Toward a Meaningful E-Discovery Standard

On the Place of Measurement in an E-Discovery Standard

Bruce Hedin
H5
San Francisco, USA
bhedin@h5.com

Dan Brassil
H5
New York, USA
dbrassil@h5.com

Christopher Hogan
H5
San Francisco, USA
chogan@h5.com

ABSTRACT

As the e-discovery industry grows and matures, there is increasing recognition of the need for certification standards that potential consumers of e-discovery products and services can use as a means of discerning what is sound from what is not. In searching for models of certification standards that might be applied to e-discovery, some have turned to the ISO 9000 family of standards, a set of standards focused on the validation of the quality management systems employed in an on-going business process; others have turned to the ISO/IEC 27000 family of standards, a set of standards focused on best practices and requirements for information security management systems (ISMS).

This paper argues that the potential benefits of a standard can be realized only if the standard addresses the central question potential consumers have when evaluating an e-discovery product or service: *how accurate are the results?* That question will be addressed only if the standard makes provision for the statistically sound measurement of the effectiveness of the review/retrieval function of an e-discovery system. The paper therefore takes the position that an e-discovery standard should require that providers of e-discovery services include the valid measurement of recall and precision as a central component of their quality management system. The paper argues that provision for such a requirement (which is a requirement, not that the provider attain any specific level of recall or precision, but only that the provider have the capability of measuring recall and precision in a meaningful way) would benefit both consumers and providers of e-discovery services, as well as the legal profession as a whole.

The paper also considers possible objections to the inclusion of measurement in a standard. The paper holds that a measurement provision need not require the specification of minimum thresholds for recall and precision, need not open the door to misuse of the results of measurement, need not entail undue cost and time in the provision of e-discovery services, and need not provide opportunities to "game the system."

The paper observes, finally, that both the ISO 9000 and the ISO 27000 frameworks are flexible enough to allow for the inclusion of a measurement provision in a standard written for e-discovery products and services.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms

Standardization, Verification, Measurement

Keywords

E-discovery, measurement, sampling design, standards

1. INTRODUCTION

As the challenge ever increasing volumes of electronically stored information (ESI) pose to the responsible execution of a response to a discovery request grows more acute (see, for example, [12]), as the commercial offerings that claim to be able to meet that challenge increase in variety and number, and as efforts to meet the challenge come under closer and better-informed scrutiny by the courts (see, for example, [4]), a consensus is emerging that there is need for a generally applicable standard for segregating sound solutions from unsound ones. The contributions to the 2011 DESI IV Workshop (on setting standards for searching electronically stored information in discovery [1]) are testimony to this emerging consensus, as are the contributions to the 2013 DESI V Workshop (on standards for using predictive coding, machine learning, and other advanced search and review methods in e-discovery [2]).

While, however, there is general consensus that a standard would be valuable, there is not a consensus on the specific form that such a standard should take. Should the scope of the standard be broad (encompassing all steps in the discovery of ESI) or narrow (focusing on just the execution of review or search)? Should the standard be of the "guidance" variety (detailing best practice recommendations) or of the "requirements" variety (specifying concrete requirements compliance with which could be audited and certified)? Should the standard focus on the process whereby a discovery response is executed (specifying that any process include certain steps) or should it simply focus on the results of that response (specifying that the outcome of any process should meet certain thresholds of quality)? Questions such as these must be answered if the industry is to move beyond the general recognition that a standard (of some sort) would be helpful and to the actual development of a meaningful standard.

In this paper, we consider the place of measurement in a meaningful standard for the discovery of ESI. We begin by briefly summarizing the potential benefits of a standard (Section 2). We next review some of the standards that have been proposed as models for an e-discovery standard (Section 3). We then turn to the question of measurement. We look first at the way in which a provision for measurement could be included in a standard (Section 4.1); we then review

Position paper submitted to DESI V: ICAIL 2013 Workshop on Standards for Using Predictive Coding, Machine Learning, and Other Advanced Search and Review Methods in E-Discovery. Consiglio Nazionale delle Ricerche, Rome, Italy. June 14, 2013. `http://www.umiacs.umd.edu/~oard/desi5/`

some of the benefits that would be gained by including such a provision (Section 4.2) and consider possible objections to the inclusion of measurement in a standard (Section 4.3). Finally, we show that any of the standard frameworks that have been proposed would be flexible enough to allow the incorporation of measurement as an element (Section 4.4).

