gTLD APPLICATION PROCESSING: INITIAL EVALUATION

QUALITY PROGRAM REPORT

26 August 2014

FOR PUBLIC RELEASE

TABLE OF CONTENTS

1 Summary 1
1.1 Program Coverage 3
1.2 Program Scope 4
1.3 Roles and Responsibilities 4
2 Program Objectives 5
3 Content Reviews 6
3.1 Process and Sampling 6
3.2 Roles and Responsibilities 6
3.3 Exceptions 6
3.4 Metrics and Reporting 6
4 Blind Content Inspections 7
4.1 Process and Sampling 7
4.2 Metrics 7
4.3 Roles and Responsibilities 8
4.4 Exceptions 8
4.5 Results 8
4.6 Analysis and Discussion 10
5 Blind Procedural Inspections 12
5.1 Process and Sampling 12
5.2 Metrics 13
5.3 Roles and Responsibilities 13
5.4 Exceptions 13
5.5 Results 13
5.6 Analysis and Discussion 13
6 Analytical System Review 14
6.1 Process 14
6.2 Analysis and Discussion 15
7 Overall Analysis, Discussion, and Recommendations 16

1 Summary

New gTLD application evaluation was a labor-intensive business process performed by multiple vendors and hundreds of individuals on a global basis. Initial Evaluation (IE) included seven distinct evaluation types: applicant background, financial capability, technical/operational capability, registry services, geographic names, DNS stability, and string similarity. For commercial and practical reasons, including application volume and handling conflicts of interest between an applicant and evaluator, multiple evaluator firms were contracted. Application evaluation was performed against detailed criteria as published in the *New gTLD Applicant Guidebook* (AGB).[1] Quality and consistency of evaluation across all applications and all evaluator firms was a key business requirement for ICANN. Given the importance of demonstrable quality, 50% of the applications were subject to quality sampling in some capacity and 100% of the applications were reviewed using analytical techniques. All application data was subject to a suite of manual and automated data consistency checks performed by ICANN staff and JAS.

At a high level, the new gTLD application evaluation training and quality program was designed to both *improve* and *measure*:

- **Consistency/Precision:** a measure of the degree of agreement between independent assessments of a particular sample. Precision is expressed in terms of the standard deviation of the consistency rating among primary and independent half-blind *de novo* assessments (calculation of the consistency rating is described in Section 5.2). Precision is important because multiple evaluator firms should produce similar results given similar applications. Situations where precision was not as expected triggered additional training, documentation, and may inform future process revisions.
- **Accuracy:** a measure of the degree of agreement of a sample with an accepted reference. In the case of application evaluation, the accepted reference is the result of "work-out" conferences between the primary evaluator firm, the quality firm, and ICANN when discrepancies occur. Accuracy is expressed in terms of percent of the samples reflecting the expected value. Situations where accuracy was not as expected triggered additional training, documentation, and may inform future process revisions.
- **Process Fidelity:** a measure of the alignment between the expected process per the vendor's contract and the actual process performed for a given application. Process fidelity is expressed in terms of a percent of the samples where a post-evaluation Procedural Inspection indicated that proper procedures were followed.

As quality measurement and improvement are typically somewhat competing goals (performing quality improvement on a process while measurement is occurring leads to a degree of Heisenberg uncertainty), the overall quality program was designed primarily to monitor, incent, and improve quality during evaluation with a secondary objective of providing analysis and a quantitative baseline to assess the process in arrears and inform future rounds.

[1] *New gTLD Applicant Guidebook*, ICANN, 4 June 2012, http://newgtlds.icann.org/en/applicants/agb

The training and quality program is comprised of six functions:

Unified Training

A unified, cross-firm approach to training was developed and implemented prior to the commencement of production evaluation. Unified training was essential in bringing together the evaluation operations of all evaluator firms – particularly the large-scale operations of the three technical/operational and financial firms – and maintaining ongoing alignment in a challenging and dynamic environment.

