad-hoc 2013	Wikipedia, DBpedia, YAGO2	72	Artificial	3
INEX jeopardy 2013	Wikipedia, DBpedia, YAGO2	77	Artificial	1

```
<URI>http://twitter.com/British_Airways</URI>
</entity>
...
</examples>
</query>
```

Types found in the topics (e.g., airlines) were mapped to the most specific class within the DBpedia ontology (e.g., dbpedia-owl:Airline). Results returned by participating systems were requested to be in the form of a ranked list of URIs which were assessed on a binary relevance scale (relevant or irrelevant). Only the first 100 results were considered and judged by the staff at NIST and by the track organisers. The evaluation measures used in the first run (pilot) were the Mean Average Precision (MAP) and R-Precision. For the second run, the normalised Discounted Cumulative Gain (NDCG) was the main measure used.

5.5. INEX Linked Data Track

INEX started in 2002 with the goal of providing the means for evaluating XML retrieval methods and systems. The function of these systems is more similar to that of semantic search tools, than to traditional information retrieval systems, since both aim at retrieving more focused content (answers) by exploiting structured information. In 2012, a new track started with the aim of investigating whether the combination of Linked Data together with textual data could lead to an improvement in the effectiveness of ad-hoc retrieval. Therefore, the evaluation test data included XML-ified Wikipedia articles enriched with RDF properties from DBpedia and YAGO2 in addition to encouraging the participating systems to use the RDF dumps of these datasets. Information about each article thus included XML-ified infobox attributes, links to the article's entity in DBpedia and YAGO2 and finally, RDF triples from both datasets containing the entity as either their subject or object.

The track has run in 2012[19] with three difference tasks: 1) faceted search; 2) classic ad-hoc retrieval and 3) jeopardy task, with only the last two being included in 2013[20].

The *classic ad-hoc retrieval* is the traditional task of returning a ranked list of Wikipedia articles (or DBpedia entities) relevant for a keyword query. Some of the queries provided in this task were randomly created by the organizers while the rest were selected from those of INEX 2009 and 2010 Ad-hoc Tracks. Examples of these queries are: 'Vietnam', 'social network', 'guitar' and 'Normandy'. The second task: *faceted search*, is more focused on assessing query refinement techniques and systems than typical search tasks. Here, the evaluated systems are required to return a list of facet-values to help the user in refining a given keyword query. The user can choose a suggested value and the system generates a new list, an iterative process which continues till the user finds a satisfying value. Five general keyword queries were created by the organizers, namely: 'Vietnam', 'guitar', 'tango', 'bicycle', and 'music', in addition to subtopics for these keywords which were generated using Google's online suggestions (e.g.: 'Vietnam war'). Finally, the *jeopardy task* is more similar to question answering tasks such as those provided in the QALD evaluations. An example of queries provided in this task is: 'Niagara Falls source lake'.

For both the ad-hoc and jeopardy tasks, systems were requested to return – for each query – a ranked list of entities

[19] https://inex.mmci.uni-saarland.de/tracks/lod/2012/
[20] https://inex.mmci.uni-saarland.de/tracks/lod/2013/

with a maximum of 1,000 results. For faceted-search, they were requested to return a hierarchy of recommended facet-values. Additionally, pooling was performed in all tasks and the top k results (100 for the jeopardy and 200 for the faceted search and ad-hoc tasks) from the submitted runs were selected for relevance assessment. Similar to SemSearch, assessment was carried out using Amazon Mechanical Turk. Finally, the measures used to evaluate and compare the performance of the participating systems in both the ad-hoc and jeopardy tasks were the mean reciprocal rank, MAP, P@5, P@10, P@20 and P@30, while for the faceted search, it was decided to use different metrics to gain a better understanding of the problem. The first metric is the NDCG of facet-values which measures the relevance of the returned hierarchy of recommended facet-values based on the relevance of the data covered by the facet-values, as measured by NDCG. The second metric is the interaction cost, which is more similar to those used in user-oriented evaluations, since it measures the amount of effort required by a user to find the answer for their information needs.

6. Analysis

The evaluations discussed previously exhibited low numbers of participants (Table 1) when compared to the number of participants in other IR evaluation exercises, such as the Web track at TREC. In addition, some of the approaches described in systems participating in the evaluations are limited and make little use of Semantic Web techniques. For example, as Halpin et al. (2010) note, the participating systems made limited use of the structure and semantics in the data and created limited indexes considering only the subjects of the triples. One might argue that this is due to the limited time available or being the first run in an evaluation series which can affect both the number of participants as well as the quality of submissions.

In this section, we analyse the five reviewed evaluations with respect to the core aspects necessary to conduct an effective evaluation (see Section 2): datasets, queries, relevance and judgements, measures, and user-based / interactive evaluations.

6.1. Datasets

6.1.1. Origin

As discussed in Section 3.2, evaluation datasets are either specially-created or operationally-derived. Of all the datasets used (see Table 2) in the evaluations described in Section 5, two were specially-created: EvoOnt, and Mooney. The EvoOnt dataset was chosen for the automated phase in SEALS since it provided the ability to assess the *scalability* criterion by creating datasets of various sizes given the same ontology. The Mooney dataset was chosen for the SEALS user-in-the-loop phase since it described a simple and common domain which allowed easily understandable questions.

The core benefit of using a bespoke dataset is the control it allows over certain features of the dataset (see Section 3.2) and, as a result, other aspects of the evaluation: the task to be evaluated (e.g., entity-retrieval), the type of the evaluation (e.g., usability), or the assessed criteria (e.g., scalability). However, this level of control comes at the cost of representativeness:

- In principle, the evaluated systems are going to be used in the real world and thus should be assessed for their ability to work with real data.
- It is difficult to simulate some of the aspects found in real data, such as the various levels of quality or the noise and errors typically found in it.
- A specially-created dataset is usually designed with specific characteristics or to test specific features defined by the evaluation organisers or by domain experts. However, there is no real guarantee that these are the right/appropriate characteristics of real data, which again raises the question of representativeness and realism.

It is important to note that even *operationally-derived* datasets may be subject to some degree of 'cleansing' before release thus reducing their representativeness. For instance, Sindice-2011 which was used in the TREC ELC task evaluation provided an entity-oriented dataset, specifically designed for supporting research in web entity search. It is based on real data crawled by Sindice, however, was not provided *as-is*, but was processed and transformed into a corpus of entities. To illustrate, the authors mention that "*we filter all entity graphs that are composed of only one or two triples. In general, such entity graphs do not bear any valuable information, and can be removed in order to reduce the noise as well as the size of the dataset*" (Campinas et al., 2011, p. 28). Furthermore, DBpedia is not as diverse or noisy as the BTC-2009 dataset which is based on data crawled by Falcon-S, Sindice, Swoogle, SWSE and Watson.

If possible, datasets ought to be *operationally-derived*. Despite the choice of EvoOnt by the SEALS initiative being due to the apparent ease of creating multiple datasets of differing sizes (Bernstein et al., 2009), one could imagine using the same tools to create test data of various sizes for the same ontology but for an operationally-derived dataset (e.g., DBpedia). The ability to create such datasets (common ontology, varying size) which are free from inconsistencies would be beneficial to the community and ought to be investigated. Similarly, *operationally-derived* datasets covering easily understandable domains — c.f., SEALS' choice of Mooney for their usability study (Bernstein et al., 2009) — are available: Geonames[21] in the geography domain, MusicBrainz[22] in the music domain or DBpedia in both, as well as other domains.

6.1.2. Domain

Of the datasets shown in Table 2, an equal split between *closed-domain* and *open-domain* can be observed. While Mooney, EvoOnt and MusicBrainz are closed-domain datasets (describing a single domain such as Geography), DBpedia, BTC-2009 and Sindice-2011 are heterogeneous datasets spanning multiple domains. Indeed, *open-domain* data can be argued to be increasingly important (at the expense of *closed-domain* data); the size of (open) linked data is continuously increasing and offers significant potential for answering various

[21] http://www.geonames.org/ontology/documentation.html
[22] http://musicbrainz.org/

Table 2: Properties/features of the datasets used in the reviewed evaluations.

Dataset	Type/Nature	Domain	Size (triples)	Creation Year
Mooney	Specially-created	Closed: Geography	5700	2001
EvoOnt	Specially-created	Closed: Software Engineering	Not fixed	2007
DBpedia	Operationally-derived	Heterogenous	1 billion	2009
MusicBrainz	Operationally-derived	Closed: Music	25 million	2007
BTC-2009	Operationally-derived	Heterogeneous	1 billion	2009
Sindice-2011	Specially-created	Heterogeneous	11 billion	2011

information needs. In response, semantic search development has begun to focus on search of the open web as opposed to traditional, closed-domain approaches. However, with the proliferation of heterogeneous datasets, more care than ever must be taken when choosing which datasets are selected to evaluate systems — the datasets must be applicable to the system and task and representative of the types of data for which the tool is designed.

6.1.3. Size

Dataset size has a strong influence on the evaluation criteria and the subsequent reliability of the evaluation and its results. In IR, the definition of a *large* dataset (i.e., sufficient for running realistic evaluations) has not been fixed and is continuously growing to reflect the increasing amount of data commonly available to organisations. For instance, when Spärck Jones and van Rijsbergen described *ideal test collections* in the 1970s, they were referring to datasets containing around 30,000 documents (Jones and Bates, 1977); current TREC test collections can contain a billion documents[23]. Open-domain datasets used in semantic search evaluations should reflect the growth of the Semantic Web and linked data; with current datasets reaching billions of triples, the Sindice-2011 dataset comprising 11 billion triple could be considered to be more suitable for evaluating scalability than EvoOnt's 10 million triples.

Similarly, closed-domain evaluations should favour larger datasets. For instance, Geonames' 150 million triples is more representative than Mooney's 5000 triples. The argument for selecting Mooney for the SEALS evaluation's usability phase was its easily understandable domain. However, one could equally argue that this affects the reliability of the usability experiment since it should assess the user experience in a real-world scenario; given larger datasets covering similarly understandable domains (to non-experts in the case of SEALS), compelling arguments must be made for selecting a small dataset instead.