2. THE NEED FOR STANDARDS

The rapid growth in the volumes of ESI held by parties involved in litigation has resulted in demands being placed on parties served with discovery requests that exceed the capabilities of the traditional approach to identifying material responsive to a request (traditional linear manual review). Fortunately, concomitant with the rapid growth in ESI has been the development of information retrieval tools and methods that hold promise as being capable of meeting even very steep discovery demands; see, for example, the results of the series of evaluations conducted in the TREC Legal Track [3] [11]. Unfortunately, however, fully understanding and evaluating these tools and methods can require kinds of expertise (computer science, linguistics, statistics, and so on) well outside the legal expertise expected of most attorneys, and so, not surprisingly, the legal profession has been somewhat hesitant to adopt what it does not fully understand. This is where standards come in.

For potential consumers of e-discovery products and services, a standard would provide relief from having to acquire the information-retrieval expertise required to conduct a thorough evaluation of the offerings; the consumer would need to know simply whether or not the offering met the standard.

For providers of e-discovery products and services, a standard would dispel some of the fog that envelopes the requirements for adoption and allow them to frame their discussions of their offerings around open and mutually understood criteria.

For the bench, a standard would provide relief from the need "to go where angels fear to tread" [5] in opining on the adequacy of a responding party's chosen review methodology; an objective standard would be available to help make that determination.

For the legal profession as a whole, a standard, by clearly marking out a path to the defensible use of new technologies and methods, would facilitate adoption of e-discovery products and services and accelerate the development of still more effective ones.

It is with benefits such as these in mind that many in the legal profession and in the e-discovery industry have come to see the development of a transparent e-discovery standard as a matter of importance and urgency.

3. PROPOSED MODELS

If there is a consensus that an e-discovery standard would be a good thing, the next question is that of the form that such a standard should take. As a first step in answering this question, some have looked to already existing standards as models for what might be articulated for e-discovery. Two such models are the ISO 9000 family of standards and the ISO/IEC 27000 family of standards.

3.1 The ISO 9000 family of standards

The ISO 9000 family of standards [6] received much attention at the 2011 DESI IV workshop; a helpful discussion of the applicability of this standard to e-discovery processes is provided by Knox and Dawson's submission to DESI IV [10].

The ISO 9000 family of standards covers requirements and best practices for the quality management systems that companies employ to ensure that their products and services meet the expectations of their customers. The most generally applicable standard in the family is ISO 9001. This standard is a "requirements" standard, meaning that it specifies principles that a quality management system should adhere to, provides for both internal and independent audits of adherence, and allows for public certification of compliance for those quality management systems found to adhere to the standard's criteria.

Examples of the quality management principles specified by the standard are: (i) a process focus (i.e., taking an end-to-end view of all resources that contribute to the delivery of a product or service and seeing that those resources are integrated into an efficient and coherent process); (ii) fact-based decision making (i.e., basing management decisions on the results of data collection and analysis); and (iii) continual improvement (i.e., continually striving to make measurable improvements in performance) [7].

The standard provides for both internal and independent audits of compliance; when the latter are conducted by an appropriate certification body, the audit can lead to a public certificate of compliance with the standard. Importantly, the standard specifies that the audits not be of the "one-and-done" sort; fresh audits must be carried out on a regular basis in order to ensure on-going compliance.

In application, either the generic ISO 9001 standard may be used or an industry may choose to develop and use an industry-specific version of the standard, one tailored to the specific conditions and objectives of the industry in question.

3.2 The ISO/IEC 27000 family of standards

Others have looked to another set of standards, the ISO/IEC 27000 family, as the appropriate home for an e-discovery standard; a helpful overview of these efforts is provided in a recent article by Steven Teppler, co-chair of the E-Discovery and Digital Evidence Committee of the ABA's Science and Technology Law Section [14].

The ISO/IEC 27000 family of standards, the joint work of the ISO and the International Electrotechnical Commission (IEC), is a set of standards that focus on information security management systems (ISMS); the general objective of the standards is to assist organizations in the development and operation of an ISMS [8]. In April of 2013, a subcommittee (SC 27) of the joint technical committee (JTC 1) that is the forum for collaboration between the ISO and IEC gave approval to the drafting of a standard focused specifically on e-discovery processes; the standard is to be called "ISO/IEC 27050, Information Technology – Security Techniques – Electronic Discovery."