For technical/operational and financial panels – the most complex evaluations – all three evaluator firms shared training materials and conducted joint training sessions. For other panels, standardized training templates were utilized.

Content Reviews

Content Reviews were discussions between two or more evaluator firms that had completed a full or partial review of the same application. Content Reviews were designed to improve consistency/precision and accuracy among the three technical/operational and financial evaluator firms. Content Reviews of selected applications were performed as a part of the comprehensive training program prior to commencement of production evaluation and additionally throughout Initial Evaluation to maintain communication and alignment between all three evaluator firms. One special case of content reviews was the applicant-facing Clarifying Question (CQ) pilot that provided immense value. Of the 1917 application IDs receiving Prioritization Draw results, 107 applications were involved in a complete or partial content review at some point.

Blind Content Inspections

Content Inspections were half-blind independent evaluation and scoring of a randomly selected set of applications. The Content Inspection included review of the primary evaluator firm's Clarifying Questions (CQs) prior to issuance, and independently generated final scoring by the quality evaluator firm. Blind Content Inspections were designed to measure and improve consistency/precision and accuracy among the three technical/operational and financial panel firms. The inspections were half-blind in that the primary panel firm did not know in advance which applications were selected for inspection and the quality firm was not aware of the primary firm's scores in advance. Content Inspections were conducted on a randomly selected 15% of the 1917 application IDs receiving Prioritization Draw results.

Blind Procedural Inspections

Procedural Inspections were half-blind reviews of the primary firm's records to gain confidence that the agreed-upon processes and procedures were performed as expected. Procedural Inspections were designed to measure the process fidelity of the panel firms. The inspections were blind in that the primary panel firm did not know in advance which applications were selected for inspection. Procedural Inspections were conducted on a randomly selected 35% of the 1917 application IDs receiving Prioritization Draw results.

Analytics

ICANN received in excess of 1900 applications, largely comprised of unstructured text and attachments. Many latent similarities existed between the applications due to common applicants, consultants, and service providers. Analytical tools were developed to highlight these latent similarities and improve confidence that applications with similar content received a similar final disposition. Moreover, in excess of 5000 Clarifying Questions (CQs) were generated as a part of evaluation; as CQ generation is labor-intensive and subject to a range of error modalities, analytical systems provided automated quality and content checks of CQs prior to issuance.

Data Consistency Checks

Application evaluation was a large-scale global operation with a number of dynamic components. Ensuring that ICANN's systems of record were both internally consistent and accurately reflective of the authoritative evaluation results as documented in numerous vendor reports was critical. Automated systems provided routine data validation and crosschecking spanning numerous systems and record types to reduce likelihood of consistency errors.

1.1 Program Coverage

While designing training and quality programs, the process of application evaluation was divided into *content* and *process* components. The process components covered each vendor's obligation to perform their contracted duties and interact with the broader system and ICANN as specified, and the general requirement to maintain data consistency across several systems given emergent and fast-moving processes. The content components covered each vendor's obligation to evaluate the application pursuant to the Applicant Guidebook and all relevant guidance. The training and quality program recognized and provided coverage to both of these at multiple points in time during application processing.

Content-oriented aspects of the training and quality program were focused on the technical/operational and financial panel types due to the nature of these evaluations and the complexity and scale of the combined evaluation operations of all three evaluator firms. For all panel types, the process-oriented aspects of the quality program were focused on ensuring that all evaluator panels followed procedures agreed upon with ICANN.