6.1.4. Age

The creation date of a dataset could also affect its suitability for a realistic and reliable evaluation. For instance, the BTC-2009 dataset was crawled during February and March 2009; naturally, this snapshot does not include subsequent updates (e.g., DBpedia's September 2011 improvements to the schema, data and mappings). Given the speed with which the Semantic Web and linked data is evolving, preference should be given to newer datasets or datasets which receive regular updates.

6.2. Queries

6.2.1. Real Versus Artificial

Queries used in IR and similarly in semantic search evaluations are either *real queries* describing genuine information needs or *artificial queries* created to simulate the first. As shown in Table 1, SEALS, QALD, TREC and INEX adopted the second approach in which queries are created either by domain experts, volunteers (usually students) or the evaluation organisers. In the SEALS user-in-the-loop phase, queries associated with the Mooney dataset were collected from university students as well as real users (*volunteers*) while in the automated phase, queries were based on templates gathered from professional programmers (*domain experts*). In QALD, queries for both the closed- and open-domain tasks were written by university students who explored the MusicBrainz and DBpedia datasets; for TREC, topics from the traditional document-retrieval REF task were adapted to the ELC task and for INEX, some of the topics were gathered from participants while others were randomly created by the organizers.

There are advantages and disadvantages for the use of *artificial queries* over *real queries*. For example, gathering queries for a software engineering dataset from professional programmers is perhaps the best way to simulate real information needs since they are actual users in this domain. Similarly, university students could be seen as potential/real users for the music domain. However, although artificial queries allow for increased control and flexibility over the query content and features, the fact that they remain 'artificial' means they do not fully reflect real information needs. For instance, it was commonly reported in the SEALS usability experiments that the questions used were too difficult and would not be typed into a search engine by real users. An example is the question "*Which states have an area greater than 1 million and border states that border states with a mountain and a river?*". The argument in favour of these complex questions is again related to the evaluation task and expectations. To fully exercise the advanced semantic search systems, there ought to be tasks to address all state-of-the-art techniques and/or user requirements which are updated to reflect advances in the field. Thus, the real challenge is to produce a set of questions that can test this ability and other features of the systems while being more natural and representative of real information needs.

[23] `http://plg.uwaterloo.ca/~trecweb/2012.html`

The approach of adapting topics from a traditional document-retrieval task (adopted in TREC ELC) was criticised by Pound et al. (2010) who opted, in the SemSearch challenge, to use *real queries* from query logs of *traditional* search engines. They argued the latter provided a better representation of real users' information needs than those from Semantic Web search engines. Additionally, such queries would not be biased towards any particular semantic search engine. However, using logs of queries searching for documents on the Web for evaluating tools searching for knowledge on the Semantic Web is not an optimal choice. Queries posed to traditional search engines, such as Google and Yahoo!, are for different tasks and use cases and thus have different characteristics. Web users are not expecting actual answers to their queries, instead they know they'll obtain a list of documents that they can use as a starting point to navigate to other relevant documents that might contain answers to their information needs or investigate the top ones to extract the necessary answers. The difference in the nature of the task and format of results expected can change the way users formulate their queries; they are continuously learning to adapt their queries to the nature and capabilities of these search engines. For example, they learn how to overcome the limitation of handling complex queries that need integration of information by issuing multiple subqueries, investigating the results separately and manually forming the answers they are looking for. Therefore, although sampling the query logs of traditional search engines provides 'real' queries, each individual query does not necessarily capture the full complexity or richness of the original information need.

6.2.2. Query Set Size

A critical factor in the logistics of the evaluation execution, analysis of the results, and ensuring representativeness and coverage is the number of queries used. Table 1 shows that most of the reviewed evaluations used approximately similar number of queries (between 23 and 140 with 50 being most frequent). The first exception is regarding the number of questions used in the user-in-the-loop phase in SEALS in the two evaluation runs: 20 and 5 respectively. Indeed, for a usability study with real users, the number of questions should be carefully chosen to have reliable results without overwhelming the users. After the first evaluation run of SEALS, it was suggested that the use of 20 questions was too many and that, to keep the experiment within a reasonable duration and also to avoid subjects' tiredness or frustration, the number of questions ought to be reduced. The other exception is with the number of queries used in TREC ELC-1 which is only eight. This was due to the adaptation of the queries from the REF task (see Section 5.4) to the ELC task which proved to be problematic: queries were excluded since the dataset did not contain relevant entities for answering them, and thus only eight queries could be adapted.

Again, for their proposed *ideal test collection*, Spärck Jones and van Rijsbergen suggested that the acceptable number should be around 250 queries while 1000 queries might be needed in some scenarios and that it was not useful to have less than 75 queries (Jones and Bates, 1977). However, this was never achieved in IR even with the huge increase in the size of document collection within earlier (e.g., Cranfield) and current evaluations (e.g., TREC). The common number of queries/topics currently used in TREC evaluations is 50[24]. This is due to the difficulty and incurred costs in obtaining a large number of topics, especially since relevance judgements are required for each topic which are very costly to produce (see Section 3.4).

The number of queries to be used should be carefully considered within the evaluation design: it directly affects the stability of the evaluation measures and in turn the reliability of the evaluation results. For instance, Buckley and Voorhees (2000) confirmed that results of evaluations using more topics are more reliable than those from evaluations using fewer topics. However, the exact number required differs according to the evaluation measure selected since they differ with respect to their stability. The authors showed that *precision at 10 (P@10)* has more than twice the error rate associated with it than the error rate associated with the *average precision*. In a later study, Voorhees and Buckley (2002) found a strong, consistent relationship between the error rate, the size of the difference in scores, and the number of topics used. To conclude, this discussion suggests that while it might be more practical and pragmatic to use a small numbers of queries, it is essential to select the appropriate combination of evaluation measures and differences in scores (to differentiate between 2 or more methods) that corresponds well to the number of queries selected in order to achieve reliable results.

6.3. Relevance judgements

As shown in Table 3, a binary scale was used in SEALS, QALD and TREC evaluations while a three-point graded scale was used in the SemSearch evaluation. As explained in Section 3.4, a binary scale is required for the use of precision and recall, the evaluation measures employed by SEALS and QALD. Alternatively, SemSearch organisers opted to use a three-point scale: *excellent* (result describes the query target specifically and exclusively), *not bad* (result mostly about the target) and *poor* (result not about the target, or mentions it only in passing). They found that, especially for expert judges, there was almost no difference between the two- and three-point scales and concluded that "*...there is thus no marked difference between a three-point scale and a binary scale*" (Halpin et al., 2010).

As described earlier, the judgment process for the relevance of documents with respect to a given query is performed either automatically using a predefined groundtruth or manually by human judges. In the former, the groundtruth generation for each query is generally performed either by human judges scanning the entire document collection (or merely a sample), or by merging the results returned by different retrieval systems.

Neither approach was used in QALD and SEALS. Instead, it was generated by executing a SPARQL query equivalent to the NL query against the dataset and using the results as the

[24] `http://plg.uwaterloo.ca/~trecweb/2012.html`

Table 3: Semantic Search Evaluation Measures

Evaluation Name	# Results evaluated	Relevancy scale	judgements	Measures
SEALS automated	All	binary	mechanised	Precision, Recall, F-Measure, Execution Time
SEALS uitl	All	binary	mechanised	Input Time, No. of attempts, No. of queries answered, SUS score, Extended score
SemSearch	10	three-point	manually (amazon mechanical turk)	MAP, P@k, NDCG
QALD	All	binary	mechanised	Precision, Recall
TREC ELC	100	binary	manually (track organisers)	MAP, R-Precision, NDCG
INEX 2012	100 (ad-hoc and faceted search) 200 (jeopardy)	graded	manually (amazon mechanical turk)	MRR, MAP, P@5, P@10, P@20, P@30 (ad-hoc and jeopardy) NDCG of facet-values, Interaction cost (faceted search)
INEX 2013	100 (ad-hoc) 10–20 (jeopardy)	graded	manually (amazon mechanical turk)	Precision, Recall, Average-Precision, MAP, MRR, NDCG

groundtruth. A disadvantage of this approach is that the transformation from natural language to SPARQL is a non-trivial task, and errors can be introduced or suboptimal SPARQL queries could be created. This is partly due to issues such as the large number of properties that can be found in the same dataset which refer to the same real-world relationship. This problem could, for instance, result in a SPARQL query that uses only a subset of these properties and misses some relevant results leading to an incomplete result set, thus, affecting precision and recall.

TREC, INEX and SemSearch used relevance judgements created by human judges. This approach is known to have a high overhead which increases in proportion to the number of results to be considered. Therefore, only the first 10 results are evaluated in SemSearch while this number increases to 100 in TREC and 100 to 200 in INEX (see Table 3). Pound et al. (2010) argue that having more queries with fewer results evaluated (SemSearch: 92 queries and 10 results evaluated) is more desirable for web search engine evaluation than having few queries with more results evaluated (TREC: 8 queries and 100 results evaluated) especially when knowing that web users tend to examine only the top few results. Further work is required to establish an optimal tradeoff between these two factors to ensure reliable evaluations. Additionally, the other challenge facing this approach is the subjectivity of relevance judgements and the degree of inter-judge agreement that needs to be established to obtain reliable results. The judgements were performed by the track organisers in TREC and by Amazon's Mechanical Turkers in SemSearch and INEX. Although it is difficult to ensure the same level of subject knowledge for all judges, the deviation can be much more with random volunteers than with evaluation organisers. It is indeed important to understand how this factor affects the reliability of an evaluation's results since it has been acknowledged in literature that the more knowledge and familiarity the judges have with the subject area, the less leniency they have for accepting documents as relevant (Rees and Schultz, 1967; Cuadra, 1967; Katter, 1968). Interestingly, Blanco et al. (2013) analysed the impact of this factor on the reliability of the SemSearch evaluations and concluded that 1) experts are more pessimistic in their scoring and thus, accept fewer items as relevant when compared to workers (which agrees with the previous studies) and 2) crowdsourcing judgements, hence, cannot replace expert evaluations.