While the specifics of ISO/IEC 27050 remain to be seen (a working draft is to be completed in July of 2013), two general features of what is planned are apparent. First, at least initially, the standard will be of the "guidance" sort rather than of the requirements sort. This means that the objective of the standard will be to normalize terminology around e-discovery processes and to provide models for the implementation and operation of those processes; the objective will not be to specify requirements compliance with which would be auditable and certifiable. It is envisioned, however, that the standard would evolve over time and it is not impossible that requirements could be specified in a future standard. Second, the standard will likely be broad in scope, covering a wide range of activities related to the discovery of ESI, from identification and collection through to production.

We believe that both the ISO 9000 and the ISO/IEC 27000

frameworks are flexible enough to be the basis for a meaningful e-discovery standard. We also believe, however, that regardless of which framework is taken as the model, an e-discovery standard must, if it is to be meaningful, include certain elements. Chief among these elements is provision for the measurement of accuracy; to this element we turn for the remainder of this paper.

4. THE PLACE OF MEASUREMENT

We begin our discussion of measurement by considering how an e-discovery standard might provide for the measurement of accuracy (Section 4.1). We then review the advantages of including the measurement of effectiveness in a standard (Section 4.2). We next discuss possible objections to including such measurement in a standard (Section 4.3). Lastly, we consider the incorporation of a measurement provision into a standard of either the ISO 9000 or ISO/IEC 27000 sort (Section 4.4).

4.1 Including measurement in a standard

Earlier (Section 2) we noted some of the benefits an e-discovery standard could bring. These benefits can be realized, however, only if the standard addresses the central question potential users have when evaluating an e-discovery product or service: *how accurate are the results?* And that question will be addressed only if the standard makes provision for the statistically sound measurement of the effectiveness of the review/retrieval function of an e-discovery system.[1]

More specifically, the user of an e-discovery product or service will have two basic questions, both of which have to do with the output of the system and not with the process whereby the system arrives at that output. First, out of all that the system was asked to find, how much did it actually find? Second, out of all that the system identified as relevant, how much was actually relevant? The former question is answered, quantitatively, by the information retrieval metric known as *recall*; the latter question is answered by the metric known as *precision*. We believe that an e-discovery standard will be meaningful only if it provides that statistically-sound estimation of recall and precision be a required element of the quality management system employed in an e-discovery system.

The standard should provide for the measurement of recall and precision in two ways: first, in the requirements specified for an e-discovery system and, second, in the protocol specified for the execution of system audits.

With regard to the requirements, the standard should specify that an e-discovery system must include documentation of the protocol whereby it measures the recall and precision achieved by its review/retrieval function on any given project. The standard need not specify the precise form that such a measurement protocol should take. Indeed, given that there is more than one way to estimate recall and precision, and given that the most efficient and appropriate way may vary with circumstance (e.g., on a single batch production *vs.* on rolling productions *vs.* on streaming preservation efforts), the standard should allow providers flexibility on this point. The standard should require, however, that the protocol be documented in sufficient detail and with sufficient transparency that an expert reviewing the documentation could reach reasonable conclusions as to whether or not the protocol would be likely to yield valid estimates.

With regard to audits, the standard should require that evaluation of the effectiveness of a system's measurement protocol be an element of any audit. Such an evaluation would involve the execution of an independent sampling and measurement protocol either on an already-completed project or, if that were not possible due to confidentiality considerations, on the results of a test exercise carried out on an evaluation data set provided by the certification body. The results of the auditor's independent measurement exercise would then be compared with the results that the provider had obtained, using its internal measurement protocol, on the same project. If the results (i.e., estimates of recall and precision) were similar, the provider's protocol would be validated. If the results were materially dissimilar, further scrutiny of the provider's measurement protocol would be triggered.

4.2 The benefits of measurement

By requiring that an e-discovery process include the sound measurement of recall and precision as a central component of its quality management system, an e-discovery standard would bring benefits to users of e-discovery systems, to providers of those systems, to the bench, and to the legal profession as a whole.

4.2.1 For consumers

For users of e-discovery products and services, the inclusion of a provision for the measurement of accuracy would provide greater certainty both when evaluating potential providers and when actually employing a chosen system.

The potential buyer, when evaluating candidate providers, could expect that any certified candidate could provide concrete data on levels of accuracy achieved on prior projects (and that information could be provided without disclosing any confidential information on the specifics of the example projects). The transparency this would impart to the buying process would help to align consumers' requirements and expectations with providers' actual capabilities.

The user of a system would have the assurance that any certified system had the capability of providing direct quantitative answers to the key questions they are likely to have about the accuracy of the review or retrieval effort (of all that was sought, how much was actually found; of all that was found, how much was what was actually sought). The consumer would know that that capability could be drawn upon both in the internal decision making process leading up to a production and, in the event of a challenge, in defending the adequacy of a production.

4.2.2 For producers

For providers of e-discovery products and services, the inclusion of a measurement requirement in a standard would require that they do no more than what they already should be doing. The accurate estimation of recall and precision is an essential element both of an iterative retrieval process and of a meaningful protocol for validating final results; to require that a provider of e-discovery services include accuracy

[1]Note that, for purposes of this discussion, the potential "user" of an e-discovery system could be either (i) a corporate entity or government agency that is party to a lawsuit or investigation or (ii) a law firm representing a party to a lawsuit or investigation. The "provider" of an e-discovery system could be either (i) a vendor of e-discovery tools, (ii) a vendor of e-discovery services, or (iii) a law firm that, supported by a vendor of e-discovery tools or services, provides e-discovery services to clients. (A law firm, it should be noted, could, depending on circumstances, be in either the user or the provider role.) Any provider, whether a law firm, a vendor of tools, or a vendor of services, that wished to be certified to an e-discovery standard would be subject to the requirements we are proposing in this section.

measurement as a component of its quality management system is nothing more than to require that the provider acquire an essential tool of the trade.

Providers would also benefit from the fact that the inclusion of such a requirement in a standard would provide common reference points for discussions of accuracy with potential clients; by adding concrete definition and clarity to such discussions, the standard will enable certified providers to answer more easily the questions about quality that buyers of new technologies will inevitably have.

4.2.3 For the bench

For the bench, the inclusion of a measurement provision in a standard would, by allowing a shift of attention from process to results, simplify the resolution of discovery disputes.

The expectation that a certified provider would have the capability of obtaining valid estimates of recall and precision would relieve judges from the need, when addressing a discovery dispute, to delve into the details of process. The specifics of a retrieval process will vary considerably from provider to provider and are likely to evolve over time as new tools and methods are developed; efforts to get parties to agree on these specifics are likely to end in frustration. The expectation that sound measures of effectiveness will be available, however, makes agreement on process unnecessary. Attention can focus instead on the results of the retrieval effort and on the more straightforward question of whether the estimates of the recall and precision attained by the effort are evidence that the production in question is reasonably accurate and complete.

4.2.4 For the legal profession as a whole

For the legal profession and the e-discovery industry, the inclusion of a measurement provision in a standard would: (i) bring questions of accuracy into the open; (ii) foster a common terminology for discussing such questions; (iii) foster realistic expectations as to levels of recall and precision that can and should be achieved; and (iv) ensure that the e-discovery standard would have a substantive impact on the quality of e-discovery products and services.

4.3 Possible objections

While the potential benefits of including a measurement provision in a an e-discovery standard are fairly clear, there are also reasons some might object to such a provision. In this section, we consider possible objections.

4.3.1 No consensus on a number

Some might object to a provision for the measurement of recall and precision on the grounds that there is as yet no consensus on the minimum levels of recall and precision that are required for a document production to be acceptable. "Should recall be at least 70%?" "How about 72%?" "Why not 75%?" Given the absence of a consensus, and thus the absence of an agreed-upon quantitative definition of what does and does not count as "accurate," some might argue that providing for the measurement of accuracy is out of place in a standard.

Note, however, that the requirement that we are proposing is a requirement, not for a *number*, but for a *capability*. That is to say, the requirement does not specify that an e-discovery process must meet some minimum level of recall and precision. What the requirement specifies is only that the process include the capability of estimating recall and precision as a component of its quality management system; that is a capability that any review or retrieval process must have if it is to enable any claims about the quality of its results.

The requirement is thus not that minimum levels of recall and precision be met; the requirement is simply that estimates of recall and precision be available. It remains up to practitioners, taking into account the specific goals and circumstances of the retrieval effort, and taking into account non-statistical data as well, to decide what levels are required to have confidence in the results of a review or retrieval effort.

We note, moreover, that, while a consensus on minimum levels of recall and precision is currently lacking, and while such a consensus is certainly not attainable for all circumstances, greater transparency as to the levels both that can realistically be achieved and that can efficiently be demonstrated may foster greater consensus about what those levels should be, at least in the most typical cases. In that regard, including a measurement provision in a standard may actually help the development of consensus.