Relevance judgements are also affected by the amount of information provided as part of the query itself. SemSearch used keyword queries such as 'american embassy nairobi' or 'ben franklin'. In contrast, TREC ELC used 'topics' which, as well as the entity name to search for, provided supplementary information (such as the URI of the given entity on the Web of Data, a DBpedia class representing the target entity type as well as URIs for examples of the expected instances). We believe this is an important aspect in an evaluation design since the *amount of information* was found to be the highest-ranked factor affecting relevance judgements: "*such information would not only help one locate relevant results but also judge their relevance subsequently*" (Chu, 2011, p. 271).

6.4. Measures

Both SEALS and QALD used set-based measures that do not take ranking into consideration (see Table 3). This is difficult to justify (barring grounds of simplicity) since one of the key features required from (semantic) search systems is to rank results according to relevance to a given query. Users will not examine hundreds, if not millions, of results even if they were actual answers rather than documents. In contrast, SemSearch, INEX and TREC used ranked-based measures. The first used precision@10, although as shown in Section 3.5, it is known to be the least stable among the other ranked-based measures (R-Precision, MAP, NDCG). However, it was used together with MAP and NDCG which provided both an overall figure of the systems' performance for the set of queries used

(MAP) as well as their performance in retrieving highly relevant documents and ranking them (NDCG). NDCG, which is also adopted in TREC and INEX, has seen increasing use in evaluations based on non-binary scales of relevance. Despite being a well-established and reliable evaluation measure, using NDCG requires deciding on the values of the gain vector and the discount function as well as producing an ideal ranked list of the results (Voorhees, 2001; Zhou et al., 2012; Kanoulas and Aslam, 2009). These details are unclear for both initiatives[25]; our best guess for the discount function is that both used the default one: log of the document's rank (see Section 3.5.2). It is worth noting that creating the ideal result list could be an even more challenging task than deciding these values, especially with the increase in the size of the datasets used which in turn increases the difficulty of manually examining them to produce this list. Finally, R-Precision was the third measure used in TREC as a more stable and reliable measure than precision at a fixed level (P@k), which is used in SemSearch.

7. Best Practices

Significant research has been conducted in the field of IR evaluation; however, similar evaluation research efforts for semantic search technologies are still in their infancy. Therefore, the aim of this section is address this gap by suggesting a set of best practices for designing semantic search evaluations which are motivated by the IR literature reviewed above and the experience of the authors in running semantic search evaluations (Wrigley et al., 2010a; Elbedweihy et al., 2012b,a). We believe there is a need for an evaluation initiative administered by a well-respected organisation which would create and distribute datasets, organise campaigns (both system- and user-oriented) and report results to the community (c.f. TREC). The following best practices could be adopted by any such initiative. However, they are equally useful for smaller-scale exercises such as individuals or companies interested in evaluating their own systems.

As discussed in Section 4, user-oriented evaluations are very important in assessing whether a search system meets the information needs of its users. This is even more important in semantic search which ultimately seeks to improve this process by understanding the search content and the user intent to produce more relevant results. Additionally, user's experience and satisfaction with this information seeking process is influenced by many aspects, including query format, performance of the search system and presentation of the results. System-oriented evaluations focus on the second aspect: system performance. However, the other aspects are equally, if not more, important. Indeed, the query format is the starting point which can affect the whole search process since the expressiveness of the query language adopted as well as the usability of the interface and the support provided during query formulation can in practice make a great deal of difference to whether users can successfully express their queries. Therefore, we believe a hybrid approach integrating both system- and user-oriented evaluations is a necessity. Sections 7.1–7.4 cover aspects related to both evaluation scenarios, while Section 7.5 discusses aspects which are specific to user-oriented evaluations.

Note that the numbers provided in the following sections, such as dataset size or number of queries, are guidelines rather than strict values. For instance, it is necessary that the dataset size should reflect the target domain(s); therefore, the evaluation designer should balance these guidelines with the requirements of their specific task and evaluation.

7.1. Datasets

Size. The size of the dataset should be large enough to be representative of datasets currently found on the Web of Data. This trend can be observed in the IR community (e.g., currently, TREC uses corpora of up to a billion documents[26]) and the growing emphasis on 'Big Data' in general means insightful evaluations must incorporate such datasets. Examples of *single/closed-domain* datasets (ones which describe a single domain) currently found on the Web of Data are *Geonames*, *PubMed* and *LinkedGeoData* which contain around 150 million, 800 million and 3 billion triples, respectively. Therefore, for a single-domain evaluation scenario, a dataset of less than 100 million triples would be small, between 100 million and 1 billion triples is acceptable and more than 1 billion triples can be required in some cases. Additionally, examples of *multiple/open-domain* datasets (heterogeneous ones spanning various domains) are DBpedia and Sindice 2011 which contain 1 billion and 11 billion triples, respectively. Therefore, for an open-domain evaluation scenario, a dataset of less than 1 billion triples would be small, between 1 billion and 10 billion triples is acceptable and more than 10 billion can be required in some cases.

Origin. Spärck Jones and Van Rijsbergen (1976) suggested that an ideal test collection should contain documents that vary in their source and origin; we believe a dataset for semantic search evaluations should also contain data collected from different sources, including triples from the datasets in the LOD cloud as well as semantic data gathered from different domains on the Web of Data.

Quality. Datasets found on the Web of Data are known to contain a certain level of noise and erroneous data, especially the operationally-derived datasets such as DBpedia. In contrast, specially-created datasets, such as Mooney (described in Section 5.1), are usually of higher quality. Ideally, evaluations would use datasets featuring different levels of quality for assessing semantic search approaches in different scenarios. For instance, operationally derived datasets can be used to test the systems' ability to work with data found in the real-world while specially-created datasets can be used when they ought to have specific characteristics (for example, to simulate high-quality datasets found in some organisations).

[25] This is a long-standing criticism of such IR evaluations: critical details of the methods and measures adopted and the justification for them are described briefly or omitted. See Section 1

[26] `http://plg.uwaterloo.ca/~trecweb/2012.html`

Data Age. Similar to how the size of the datasets used in evaluations should be representative of datasets found in the real-world, these datasets should also be up-to-date to capture any changes in the Semantic Web standards and technologies. Outdated datasets can directly affect the reliability of evaluations.

7.2. Queries

Size. The number of queries used should be large enough to produce reliable results especially since the stability of evaluation measures is directly influenced by this number. It has been common to use 50 queries in IR evaluations[27] and similarly in the semantic search evaluations discussed in this paper (see Table 1).

Therefore, for system-based evaluations (focusing on assessing retrieval performance) which can be done in an offline mode, between 50 and 100 queries would be acceptable. In contrast, in user-based evaluations, this number is usually much smaller since it directly influences the amount of resources (time, cost, effort) required and the subjects recruited. Most of the IIR and semantic search user-based studies discussed above used between four and 20 queries. However, using a large number of queries such as 20 was found to be too many for the subjects (Wrigley et al., 2010b). Wrigley et al. explained that in the final quarter of the experiment, the subjects tended to become tired, bored and sometimes frustrated. Therefore, we believe a best practice would be to use a number of queries in the range of 5 and 15. If more queries are necessary for a specific scenario, multiple sessions could be used to alleviate the previously mentioned effects on subjects.

Origin. Queries should describe genuine user information needs. These could be collected from interviewing users, or based on analysing the query logs of different state-of-the-art semantic search systems. However, given the infancy of the field, this is challenging beacuse there are currently not enough human-focussed systems or users to provide representative query histories (let alone users who are 'non-experts' in the Semantic Web field). An alternative source are the semantic search engines, such as Sindice (Tummarello et al., 2007), and federated query architectures, such as SQUIN (Hartig et al., 2009), which are commonly used programmatically by other Semantic Web applications. However, there are three problems with using their query logs. Firstly, the tasks / use cases directly influence the characteristics of the queries (see Section 6.2) and are usually very different from queries issued by human users. Secondly, it is usually difficult to find such systems, which are actively developed and frequently used by the community. The third problem is with respect to the representation of the queries. Most of the queries issued to the first type of applications (semantic search engines) are usually given as keywords, while those issued to the second type of applications (federated query architectures) are usually written in SPARQL. On one hand, keywords can lack important information such as the relations between concepts or entities found in the queries which can affect the subjects' understanding and interpretation. On the other hand, SPARQL queries are not suitable for subjects who are not Semantic Web experts. Therefore, there is a step required to translate either of these types of queries into NL verbalised statements to be used in semantic search evaluations.

Complexity/Difficulty. Since evaluations aim to assess systems in real-world scenarios, a requirement is, therefore, to use queries of different levels of complexity (e.g., different number of concepts and properties) and comprising different features (such as negation) that are typically found in real queries. Using only simple or complex queries could affect the realism and in turn reliability and efficacy of the evaluations.

Type. Queries can be broadly categorised into *factual queries*, in which the information needed is about a particular fact, and *exploratory queries* in which the information need is more generic and does not specify a particular fact. According to the classification used in Section 4.2.4, the first type covers both *specific fact-finding* and *extended fact-finding* tasks, while the second type covers *open-ended browsing* and *exploration* tasks. An example of factual queries is "Give me all the capitals of the USA" (taken from Mooney), while an example of exploratory queries is "Provide information on the Bengal cat breed" (taken from TREC). While both types have been used in IR evaluations, semantic search evaluations have been mostly adopting factual queries. A justification for using this type of queries could be related to the current ability of semantic search approaches in coping with exploratory queries. A more probable justification is the fact that it is much easier to generate groundtruth for factual queries which allows measuring precision and recall for the evaluated approaches. Indeed, it is very challenging to generate groundtruth for exploratory queries since it is not clear what would represent a *correct* answer. In this scenario, the same approach adopted in IR can be used in which human judges are asked to evaluate a sample pooled from the top results of different retrieval systems to construct the final groundtruth. Yet again, the difficulty here would be to determine the relevance of an entity URI or a literal value, which are the types of answers returned by semantic search systems, to the given query, as noted by Pound et al. (2010). This decision can be highly subjective which would increase the number of judges required to allow a level of inter-judge agreement. Altogether, this causes the evaluation process to be resource intensive (in terms of time, effort and money). We believe this challenge should be addressed since both types of queries are found in the real-world search scenarios and represent genuine information needs.