4.3.2 The glass is 20% empty

Some might object to a measurement provision on the grounds that, in an adversarial system, quantitative measures will only provide fodder for distracting and unproductive argument. Given a recall estimate of 80%, for example, opposing counsel may try to focus all attention on the 20% that has not been retrieved; or opposing counsel may use the absence of a well-defined minimum threshold to demand that the producing party continue to work to increase recall to higher and higher levels.

We note, however, that there are reasonable and empirically well-grounded responses to such objections. To the "glass is 20% empty" argument, for example, one might respond by drawing the distinction between *document* recall and *information* recall. Our recall measures are almost always of the former variety; i.e., we measure the proportion of relevant *documents* retrieved out of all relevant *documents* that reside in a population. Once, however, a reasonably high level of document recall has been achieved (e.g., 80%), we typically find that, when additional relevant documents are found, those new documents in fact add no new information; that is to say, the information in the unretrieved 20% of relevant documents is almost entirely redundant to information in the already-retrieved 80% of relevant documents. When, therefore, one has achieved 80% *document* recall, one will have typically achieved a much higher level of *information* recall. Put in terms of information, then, the glass is not in fact 20% empty; it is in fact almost entirely full.

To a demand for ever-higher levels of recall ("If you have achieved 80% recall, why not make the additional effort to get to 81%?" and "If you have achieved 81% recall, why not make the additional effort to get to 82%?" and so on), one might respond by pointing to the fact of diminishing returns. Given the composition of most document collections and given the nature of most topics pertinent to those collections, it is typically the case that, as more relevant documents have been retrieved, the cost of finding additional relevant documents increases; indeed, it could well be as costly, in terms of time and resources, to go from 80% to 85% recall as it was to go from 0% to 80% recall. Each percentage of recall, therefore, tends to come at greater cost, and, at some point, that cost will outweigh the value of additional documents.

We cite these responses simply to illustrate that there are available reasonable and substantive answers to potentially distracting cavils about measurement. We would argue, in

fact, that the legal profession and the e-discovery industry would be better served by getting such arguments out in the open rather than avoiding them. If a measurement provision in a standard helped to bring such discussions into the light of day, that would not be a bad thing.

4.3.3 Time and cost

A third objection to the inclusion of a measurement provision in an e-discovery standard centers on time and cost. Obtaining meaningful estimates of precision and recall, and especially of the latter, can require very large samples; given real-world time constraints and the stakes at issue in a matter, the cost and time it would take to draw and assess samples of such size would often be prohibitive. Practitioners cannot be expected to run TREC-like exercises every time they produce a set of documents.

In response to this objection, we note that measurement need not imply a TREC-like evaluation. There are in fact a number of ways to arrive at sound measures of recall, and some of these are less costly than others. If care is taken to weigh the actual information need in a given circumstance against the sampling and review costs required to meet that need, one can, in most cases, arrive at a sampling design that will, in a cost-effective manner, contribute meaningful empirical data to the validation of the results of a retrieval effort. Indeed, a properly executed measurement program can, by illuminating in real time where improvements do and do not need to be made, often reduce the overall time and cost of a retrieval effort

By way of illustration, we take a brief look at the application of the acceptance-testing paradigm to the validation of retrieval results. Our goal here is not to provide a full discussion of the application of the approach to e-discovery (we plan to do that elsewhere); our goal is simply to illustrate some of the flexibility that the approach affords practitioners wishing to validate their results (for a discussion of acceptance sampling approaches in general, see [13]; for an **R** package helpful in designing and evaluating acceptance tests, see [9]).

Under the acceptance-testing paradigm, the aim of the sampling exercise is not to arrive at a precise estimate of recall (i.e., an estimate associated with a very narrow confidence interval); the aim is rather to establish whether or not it can be stated, at a given level of confidence, that recall is at or above some pre-specified level. The exercise is thus that of a pass/fail test: a passing result means that one can state, at the specified level of confidence, that one's recall is at or above the specified threshold; a failing result means that one cannot make that statement at that level of confidence. To be sure, this approach provides less information than does an approach that provides narrow confidence bounds around a recall estimate; the approach also, however, generally entails lower costs in terms of sampling and review. As long as the pre-specified recall threshold is a meaningful one, the acceptance-testing approach to validation may provide all the information one really needs and may do so in a more economical way than could alternative approaches.