Representation. A common approach currently used in most IR evaluations, such as TREC, is to include a verbal statement of the information need together with additional information, such as a description and a narrative (i.e., topic). Semantic search evaluations have been using only verbal statements to describe their queries. Topics provide more information which helps subjects/judges identify the relevant answers in the results. Additionally, as discussed earlier, the *simulated work*

[27] `http://sites.google.com/site/trecfedweb/` `http://plg.uwaterloo.ca/~trecweb/`

task situation which was proposed by Borlund and Ingwersen (1997) provides more realism by describing the situation that led to the information need. Indeed, the semantic search community should investigate the possibility of using these representations to increase the reliability of evaluations and their results.

It is interesting to note that the importance of the aspects presented above has been recently recognized within the community. Issues related to some of these aspects have been discussed in the JIWES workshop, first held in 2012, which focused on challenges and tasks in entity search (Balog et al., 2012). To address some of these issues, Balog and Neumayer (2013) released an entity search test collection featuring DBpedia as a dataset together with a large number of queries and relevance judgments from previous evaluation campaigns. With regards to queries, the authors tried to provide a representative set of information needs with varying length, type, origin and complexity. These were gathered from four different evaluation campaigns: NL queries from QALD-2 (Section 5.3), type and entity-queries from SemSearch 2010 and 2011 (Section 5.2), and entity-oriented queries from both TREC-2009 entity track (Balog et al., 2010a) and INEX-2009 entity ranking track (Demartini et al., 2010).

7.3. Groundtruth

Some of the semantic search evaluations discussed above generate a SPARQL query equivalent to the NL query and execute the former to produce the groundtruth. However, this might not produce very accurate results since the mapping from the NL to a SPARQL query is manually performed by the organisers and is subjective; there is not always one *right* way to do it since there can be different paths in the dataset that lead to the same result (concepts can be linked to each other through different paths). It is therefore difficult to guarantee the completeness of the groundtruth generated by following this approach. This could result in some of the evaluated systems to have recall *less than one* if they follow different paths, generate different SPARQL queries and therefore get different results, which may still be relevant to the query. We believe that a more accurate approach could be to inherit the pooling technique from IR, in which different systems are asked to submit their results and the top K results are merged and assessed by human judges to generate the groundtruth. Recently, crowdsourcing this and similar tasks has received an interest within the research community, for example, using Amazon Mechanical Turk. However, this should be further investigated since the feasibility and reliability of this approach are not yet agreed on (Kazai, 2011; Carvalho et al., 2011; Clough et al., 2012).

7.4. Evaluation Criteria and Measures

Ranking is a necessity for search. It is important to encourage adopting ranked-based measures (as opposed to set-based measures such as precision and recall). Also, it is important to distinguish between systems based on their different levels of performance in retrieving relevant answers. Graded-relevance scale and the suitable measures (e.g., NDCG) should be used (as opposed to 'relevant' and 'non-relevant' mode). Moreover, relevance assessment should be performed by human judges, with careful consideration to what affects judges' assessments, especially key aspects such as the difference in their knowledge (experts in the domain versus non-experts) or the order of presentation of the results (which can be normalised by randomising this order). Also, measures ought to be chosen while taking into account the number of queries used in the evaluation design and the number of results that will be assessed – per query – since both aspects influence the stability of the measures used.

7.5. User-based evaluation

In designing user-based evaluations of semantic search systems, the following aspects are important and require careful consideration.

Evaluation Setup. Running user-based studies is known to have a high overhead with the need to allocate resources (time, cost, effort) and recruit subjects. This is in addition to the logistics required for carefully organising and scheduling an evaluation, as well as the overhead incurred in the data analysis which is acknowledged to be extremely time consuming and labor-intensive (Kelly, 2009). This has led to researchers refraining from evaluating their own systems in a user-oriented scenario, let alone comparing them with others. Thus, having a central organisation responsible for evaluating different systems and approaches is required. It would also facilitate adopting a within-subjects design to allow comparison of results and guarantee fairness of the evaluation process since systems would be evaluated in equal time periods and by external people, which sidesteps any possible bias that could be introduced by having developers evaluating their own systems.

In addition to the requirements specified above for the choice of the evaluation dataset, our experience with user-based evaluations showed another issue to consider. We found that inconsistencies in the dataset, as well as naming techniques used within the Semantic Web community, could affect the user's experience and their ability to perform the search tasks. This in turn would affect the reliability of the results of the evaluation. For instance, users in one evaluation were confused with inverse properties (e.g., *runsthrough* and *hasRiver* found in Mooney dataset) when building queries using a view-based query approach. Similarly, a property like `dbo:isPartOf` (found in DBpedia), connecting places like regions and cities found in them, was confusing and not self-explainable for users. This introduces an additional difficulty while choosing a dataset to ensure a balance between evaluation realism (by choosing datasets representative of those found in the Semantic Web) and evaluation reliability (by trying to avoid these characteristics in the chosen datasets).

Moreover, in the choice of the evaluations subjects, it is mostly important that they well represent the population, which mainly depends on who is targeted by the evaluated systems. In the literature, users have been usually categorised as expert users and casual users (see Section 4.2.3). Many systems developed within the Semantic Web community have been evaluated with experts. This is usually due to the difficulty of finding

casual users who are able to understand and assess these systems. However, this needs careful consideration since ideally, the goal for the Semantic Web and similarly semantic search is to reach a wider population of users, not limited to the Semantic Web community. Indeed, evaluating systems with both types of users and comparing their results is beneficial and could provide an understanding of the suitability of certain search approaches to specific types of users and, furthermore, the different requirements and preferences to cater for when targeting a particular type of users. Additionally, deciding the number of subjects to recruit for a user-based evaluation is an open question and is influenced by the availability of resources (time, cost, effort) and subjects with the required characteristics. It can also affect the reliability of the evaluation results. Based on IIR and HCI literature (see Section 4.2.3) and also our experience, we believe that a number ranging between 8 and 12 subjects would be acceptable.

Finally, with respect to data collection, in addition to using individual questionnaires to assess certain aspects for each system, we found that an overall questionnaire (presented after evaluating all the approaches) asking the user to rank the systems with respect to certain aspects can produce more accurate comparisons since the rankings are an inherently relative measure. Such comparisons using the individual questionnaires may be less reliable since the questionnaire is completed after evaluating each approach's experiment (and thus temporally spaced) with no direct frame of reference to any of the other approaches.

What to evaluate. As argued in IIR literature (see Section 4.1), utility could be a more appropriate measure for evaluating IIR systems and their ability to support users in their search tasks. Assessing utility and usefulness of results as opposed to relevance would capture how the user judgment is influenced by other aspects beside the relevance of the answer to the query. Examples of these aspects are users' background and knowledge (what is already known about the query subject), the interaction between the system and the user, or the representation of the answer itself (it can be understood for instance by one user and not by another). Goffman (1964) notes that "any measure of information must depend on what is already known". Therefore, to assess utility, one could use questions with an overall goal – as opposed to ones which are not part of any overarching information need – and compare users' knowledge before and after the search task. Since the usefulness – in this scenario – of a result will be evaluated by the user, exploratory queries could be used in addition to factual queries since there is no need to worry about generating the groundtruth for the queries. Furthermore, as discussed earlier, using *simulated work tasks* is intended to develop simulated information needs by allowing for user interpretations of the situation. All of the above together would, indeed, add more realism to the evaluation and increase its reliability.

Moreover, one of the most used evaluation criteria in IIR and usability studies is efficiency (commonly assessed using time- or effort-based measures). From previous evaluations, we found that both measures should be used in order to obtain a full impression of efficiency (either measure alone provided only a partial account). For instance, the time required by users to formulate queries in a system can be low but the number of trials performed to answer a question is high. In such a situation, using both measures would provide more reliable results and comparisons. Additionally, we believe that it is important to evaluate system-level efficiency (e.g., response time) since this strongly influences user satisfaction. Despite its importance, this aspect has been omitted from previous user-based semantic search evaluations.

Furthermore, any new interface, application, or product is expected to require some amount of learning. Nielsen (1993) notes that a system that is initially hard to learn could be eventually efficient. Certainly, investigating the ease of learning how to use a system would be even more beneficial when evaluating a new or advanced approach or interface that users are not familiar with (such as the different visual approaches consuming Linked Data). This can be achieved by evaluating *extended learnability*, which refers to the change in performance over time (Grossman et al., 2009) as opposed to *initial learnability*, which refers to the initial performance with the system. This aspect, despite its importance, has been missing from user-based evaluations of semantic search systems. Studies focusing on extended learnability are usually referred to as *longitudinal* studies, and they are conducted over an extended period of time, with evaluation measures taken at fixed intervals, both of which determine the number of sessions required (Kelly, 2009). It is thus important to decide on this period of time as well as the interval between the sessions. Similarly to the choice of the number of subjects required for a usability study, the number of sessions presents a tradeoff between reliability of the evaluation (since it directly affects the amount of collected data and results), and its overhead. On the other hand, the interval between two evaluation sessions should be influenced by the expected/actual use of the evaluated system or interface. For instance, since search tools are often used everyday, the evaluation sessions should be placed over consecutive days (with the same users).

7.6. Repeatability and Reliability

Experiments should be repeatable and reproducible, as well as being carried out with scientific rigour (Vitek and Kalibera, 2011). Repeatability and reproducibility of an evaluation is concerned with whether its repetition over a period of time produces the same results and rankings of systems. A key factor in achieving this repeatability is the control over the experimental variables. Hence, the user-oriented approach to evaluations has been acknowledged to face difficulties with being repeatable. One of the main reasons is the variability introduced by human factors, such as differing search behavior and strategies, their capabilities in expressing the information need, as well as their satisfaction degrees and criteria. On the other hand, in the system-oriented approach, the main factor is the consistency of relevance assignments for a specific result.

As discussed in Section 7.3, in the SEALS and QALD evaluations, generating groundtruth for a specific query is performed by running its equivalent SPARQL query over the evaluation

dataset. Although we argued against the reliability of this approach, it guarantees the repeatability of the results. Indeed, this requires using the exact query on the same version of the dataset to avoid any changes in the resulting assessments. In contrast, TREC and SemSearch used human judges to assess the relevance of results. The repeatability here thus depends on the degree of inter-judge agreement. The difference between the two evaluations is that TREC used expert judges, whereas SemSearch used Amazon Mechanical Turk workers in the assessment process. Blanco et al. (2013) pointed out the limitation – in terms of scalability – of depending on a limited number of expert judges since, in repeating the evaluation by other researchers, it would be difficult if not impossible to use the same judges. Additionally, they showed that repeatability was successfully achieved through crowdsourced judgements since conducting the same experiment with two different pools of workers over a six-month period produced the same assessments and same rankings for the evaluated systems.

Another important aspect that could influence repeatability is the cost of conducting an evaluation. Clough et al. (2012) note that crowdsourcing potentially lowers this cost, and thus, is an advantage of using this approach. They conducted an experiment in which they showed that the cost of recruiting 73 workers on Amazon Mechanical Turk, for around 45 hours to judge the relevance of 924 results, was $43.00, while the expert judge cost was $106.02 for around 3 hours of work. Additionally, Blanco et al. (2013) showed that, using Amazon Mechanical Turk, an entire SemSearch challenge was conducted for a total cost of $347.16. In this competition, 65 workers judged a total of 1737 assignments, covering 5786 submitted results from 14 different runs of 6 semantic search systems. Blanco et al. (2013) thus considered this approach to be cost-effective. However, arguably, being "*cost-effective*" is very subjective: while it could be affordable for an organization or an evaluation campaign, it is more likely to cause difficulties for an individual researcher (e.g., a PhD student) and thus affect the repeatability criterion.

With regards to reliability, we illustrated how the approach adopted by SEALS and QALD is the most problematic since it does not guarantee the completeness of results and, in turn, the reliability of the assessments. The use of expert judges (as in TREC) is found on the other end of the spectrum as the most reliable approach. However, due to the issues with this approach described above (scalability limitation and high cost), several experiments have been recently conducted to investigate the reliability of using non-expert judges (e.g., Amazon Mechanical Turkers) as an alternative. However, a consensus on the reliability of this approach does not seem to be yet reached. On one hand, Alonso and Mizzaro (2009) showed that crowdsourcing was a reliable way of providing relevance assessments, the same conclusion of a more recent study by Carvalho et al. (2011). On the other hand, Clough et al. (2012) and Blanco et al. (2013) showed that, while crowdsourced assessments and expert-judges' assessments produce similar rankings of evaluated systems, they do not produce the same assessment scores. Blanco et al. (2013) found that, in contrast to experts who are pessimistic in their scoring, non-expert judges accept more items as relevant. Additionally, the level of inter-judge agreement was much higher for expert judges than for non-experts (0.57 versus 0.36). Despite this, they concluded that the reliability of non-expert judges is still sufficient since they provided the same overall ranking of the systems as the expert judges. We suggest, however, that the ranking of systems is not the sole goal of an evaluation: understanding and investigating the real performance of systems is equally important. Indeed, Clough et al. (2012) note that crowdsourced workers are not guaranteed to provide assessments that enable accurate measurement of the differences between two systems. Furthermore, being lenient in the assessment process and producing high results for evaluated systems could, indeed, mask the systems' true performance.

Finally, it is important to re-emphasise the need for more work towards evaluation platforms that enable persistent storage of datasets and results that guarantee their reusability and the repeatability of tests. Forming agreement on the approaches and measures to be used within the search community for evaluating similar categories of tools is a key aspect of standardised evaluation. In addition to the resources created, the value of organised evaluation campaigns in bringing together members of the research community to tackle problems collectively would help in accelerating progress in the semantic search field.

8. Conclusions and Future Work

In this paper, we have analysed existing semantic search evaluation campaigns with respect to a number of critical aspects such as the datasets and queries used; the process of the result relevance decision; and the performance measures and how they are computed. Based upon this analysis, we have discussed limitations to the approaches followed in these evaluations. We have also described additional elements currently missing from many evaluations that we believe should be addressed in current and future evaluations.

Recently, more attention is being driven to evaluating semantic search tools especially with the growth in development and research in this area. However, these efforts have largely been developed in isolation with no coherent overall design leading to slow progress and low interest especially when compared to other established evaluation series such TREC in IR. Our work is thus a first step towards identifying the adequacy and deficiencies of current evaluations as well as missing aspects in order to arrive at a more comprehensive and improved evaluation methodology and framework. This would enable more reliable and thorough assessments of semantic search tools and highlight their strengths and weaknesses which in turn drives progress and improvements in this area.

Indeed, we plan to make use of this analysis in designing and proposing an improved evaluation framework to the community. According to the recommendations presented in this paper, some important aspects to be considered are the type of queries used and their level of complexity as well as how they are generated: synthetic versus real-world. Also, the overhead and difficulty of having human judges assess the relevance of results (which is also affected by the size of the dataset, the

number of queries as well as the number of results considered) should not be underestimated. Additionally, the choice of the evaluation measures to be used is critical as it directly affects the reliability and worthiness of the evaluation results. Finally, user-centred evaluation should be considered an essential part of the assessment process since the usability of semantic search tools and users satisfaction is critical to a user-centric activity such as search.

References

Albakour, M.-D., Kruschwitz, U., Nanas, N., Kim, Y., Song, D., Fasli, M., De Roeck, A., 2011. Autoeval: an evaluation methodology for evaluating query suggestions using query logs. In: Proceedings of the 33rd European conference on Advances in information retrieval. pp. 605–610.

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

Aslam, J. A., Pavlu, V., Yilmaz, E., 2006. A statistical method for system evaluation using incomplete judgments. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '06. ACM, pp. 541–548.

Azzopardi, L., Järvelin, K., Kamps, J., Smucker, M., 2010. Report on the SIGIR 2010 workshop on the simulation of interaction. SIGIR Forum 44, 35–47.

Baeza-Yates, R., Ribeiro-Neto, B., 2011. Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley Professional, Ch. 4.

Bailey, P., Thomas, P., Craswell, N., Vries, A. P. D., Soboroff, I., Yilmaz, E., 2008. Relevance assessment: are judges exchangeable and does it matter. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 667–674.

Bainbridge, L., 1979. Verbal reports as evidence of the process operator's knowledge. International Journal of Man-Machine Studies.

Balog, K., Carmel, D., de Vries, A. P., Herzig, D. M., Mika, P., Roitman, H., Schenkel, R., Serdyukov, P., Duc, T. T., 2012. The first joint international workshop on entity-oriented and semantic search (jiwes). SIGIR Forum 46 (2), 87–94.

Balog, K., de Vries, A. P., Serdyukov, P., Thomas, P., Westerveld, T., 2010a. Overview of the TREC 2009 entity track. In: Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009). NIST.

Balog, K., Meij, E., de Rijke, M., 2010b. Entity search: Building bridges between two worlds. In: SemSearch2010: Semantic Search 2010 Workshop at WWW 2010. pp. 91–95.

Balog, K., Neumayer, R., 2013. A test collection for entity search in dbpedia. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13. ACM, pp. 737–740.

Balog, K., Serdyukov, P., de Vries, A., 2010c. Overview of the TREC 2010 Entity Track. In: TREC 2010 Working Notes. NIST.

Balog, K., Serdyukov, P., de Vries, A. P., 2011. Overview of the TREC 2011 Entity Track. In: TREC 2011 Working Notes. NIST.

Belkin, N., Vickery, A., 1985. Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-Based Systems. British Library.

Bernstein, A., Reinhard, D., Wrigley, S. N., Ciravegna, F., November 2009. SEALS Deliverable D13.1 Evaluation design and collection of test data for semantic search tools. Tech. rep., SEALS Consortium.
URL `http://about.seals-project.eu/downloads/category/1-`

Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D., 2008. Hybrid Search: Effectively Combining Keywords and Ontology-based Searches. In: Proceedings of the 5th European Semantic Web Conference (ESWC2008). pp. 554–568.

Bitton, D., DeWitt, D. J., Turbyfill, C., 1983. Benchmarking Database Systems – A Systematic Approach. In: Proceedings of the 9th International Conference on Very Large Data Bases. pp. 8–19.

Bitton et al., D., 1985. A Measure of Transaction Processing Power. Datamation.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. DBpedia - A crystallization point for the Web of Data. Journal of Web Semantics.

Bizer, C., Schultz, A., 2009. The Berlin SPARQL Benchmark. Group.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., Tran, T., 2013. Repeatable and reliable semantic search evaluation. Web Semantics 21, 14–29.

Borland, P., 2013. Interactive information retrieval: An introduction. Journal of Information Science Theory and Practice 13, 12–32.

Borlund, P., 2000. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation.

Borlund, P., Ingwersen, P., 1997. The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation.

Borlund, P., Ingwersen, P., 1998. Measures of relative relevance and ranked half-life: performance indicators for interactive ir. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 324–331.

Brooke, J., 1996. SUS: a quick and dirty usability scale. In: Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland, I. L. (Eds.), Usability Evaluation in Industry. Taylor and Francis, pp. 189–194.

Buckley, C., Voorhees, E. M., 2000. Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR2000). pp. 33–40.

Buckley, C., Voorhees, E. M., 2004. Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 25–32.

Büttcher, S., Clarke, C. L. A., Yeung, P. C. K., Soboroff, I., 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 63–70.

Campinas, S., Ceccarelli, D., Perry, T. E., Delbru, R., Balog, K., Tummarello, G., 2011. The Sindice-2011 dataset for entity-oriented search in the web of data. In: Proceedings of the 1st International Workshop on Entity-Oriented Search (EOS). pp. 26 – 32.