The acceptance-testing approach, moreover, puts a number of test parameters at the discretion of the designer of the test; by varying the values of these parameters, the designer can manage the balance between the levels of recall gauged, the sampling uncertainty tolerated, and the sample size required.

To illustrate with a concrete example, suppose we had a population of 1,000,000 documents. Suppose that a retrieval effort had been conducted on that population and that the effort had coded 30,000 of the documents as responsive (3% of the full population, a not atypical yield). Suppose, further, that an exhaustive privilege review had been conducted on the 30,000 documents coded as responsive and that that review had ascertained that 24,000 of the 30,000 were actually responsive (i.e., the privilege review had found that the original retrieval effort had attained 80% precision).[2] We now want to design a test that will help us determine whether or not the retrieval effort had met some minimum level of recall

In order to see our options, we need to specify values for four parameters:

1. The level of recall at or above which, if that is what the retrieval effort actually achieved, we would like to have a strong probability of passing the test;
2. The minimum probability with which we would like to pass at parameter (1);
3. The level of recall at or below which, if that is what the retrieval effort actually achieved, we would like to have a strong probability of failing the test; and
4. The maximum probability with which we would like to pass at parameter (3).

Parameters (1) and (2) are used to guard against failing the test, simply due to sampling error, when the retrieval effort has in fact achieved high recall. Parameters (3) and (4) are used to guard against passing the test, simply due to sampling error, when the retrieval effort has in fact achieved low recall.

Table 1 shows the test designs that result from five different combinations of settings for these four parameters. For each design, the table shows: a plan ID, data on the specified "should-pass" point (i.e., the values specified for parameters (1) and (2) above (along with a translation of the specified recall level to the corresponding density of responsive documents in the unretrieved set)), data on the specified "should-fail" point (i.e., the values specified for parameters (3) and (4) above (along with a translation of the specified recall level to the corresponding density of responsive documents in the unretrieved set)), and the sampling specification that results from these settings of the input parameters (sample size and the maximum number of responsive documents that can reside in the sample if a passing result is to be attained).

As can be seen from the table, variation of the input specifications can result can result in significant variation in the size of sample required for a validation test.

Under Plan 1, if our retrieval efforts have achieved 85% recall or better, we are very likely to pass the test (i.e., at least 95% of the time). If our recall efforts have achieved less than 80% recall, we are very likely to fail the test (again, at least 95% of the time); put in other words, this means that a passing result will allow us to state, with 95% confidence, that our retrieval efforts have achieved 80% recall or better. The sample size required to meet these specifications is large. We must draw and review a sample of 17,153 documents from the unretrieved set; we pass if the review finds 89 or fewer responsive documents in the sample; we fail if the review finds 90 or more responsive documents in the sample. This plan tightly constrains the scope of sampling error: the

[2]Note that we could design a test either with a known precision value (resulting from exhaustive review of the retrieved set) or with an estimated precision value (resulting from a review of a sample of the retrieved set); for the current example, in order to keep the discussion simpler, we have assumed the former scenario.

Plan	Should-Pass Point			Should-Fail Point			Sample Spec	
	Recall	Density	Prob Pass	Recall	Density	Prob Pass	Size	Max R
1	0.85	0.0044	0.95	0.80	0.0062	0.05	**17,153**	**89**
2	0.90	0.0027	0.95	0.80	0.0062	0.05	**3,925**	**16**
3	0.90	0.0027	0.95	0.80	0.0062	0.10	**3,062**	**13**
4	0.90	0.0027	0.95	0.75	0.0082	0.05	**1,901**	**9**
5	0.95	0.0013	0.90	0.75	0.0082	0.10	**644**	**2**

Table 1: Illustration of design options (coded R = 0.03; precision = 0.80).

band within which the outcome of the test is more or less unpredictable is narrow (between 80% and 85% recall). The plan does so, however, at a high cost in terms of sample size.

Under Plan 2, we increase the distance between the "should-pass" point and "should-fail" point and, by doing so, realize a significant reduction in the size of sample required. Under this plan, if our retrieval efforts have achieved 90% recall or better, we are very likely to pass the test (i.e., at least 95% of the time). If our recall efforts have achieved less than 80% recall, we are very likely to fail the test (again, at least 95% of the time); as with Plan 1, a passing result will allow us to state, with 95% confidence, that our retrieval efforts have achieved 80% recall or better. Under Plan 2, we must draw and review a sample of 3,925 documents from the unretrieved set; we pass if the review finds 16 or fewer responsive documents in the sample; we fail if the review finds 17 or more responsive documents in the sample.