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., Allan, J., 2008. Evaluation over thousands of queries. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '08. pp. 651–658.

Carvalho, V. R., Lease, M., Yilmaz, E., 2011. Crowdsourcing for search evaluation. SIGIR Forum 44, 17–22.

Cattell, R. G. G., 1994. The engineering database benchmark. In: Readings in database systems (2nd ed.). Morgan Kaufmann Publishers Inc., pp. 247–281.

Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P., 2009. Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). pp. 621–630.

Cheng, G., Wu, H., Ge, W., Qu, Y., 2008. Searching semantic web objects based on class hierarchies. In: Proceedings of of Linked Data on the Web Workshop (LDOW).

Chin, J., Diehl, V., Norman, K., 1987. Development of an Instrument Measuring User Satisfaction of the Human-computer Interface. University of Maryland.

Chu, H., 2011. Factors affecting relevance judgment: a report from TREC Legal track. Journal of Documentation.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I., 2008. Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval. pp. 659–666.

Clemmer, A., Davies, S., 2011. Smeagol: A Specific-to-General Semantic Web Query Interface Paradigm for Novices. In: Proceedings of the 22nd international conference on Database and expert systems applications (DEXA2011). pp. 288–302.

Cleverdon, C. W., 1960. Report on the first stage of an investigation onto the comparative efficiency of indexing systems. Tech. rep., The College of Aeronautics, Cranfield, England.

Cleverdon, C. W., 1970. The effect of variations in relevance assessments in comparative experimental tests of index languages. Tech. rep., The College of Aeronautics, Cranfield, England.

Cleverdon, C. W., 1991. The significance of the cranfield tests on index lan-

guages. In: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '91. pp. 3–12.
Cleverdon, C. W., Mills, J., Keen, M., 1966. Factors determining the performance of indexing systems. Aslib Cranfield Research Project Cranfield England.
Clough, P., Goodale, P., 2013. Selecting success criteria: Experiences with an academic library catalogue. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (Eds.), CLEF. Springer, pp. 59–70.
Clough, P., Sanderson, M., 2013. Evaluating the performance of information retrieval systems using test collections. Information Research 18.
Clough, P., Sanderson, M., Tang, J., Gollins, T., Warner, A., 2012. Examining the limits of crowdsourcing for relevance assessment. Internet Computing, IEEE PP, 1–1.
Cooper, W. S., 1968. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. Journal of the American Society for Information Science.
Cooper, W. S., 1973a. On selecting a measure of retrieval effectiveness. Journal of the American Society for Information Science.
Cooper, W. S., 1973b. On selecting a measure of retrieval effectiveness, Part 1: The "subjective" philosophy of evaluation. Journal of the American Society for Information Science.
Cooper, W. S., 1973c. On selecting a measure of retrieval effectiveness part II. Implementation of the philosophy. Journal of the American Society for Information Science.
Cormack, G. V., Palmer, C. R., Clarke, C. L. A., 1998. Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '98. ACM, pp. 282–289.
Craswell, N., Zoeter, O., Taylor, M., Ramsey, B., 2008. An experimental comparison of click position-bias models. In: Proceedings of the international conference on Web search and web data mining. pp. 87–94.
Crawford, G. A., Lee, A., Connolly, L., Shylaja, Y., 1992. Opac user satisfaction and success: a study of four libraries. In: Proceedings of the 7th Conference on Integrated on-line Library Systems, IOLS 1992: Integrated on-line Library Systems. pp. 81–89.
Croft, B., Metzler, D., Strohman, T., 2009. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company.
Cuadra, C., 1967. Experimental Studies of Relevance Judgments. Final Report [by Carlos A. Cuadra and Others]. System Development Corporation.
Damljanovic, D., Agatonovic, M., Cunningham, H., 2010. Natural Language Interface to Ontologies: combining syntactic analysis and ontology-based lookup through the user interaction. In: Proceedings of the 7th Extended Semantic Web Conference (ESWC2010). pp. 106–120.
d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E., 2007. Characterizing knowledge on the semantic web with watson. In: EON. pp. 1–10.
Davis, F. D., Bagozzi, R. P., Warshaw, P. R., 1989. User acceptance of computer technology: a comparison of two theoretical models. Manage. Sci., 982–1003.
de Alwis, B., Murphy, G. C., 2008. Answering conceptual queries with ferret. In: Proceedings of the 30th international conference on Software engineering (ICSE 2008). pp. 21–30.
Demartini, G., Iofciu, T., De Vries, A. P., 2010. Overview of the inex 2009 entity ranking track. In: Proceedings of the Focused Retrieval and Evaluation, and 8th International Conference on Initiative for the Evaluation of XML Retrieval. INEX'09. Springer-Verlag, pp. 254–264.
Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. C., Sachs, J., 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management. pp. 652–659.
Dong, X., Su, L. T., 1997. Search engines on the world wide web and information retrieval from the internet: A review and evaluation. Online and CD-ROM Review.
Draper, S., 1996. Overall task measurement and sub-task measurements. In: Proceedings of the 2nd Mira Workshop. pp. 17–18.
Dunlop, M. D., 1986. Experiments on the Cognitive Aspects of Information Seeking and Information Retrieving. Final Report and Appendices. Research report, Rutgers, The State Univ., New Brunswick, NJ. School of Communication, Information, and Library Studies.
Dunlop, M. D., 1996. Proceedings of the second mira workshop. Research report tr-1997-2, University of Glasgow.
Eisenberg, M. B., 1988. Measuring relevance judgments. Information Processing & Management.
Elbedweihy, K., Wrigley, S. N., Ciravegna, F., 2012a. Evaluating Semantic Search Query Approaches with Expert and Casual Users. In: Proceedings of the Evaluations and Experiments Track, 11th International Semantic Web Conference (ISWC2012). pp. 274–286.
Elbedweihy, K., Wrigley, S. N., Ciravegna, F., Reinhard, D., Bernstein, A., 2012b. Evaluating semantic search systems to identify future directions of research. In: Proceedings of the 2nd International Workshop on Evaluation of Semantic Technologies (IWEST). pp. 25–36.
Ericsson, K. A., Simon, H. A., 1993. Protocol analysis: Verbal reports as data. MIT Press.
Faulkner, L., 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments and Computers.
Foskett, D. J., 1972. A note on the concept of relevance. Information Storage and Retrieval.
Fuhr, N., Gövert, N., Kazai, G., Lalmas, M., 2002. Inex: Initiative for the evaluation of xml retrieval. Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval 2006, 1–9.
Goffman, W., 1964. On relevance as a measure. Information Storage and Retrieval 2, 201–203.
Goffman, W., 1967. Communication and epidemic processes. Proceedings of the Royal Society.
Goffman, W., Newill, V. A., 1966. A methodology for test and evaluation of information retrieval systems. Information Storage and Retrieval.
Gonzalo, J., Clough, P., Vallin, A., 2006. Overview of the clef 2005 interactive track. In: Proceedings of the 6th international conference on Cross-Language Evaluation Forum: accessing Multilingual Information Repositories. pp. 251–262.
Gonzalo, J., Oard, D., 2004. iCLEF 2004 track overview: Interactive Cross-Language Question Answering. In: Results of the CLEF 2004 Evaluation Campaign. pp. 310–322.
Griffiths, J. R., Johnson, F., Hartley, R. J., 2007. User Satisfaction as a Measure of System Performance. Journal of Librarianship and Information Science.
Grossman, T., Fitzmaurice, G., Attar, R., 2009. A survey of software learnability: metrics, methodologies and guidelines. In: Proceedings of the 27th international conference on Human factors in computing systems. pp. 649–658.
Grubinger, M., Clough, P., 2007. On the creation of query topics for imageclefphoto. In: Proceedings of the third MUSCLE / ImageCLEF workshop on image and video retrieval evaluation. pp. 19–21.
Guo, Y., Pan, Z., Heflin, J., 2005. LUBM: A benchmark for OWL knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web.
Halpin, H., Herzig, D. M., Mika, P., Blanco, R., Pound, J., Thompson, H. S., Duc, T. T., 2010. Evaluating Ad-Hoc Object Retrieval. In: Proceedings of the 1st International Workshop on Evaluation of Semantic Technologies (IWEST).
Harman, D., 2011. Information Retrieval Evaluation. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.
Harper, D., 1996. User, task and domain. In: Proceedings of the 2nd Mira Workshop.
Harter, S., Hert, C., 1997. Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. Annual Review of Information Science and Technology (ARIST).
Harth, A., 2010. VisiNav: A system for visual search and navigation on web data. Journal of Web Semantics 8, 348–354.
Harth, A., Hogan, A., Delbru, R., Umbrich, J., O'Riain, S., Decker, S., 2007. Swse: Answers before links! In: Semantic Web Challenge. Vol. 295.
Hartig, O., Bizer, C., Freytag, J. C., 2009. Executing SPARQL Queries over the Web of Linked Data. In: Proceedings of the 8th International Semantic Web Conference (ISWC 2009). pp. 293–309.
Hearst, M., 2009. Search User Interfaces. Cambridge University Press.
Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.-P., 2002. Finding the flow in web site search. Commun. ACM, 42–49.
Hersh, W., Over, P., 2000. SIGIR workshop on interactive retrieval at TREC and beyond. SIGIR Forum 34.
Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., Olson, D.,

2000a. Do batch and user evaluations give the same results? In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 17–24.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., Olson, D., 2000b. Do batch and user evaluations give the same results? In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 17–24.

Hersh, W. H., Over, P., 1999. Trec-9 interactive track report. In: Proceedings Text REtrieval Conference (TREC-9. pp. 41–50.

Hersh, W. R., Crabtree, M. K., Hickam, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, P., Mosbaek, C., Kraemer, D., 2002. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. American Medical Informatics Association.

Hildreth, C., 2001. Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. Information Research: an international electronic journal.