Plans 3 through 5 further vary the input parameters in order to realize additional reductions in sample size. Plan 3 increases the probability of passing at the minimum recall threshold (again 80%) from 5% to 10%, meaning that a passing result will again allow us to state that our retrieval efforts have achieved 80% recall or better, but it will allow us to do so with a lower level of confidence (90%); this plan requires a sample of 3,062 documents. Plan 4 lowers the minimum recall threshold to 75%, but returns the probability of passing at that point to 5%, meaning that a passing result will allow us to state, with 95% confidence, that our retrieval efforts have achieved 75% recall or better; this plan requires a sample of 1,901 documents. Plan 5 further loosens constraints on sampling error, increasing the distance between the "should-pass" point and the "should-fail" point, decreasing the probability of passing at the former point, and increasing the probability of passing at the latter point; this plan requires a sample of just 644 documents. While Plan 5 is looser, a passing result on the test will still provide meaningful empirical validation of the effectiveness of our retrieval efforts: a passing result means that a passing result will allow us to state, with 95% confidence, that our retrieval efforts have achieved 75% recall or better.

Each of the five designs considered could yield meaningful results in the scenario under consideration (3% of the population coded as responsive at a precision level of 80%). Each test design, however, is sensitive to different levels of recall, places different controls on sampling error, and requires different sample sizes. It is up to the practitioner to decide which of these (or some other variant) will be optimal in the specific circumstances for which the test is designed.

Figure 1 shows the implications of the designs we have been considering. The chart shows, for each test design, the probability of realizing a passing result (represented on the vertical axis) at any given level of actual recall achieved by the retrieval effort (represented on the horizontal axis).

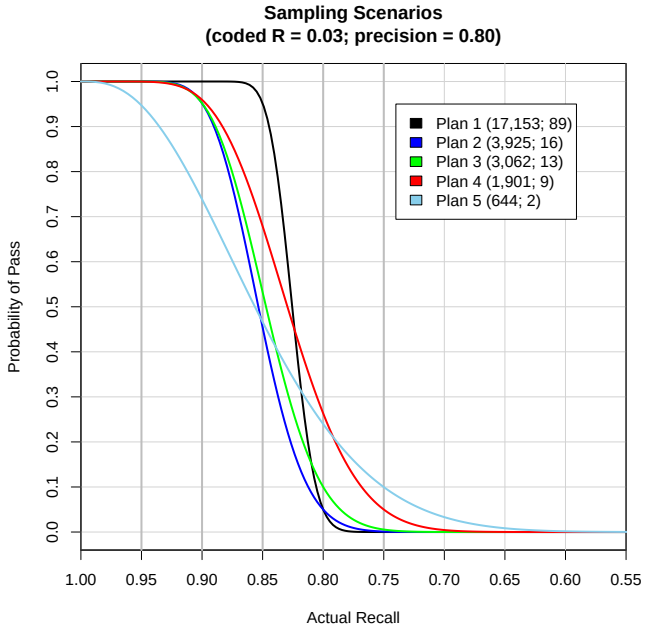

Figure 1: Implications of design options.

We have taken this brief look at the application of the acceptance testing paradigm to the validation of retrieval results, not in order to argue that this approach to validation should be adopted in all circumstances, but simply in order to show that, if one weighs carefully the information one really needs, the tolerance one has for sampling error, and the budget one has for sampling and review, one can almost always arrive at a sampling design that suits one's objectives and constraints. Objections to measurement on the basis of the time and cost entailed are simply not valid in the majority of cases.

In fact, a well thought out and executed measurement program can often lead to reductions in the time and cost required by a retrieval effort. Accurate real-time measurements can provide a project team with valuable information as to where improvements do and do not need to be made, thus allowing the team to reach its objectives more efficiently than it would be able to without that information.

Now, this is not to say that there may not be some, relatively infrequent, cases in which sampling simply cannot generate any useful information in a cost-effective manner (e.g., cases in which the prevalence of relevant material is

extremely low). We would argue that, even in such cases, there are often options for gathering at least some meaningful empirical information from at least some parts of the population, and some empirical data are better than no empirical data. The fact that such cases exist, however, does not argue against the inclusion of a measurement provision in an e-discovery standard. A requirement that a provider of e-discovery products and services have a capability that it can and should apply in the vast majority of cases is a reasonable, and essential, element in a meaningful standard.