Hitchingham, E., 1979. A Study of the Relationship between the Search Interview of the Intermediary Searcher and the Online System User, and the Assessment of Search Results as Judged by the User. Final Report. Research report, Oakland Univ., Rochester, MI. Kresge Library., project No. 475 AH 70111. Grant No. GO0 7702309) U.S. Office of Education. Department of Health. Education and Welfare.

Hoeber, O., Yang, X. D., 2007. User-oriented evaluation methods for interactive web search interfaces. In: Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops. IEEE Computer Society, pp. 239–243.

Hölscher, C., Strube, G., 2000. Web search behavior of Internet experts and newbies. Computer Networks 33, 337–346.

Hsieh-Yee, I., 2001. Research on web search behavior. Library & Information Science Research.

Huuskonen, S., Vakkari, P., 2008. Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine. Journal of Documentation.

Ingwersen, P., 1992. Information retrieval interaction. Taylor Graham.

Ingwersen, P., Järvelin, K., 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Springer.

Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Möller, T., 2013. A Systematic Review on the Practice of Evaluating Visualization. IEEE Trans. Vis. Comput. Graph. 19 (12), 2818–2827.

ISO, 1998. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. Tech. rep., International Organization for Standardization.

Janes, J. W., 1993. On the distribution of relevance judgements. In: Proceedings of the ASIS Annual Meeting. Vol. 31. pp. 104–114.

Järvelin, K., 2011. Evaluation. In: Ruthven, I., Kelly, D. (Eds.), Interactive Information Seeking, Behaviour and Retrieval. Facet Publishing, pp. 113–138.

Järvelin, K., Kekäläinen, J., 2000. Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 41–48.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst.

Jenkins, C., Corritore, C. L., Wiedenbeck, S., 2003. Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise. IT and Society 1, 64–89.

Johnson, F. C., Griffiths, J. R., Hartley, R. J., 2003. Task dimensions of user evaluations of information retrieval systems. Information Research.

Joho, H., Hannah, D., Jose, J. M., 2008. Comparing collaborative and independent search in a recall-oriented task. In: Proceedings of the second international symposium on Information interaction in context. pp. 89–96.

Jones, K., Bates, R., 1977. Report on a Design Study for the 'ideal' Information Retrieval Test Collection. Computer Laboratory, University of Cambridge.

Jose, J. M., Furner, J., Harper, D. J., 1998. Spatial querying for image retrieval: a user-oriented evaluation. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 232–240.

Kaki, M., Aula, A., 2008. Controlling the complexity in comparing search user interfaces via user studies. Information Processing & Management.

Kamps, J., Geva, S., Peters, C., Sakai, T., Trotman, A., Voorhees, E., 2009. Report on the SIGIR 2009 workshop on the future of information retrieval evaluation. SIGIR Forum 43, 13–23.

Kanoulas, E., Aslam, J. A., 2009. Empirical justification of the gain and discount function for ndcg. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). pp. 611–620.

Kanoulas, E., Carterette, B., Clough, P., Sanderson, M., 2011. Evaluating multi-query sessions. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1053–1062.

Kantor, P. B., Voorhees, E. M., 2000. The trec-5 confusion track: Comparing retrieval methods for scanned text. Inf. Retr.

Katter, R., 1968. The influence of scale form on relevance judgments. Information Storage and Retrieval.

Kaufmann, E., 2007. Talking to the semantic web — natural language query interfaces for casual end-users. Ph.D. thesis, University of Zurich.

Kaufmann, E., Bernstein, A., Fischer, L., 2007. NLP-Reduce: A "naïve" but Domain-independent Natural Language Interface for Querying Ontologies. In: Proceedings of the 4th European Semantic Web Conference (ESWC2007). Vol. 4519.

Kaufmann, E., Bernstein, A., Zumstein, R., 2006. Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs. In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006). pp. 980–981.

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. Vol. 6611. Springer-Verlag, pp. 165–176.

Kelly, D., 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. Found. Trends Inf. Retr. 3, 1–224.

Kelly, D., Dumais, S., Pedersen, J., 2009. Evaluation challenges and directions for information-seeking support systems. Computer 42 (3), 60–66.

Kelly, D., Lin, J., 2007. Overview of the TREC 2006 ciQA task. SIGIR Forum 41, 107–116.

Kemp, D., 1974. Relevance, pertinence and information system development. Information Storage and Retrieval.

Kent, A., Berry, M. M., Luehrs, Perry, J. W., 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. American Documentation.

Kinney, K. A., Huffman, S. B., Zhai, J., 2008. How evaluator domain expertise affects search result relevance judgments. In: Proceedings of the 17th ACM conference on Information and knowledge management. CIKM '08. ACM, pp. 591–598.

Koenemann, J., Belkin, N. J., 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '96. ACM, pp. 205–212.

Lewis, J. R., 1994. Sample sizes for usability studies: Additional considerations. The Journal of the Human Factors and Ergonomics Society.

Lewis, J. R., 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 7, 57–78.

López, V., Motta, E., Uren, V., 2006. PowerAqua: Fishing the Semantic Web. In: The Semantic Web: Research and Applications. pp. 393–410.

López, V., Pasin, M., Motta, E., 2005. AquaLog: An Ontology-Portable Question Answering System for the Semantic Web. In: The Semantic Web: Research and Applications. Springer, pp. 269–272.

Losee, R. M., 1998. Text retrieval and filtering: analytic models of performance. Kluwer Academic Publishers, Norwell, MA, USA.

Loupy, C. D., Bellot, P., 1997. Evaluation of document retrieval systems and query difficulty.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Penas, A., Peinado, V., Verdejo, F., de Rijke, M., Vallin, R., 2003. The multiple language question answering track at clef 2003. In: CLEF 2003 Workshop. pp. 471–486.

Manning, C., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

McGrath, J. E., 1995. Methodology matters: doing research in the behavioral and social sciences. In: Human-computer interaction. Morgan Kaufmann Publishers Inc., pp. 152–169.

Meij, E., Bron, M., Hollink, L., Huurnink, B., Rijke, M., 2009. Learning semantic query suggestions. In: Proceedings of the 8th International Semantic Web Conference (ISWC2009). pp. 424–440.

Mizzaro, S., 1998. How many relevances in information retrieval? Interacting with Computers.

Müller, H., 2010. Creating realistic topics for image retrieval evaluation. In: Müller, H., Clough, P., Deselaers, T., Caputo, B. (Eds.), ImageCLEF. Vol. 32 of The Information Retrieval Series. Springer Berlin Heidelberg, pp. 45–61.

Navarro-Prieto, N., Scaife, M., Rogers, Y., 1999. Cognitive strategies in web searching. Proceedings of the 5th Conference on Human Factors and the Web 2004, 1–13.

Nielsen, J., 1993. Usability Engineering. Morgan Kaufmann Publishers Inc.

Nielsen, J., 1994. Estimating the number of subjects needed for a thinking aloud test. International Journal of Human-Computer Studies.

Over, P., 1997. TREC-6 Interactive track report. In: TREC. pp. 57–64.

Over, P., 2001. The trec interactive track: an annotated bibliography. Information Processing & Management.

Page, S., 2000. Community research: The lost art of unobtrusive methods1. Journal of Applied Social Psychology.

Perfetti, C., Landesman, L., 2001. Eight is not enough. `http://www.uie.com/articles/eight_is_not_enough`, retrieved: August 2012.

Peters, C., Braschler, M., 2001. Cross-language system evaluation: The CLEF campaigns. Journal of the American Society for Information Science and Technology 52 (12), 1067–1072.

Peters, T., 1993. The history and development of transaction log analysis. Library Hi Tech.

Petrelli, D., 2008. On the role of user-centred evaluation in the advancement of interactive information retrieval. Information Processing & Management 44, 22 – 38.

Pound, J., Mika, P., Zaragoza, H., 2010. Ad-hoc object retrieval in the web of data. In: Proceedings of the 19th international conference on World wide web (WWW2010). pp. 771–780.

Radlinski, F., Craswell, N., 2010. Comparing the sensitivity of information retrieval metrics. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 667–674.

Rees, A., Schultz, D., 1967. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Case Western Reserve University.

Rhenius, D., Deffner, G., 1990. Evaluation of Concurrent Thinking Aloud using Eye-tracking Data. Human Factors and Ergonomics Society Annual Meeting Proceedings.

Rice, R., Borgman, C., 1983. The use of computer monitored data in information science and communication research. Journal of the American Sociey for Information Science.

Roberts, T. L., Moran, T. P., 1983. The evaluation of text editors: methodology and empirical results. Commun. ACM, 265–283.

Robertson, S., 2008. On the history of evaluation in ir. Journal of Information Science 34, 439–456.

Robertson, S. E., 1981. The methodology of information retrieval experiment. Butterworths, pp. 9–31.

Robertson, S. E., Hancock-Beaulieu, M. M., 1992. On the evaluation of ir systems. Information Processing and Management 28, 457–466.

Sakai, T., 2006. Evaluating evaluation metrics based on the bootstrap. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 525–532.

Sakai, T., 2007. On the reliability of information retrieval metrics based on graded relevance. Information Processing & Management.

Sanderson, M., 2010. Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval.

Sanderson, M., Paramita, M., Clough, P., Kanoulas, E., 2010. Do user preferences and evaluation measures line up? In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 555–562.

Sanderson, M., Zobel, J., 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR2005). pp. 162–169.

Saracevic, T., 1995. Evaluation of evaluation in information retrieval. In: Proceedings of SIGIR. pp. 138–146.

Saracevic, T., 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. J. Am. Soc. Inf. Sci. Technol.

Saracevic, T., Kantor, P., Chamis, A. Y., Trivison, D., 1988. "a study of information seeking and retrieving. i. background and methodology". Journal of the American Society for Information Science 39, 161–176.

Schamber, L., 1994. Relevance and information behavior. Annual Review of Information Science and Technology (ARIST).

Schmidt, M., Hornung, T., Lausen, G., Pinkel, C., 2008. SP2Bench: A SPARQL Performance Benchmark. CoRR.

Shackel, B., 1986. Ergonomics in design for usability. In: People & computers: Designing for usability. Proceedings of the second conference of the BCS HCI specialist group. Cambridge University Press., pp. 44–64.

Shneiderman, B., 1986. Designing the user interface: strategies for effective human-computer interaction. Addison-Wesley Longman Publishing Co., Inc.

Sillito, J., Murphy, G. C., De Volder, K., 2006. Questions programmers ask during software evolution tasks. In: Proceedings of the 14th ACM SIGSOFT international symposium on Foundations of software engineering. pp. 23–34.

Sillito, J., Murphy, G. C., Volder, K. D., 2008. Asking and answering questions during a programming change task. IEEE Transactions on Software Engineering 34, 434–451.

Smucker, M., Clarke, C., 2012. Time-based calibration of effectiveness measures. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 95–104.

Soergel, D., 1994a. Indexing and retrieval performance: the logical evidence. J. Am. Soc. Inf. Sci.

Soergel, D., 1994b. Indexing and retrieval performance: the logical evidence. Journal of the American Society for Information Science.

Spärck Jones, K., 2000. Further reflections on TREC. Information Processing & Management 36, 37–85.

Spärck Jones, K., Van Rijsbergen, C. J. K., 1976. Information Retrieval Test Collections. Journal of Documentation.

Spink, A., 2002. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. Information Processing & Management.

Spink, A., Greisdorf, H., Bateman, J., 1998. From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management.

Spyridakis, J., 1992. Conducting research in technical communication: The application of true experimental designs. Technical Communication.

Stonebraker, M., Frew, J., Gardels, K., Meredith, J., 1993. The SEQUOIA 2000 storage benchmark. SIGMOD Rec.

Su, L. T., 1992. Evaluation measures for interactive information retrieval. Information Processing & Management 28, 503 – 516.

Su, L. T., 1998. Value of search results as a whole as the best single measure of information retrieval performance. Information Processing % Management.

Su, L. T., 2003. A comprehensive and systematic model of user evaluation of web search engines: Ii. an evaluation by undergraduates. Journal of the American Society for Information Science and Technology.

Tabatabai, D., Shore, B. M., 2005. How experts and novices search the Web. Library & Information Science Research 27, 222 – 248.

Tagliacozzo, R., 1997. Estimating the satisfaction of information users. Bulletin of the Medical Library Association.

Tague, J., Schultz, R., 1989. Evaluation of the user interface in an information retrieval system: A model. Information Processing & Management.

Tague-sutcliffe, J., 1992. The pragmatics of information retrieval experimentation, revisited. Information Processing & Management.

Tague-sutcliffe, J., Blustein, J., 1994. A statistical analysis of the trec-3 data. In: Overview of the Third Text REtrieval Conference (TREC-3). pp. 385–398.

Tang, L. R., Mooney, R. J., 2001. Using Multiple Clause Constructors in Inductive Logic Programming for Semantic Parsing. In: Proceedings of the 12th European Conference on Machine Learning. pp. 466–477.

Tang, R., Shaw, W. M., Vevea, J. L., 1999. Towards the identification of the optimal number of relevance categories. Journal of the American Society for Information Science.

Taube, M., 1956. Cost as the measure of efficiency of storage and retrieval systems. In: Studies in Coordinate Indexing. Defense Technical Information Center, pp. 18–33.

Taube, M., 1965. A note on the pseudo-mathematics of relevance. American Documentation.

Tessier, J., Crouch, W., Atherton, P., 1977. New Measures of User Satisfaction With Computer Based Literature Searches. Special Libraries.

Thomas, P., Hawking, D., 2006. Evaluation by comparing result sets in context. In: Proceedings of the 15th ACM international conference on Information and knowledge management. CIKM '06. pp. 94–101.

Tsakonas, G., Kapidakis, S., Papatheodorou, C., 2004. Evaluation of user inter-

action in digital libraries. In: Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries.

Tullis, T., Albert, W., 2008. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Elsevier/Morgan Kaufmann.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., Decker, S., 2010. Sig.ma: live views on the web of data. In: Proceedings of the 19th international conference on World wide web (WWW2010). pp. 355 – 364.

Tummarello, G., Oren, E., Delbru, R., 2007. Sindice.com: Weaving the Open Linked Data. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007). pp. 552–565.

Turner, C. W., Lewis, J. R., Nielsen, J., 2006. Determining usability test sample size. International Encyclopedia of Ergonomics and Human Factors.

Turpin, A. H., Hersh, W., 2001. Why batch and user evaluations do not give the same results. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 225–231.

Unger, C., Cimiano, P., Lopez, V., Motta, E., 2011. Proceedings of the 1st Workshop on Question Answering Over Linked Data (QALD-1).
URL `http://www.sc.cit-ec.uni-bielefeld.de/sites/www.sc.cit-ec.uni-bielefeld.de/files/proceedings.pdf`

Vickery, B. C., 1959a. Subject analysis for information retrieval. In: Proceedings of the International Conference on Scientific Information. pp. 41–52.

Vickery, B. C., 1959b. The structure of information retrieval systems. In: Proceedings of the International Conference on Scientific Information. pp. 1275–1290.

Virzi, R. A., 1990. Streamlining the design process: Running fewer subjects. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. pp. 291–294.

Virzi, R. A., 1992. Refining the test phase of usability evaluation: how many subjects is enough? Human Factors.

Vitek, J., Kalibera, T., 2011. Repeatability, reproducibility, and rigor in systems research. In: Chakraborty, S., Jerraya, A., Baruah, S. K., Fischmeister, S. (Eds.), EMSOFT. ACM, pp. 33–38.

Voorhees, E., 2001. Evaluation by highly relevant documents. In: In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 74–82.

Voorhees, E., 2002. The Philosophy of Information Retrieval Evaluation. In: Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems (CLEF2001). pp. 355–370.

Voorhees, E. M., 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 315–323.

Voorhees, E. M., 1999. The TREC-8 Question Answering Track Report. In: In Proceedings of TREC-8. pp. 77–82.

Voorhees, E. M., 2003. Overview of TREC 2003. In: TREC. pp. 1–13.

Voorhees, E. M., 2009. Topic set size redux. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '09. ACM, pp. 806–807.

Voorhees, E. M., Buckley, C., 2002. The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 316–323.

Voorhees, E. M., Harman, D., 2000. Overview of the Ninth Text REtrieval Conference (TREC-9). In: In Proceedings of the Ninth Text REtrieval Conference (TREC-9. pp. 1–14.

Voorhees, E. M., Harman, D. K., 2005. TREC: Experiment and Evaluation in Information Retrieval. Digital Libraries and Electronic Publishing. MIT Press.

Walker, S., DeVere, R., 1990. Improving subject retrieval in online catalogues. 2. relevance feedback and query expansion. London: British Library.

Walker, S., Hancock-Beaulieu, M., City University (London, E. C. f. I. S. R., 1991. Okapi at City: An Evaluation Facility for Interactive IR. Centre for Interactive Systems Research, City University.

Webb, E. J., Campbell, D. T., Schwarz, R. D., Sechrest, L., 2000. Unobtrusive measures. Sage Publications.

White, R. W., Bilenko, M., Cucerzan, S., 2007. Studying the use of popular destinations to enhance web search interaction. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 159–166.

Whiteside, J., Jones, S., Levy, P. S., Wixon, D., 1985. User performance with command, menu, and iconic interfaces. SIGCHI Bull., 185–191.

Wilson, M. L., Kules, B., schraefel, m. c., Shneiderman, B., 2010. From keyword search to exploration: Designing future search interfaces for the web. Foundations and Trends in Web Science 2 (1).

Wrigley, S. N., Elbedweihy, K., Reinhard, D., Bernstein, A., Ciravegna, F., 2010a. Evaluating semantic search tools using the seals platform. In: International Workshop on Evaluation of Semantic Technologies (IWEST 2010), ISWC 2010.

Wrigley, S. N., Elbedweihy, K., Reinhard, D., Bernstein, A., Ciravegna, F., November 2010b. Results of the first evaluation of semantic search tools. Technical report, SEALS Consortium.

Wrigley, S. N., García-Castro, R., Trojahn, C., 2011. Infrastructure and workflow for the formal evaluation of semantic search technologies. In: Proceedings of the workshop on Data infrastructures for supporting information retrieval evaluation. pp. 29–34.

Wrigley, S. N., K.Elbedweihy, Gentile, A., Lanfranchi, V., Dadzie, A.-S., 2012. Results of the second evaluation of semantic search tools. Tech. Rep. D13.6, SEALS Consortium.

Wrigley, S. N., Reinhard, D., Elbedweihy, K., Bernstein, A., Ciravegna, F., 2010c. Methodology and campaign design for the evaluation of semantic search tools. In: Proceedings of the 3rd International Semantic Search Workshop. SEMSEARCH '10. pp. 10:1–10:10.

Xie, H., 2003. Supporting ease-of-use and user control: desired features and structure of web-based online IR systems. Information Processing & Management 39, 899–922.

Yilmaz, E., Aslam, J. A., 2006. Estimating average precision with incomplete and imperfect judgments. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp. 102–111.

Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S., 2010. Expected browsing utility for web search evaluation. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 1561–1564.

Zhang, J., Kamps, J., 2010. A search log-based approach to evaluation. In: Proceedings of the 14th European conference on Research and advanced technology for digital libraries. ECDL'10. Springer-Verlag, pp. 248–260.

Zhou, K., Zha, H., Xue, G.-R., Yu, Y., 2012. Learning the gain values and discount factors of dcg. CoRR abs/1212.5650.

Zobel, J., 1998. How reliable are the results of large-scale information retrieval experiments? In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 307–314.

Zobel, J., Webber, W., Sanderson, M., Moffat, A., 2011. Principles for robust evaluation infrastructure. In: Proceedings of the workshop on Data infrastructures for supporting IR evaluation. pp. 3–6.

*Detailed Response to Reviewers

We thank the reviewers for their thorough and helpful critique of the paper. We have implemented changes relating to their suggestions, and believe that, thanks to their input, the paper is now substantially improved.