4.3.4 *Opportunity to game the system*

Some might argue that an undue focus on quantitative measures may open the door to "gaming the system." If the validation of the results of a review or retrieval effort rests solely upon a number, a party might seek bad-faith methods of arriving at that number that are not in keeping with its true obligations.

To this objection we note, first, that, as noted earlier, we do not argue for the specification of minimum quantitative thresholds in a standard. We believe that practitioners, taking into consideration the requirements and conditions specific to a given matter, are best placed to decide on the role quantitative measures will play and on the levels expected. Second, we note that the statistical estimation of recall and precision is just one element in the validation of an e-discovery process. Sound validation also includes assessment of the protocol whereby the measures were obtained as well as consideration of non-statistical quantitative and qualitative aspects of the review or retrieval effort. When validation is viewed as a comprehensive exercise, in which statistical estimation of recall and precision is just one component, the opportunity to game the system via statistical legerdemain is significantly diminished.

4.4 Incorporation into a standard

We believe that a measurement provision, like that which we have described in this paper, could readily be incorporated into a standard of either the ISO 9000 or the ISO/IEC 27000 variety; an ISO 9000 standard might be the most appropriate home, however, given the focus of that family of standards on quality management systems.

We also note that we believe that an e-discovery standard, whatever the ISO family to which it belongs, must be of the "requirements" sort, if it is to be truly meaningful. We recognize that an initial "guidance" standard may be necessary as an instrument for building and validating consensus around the standard's provisions. It is only with the development of the standard into a set of certifiable requirements, however, that the potential benefits of a standard will be realized.

5. CONCLUDING REMARKS

We believe that an e-discovery standard could bring considerable benefits to the legal profession and the e-discovery industry. We believe, however, that a standard will bring those benefits only if it requires that those executing the activities covered by the standard demonstrate that they indeed have the capabilities essential to the effective execution of those activities. In e-discovery, one of those capabilities is the measurement of the effectiveness of the review/retrieval function. We believe that an e-discovery standard should require that a provider demonstrate that it has this essential capability.

6. ACKNOWLEDGMENTS

We would like to thank the organizers of the DESI V workshop for providing a forum at which we could express our views and hear those of others interested in these topics.

7. REFERENCES

[1] DESI IV – ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings, 2011. `http://www.umiacs.umd.edu/~oard/desi4/`.

[2] DESI V – ICAIL 2013 Workshop on Standards for Using Predictive Coding, Machine Learning, and Other Advanced Search and Review Methods in E-Discovery, 2013. `http://www.umiacs.umd.edu/~oard/desi5/`.

[3] TREC Legal Track Home Page. `http://trec-legal.umiacs.umd.edu/`.

[4] Kleen Products, LLC v. Packaging Corp. of America, 10 C 5711 (N.D. Ill.) (Nolan, M.J.).

[5] United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008) (Facciola, M.J.).

[6] International Organization for Standardization. Quality management systems – fundamentals and vocabulary, 2005. ISO 9000:2005.

[7] International Organization for Standardization. Quality management principles, 2012. `http://www.iso.org/iso/qmp_2012.pdf`.

[8] International Organization for Standardization and International Electrotechnical Commission. Information technology – security techniques – information security management systems – overview and vocabulary, 2012. ISO/IEC 27000:2012.

[9] A. Kiermeier. Visualizing and Assessing Acceptance Sampling Plans: The R Package AcceptanceSampling. *Journal of Statistical Software*, 26(6), 2008.

[10] C. Knox and S. Dawson. ISO 9001: A foundation for e-discovery. In *Proceedings of DESI IV: The ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery*, 2011.

[11] D. Oard, J. Baron, B. Hedin, D. Lewis, and S. Tomlinson. Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law*, 18:347–386, 2010.

[12] N. M. Pace and L. Zakaras. *Where the Money Goes – Understanding Litigant Expenditures for Producing Electronic Discovery*. RAND Institute for Civil Justice, 2012.

[13] E. G. Schilling. *Acceptance Sampling in Quality Control*. ASQC Quality Press, 1982.

[14] S. Teppler. International standard project for e-discovery approved. *Law Technology News*, April 2013.