Panel Type	Prior to CQ Release	Final Scoring (IE)	
	Content	Content	Process
Financial	Training Content Review Blind Content Inspection Analytics	Ongoing Training & Communication Content Review Blind Content Inspection Analytics	Training Blind Procedural Inspection Data Consistency Checks
Technical/Operational	Training Content Review Blind Content Inspection Analytics	Ongoing Training & Communication Content Review Blind Content Inspection Analytics	Training Blind Procedural Inspection Data Consistency Checks
Registry Services	Training	Analytics	Training Blind Procedural Inspection Data Consistency Checks
DNS Stability	Training		Training Blind Procedural Inspection Data Consistency Checks
String Contention	Training		Training Blind Procedural Inspection Data Consistency Checks
Geographic	Training		Training Blind Procedural Inspection Data Consistency Checks

Table 1: Training and Quality Program Coverage

1.2 Program Scope

The training and quality programs were operational prior to the commencement of production evaluation and continued through the completion of Initial Evaluation. Extended Evaluation was not included in the scope of the quality program.

1.3 Roles and Responsibilities

JAS Global Advisors LLC ("JAS") was responsible for designing the overall training and quality programs based on requirements developed with ICANN. JAS was responsible for administering the quality program during execution, coordinating content reviews, performing Content Inspections, performing Procedural Inspections, implementing analytical and consistency checking systems, and reporting results. JAS was the primary technical/operational and financial reviewer for fewer than 50 applications and only in situations where no other technical/operational and financial firms were available due to a conflict of interest with the applicant. Related to the training and quality programs, all evaluator firms had obligations to provide data, participate in training activities, produce documentation, and generally cooperate with training and quality activities.

2 Program Objectives

The training and quality program was designed to achieve multiple objectives. The most important objective was to provide confidence that applications with similar content received a similar final pass/fail disposition. It's important to note that with respect to scoring, the quality program viewed Initial Evaluation as a pass/fail exercise consistent with the description in the Applicant Guidebook. No meaning is or should be imparted to numerical differences in score between two passing (or two failing) applications.

To achieve this objective, training and quality programs focused on:

- Upfront "calibration" among evaluator firms via unified training, discussion, scoring exercises, and pilots;
- Encouraging and maintaining ongoing communication among evaluator firms throughout the process via training, scoring exercises, and comparison of evaluation results;
- Leveraging analytics to identify latent similarities and determine potential scoring inconsistencies; and
- Providing visibility and early notification to ICANN in the event inconsistencies were discovered.

Clearly, *communication* and *visibility* are the central themes. Given the scale and nature of evaluation, absent active mechanisms to maintain communication between firms and with ICANN, there was a risk that evaluator firms would become isolated and produce increasingly divergent results over time. A central objective was to maintain open communication among all participants during the entire evaluation process.

A second central objective was to provide ICANN visibility into evaluation quality throughout the evaluation time period. Absent active mechanisms to assess quality during evaluation, it would be hard to quickly determine if quality was acceptable or unacceptable, converging or diverging, or if process improvements or additional training was required, leading to a sort of unmanaged Markov process.

By creating active communication and visibility mechanisms, ICANN was able to successfully keep the evaluation process under control.

Additionally, the program had the following secondary objectives:

- Improve quality of issued CQs
- Reduce data and clerical errors
- Provide quantitative baseline for future rounds

3 Content Reviews

Content Reviews were discussions between two or more firms that had completed a full or partial review of the same application. Content Reviews were designed to improve consistency/precision and accuracy among the three technical/operational and financial evaluator firms.

Content Reviews were performed early in the process – during training and early in Initial Evaluation – in order to add maximum value to the calibration process; subsequent and less frequent Content Reviews were performed throughout Initial Evaluation to encourage continued communication and alignment, particularly around emergent issues. Content Reviews were performed on technical/operational and financial panel results.

One special case of content reviews was the applicant-facing Clarifying Question (CQ) pilot that provided immense value; multiple pilots that were not applicant-facing were also conducted.

3.1 Process and Sampling

Content Reviews leveraged approximately 107 applications that both a primary reviewer and a secondary reviewer had evaluated (in part or in full) in some capacity. An effort was made to select applications for Content Review that represented a wide range of applicants and service providers to maximize the value of the exercise. Applications utilized for Content Reviews were not eligible for selection for Content Inspection.

3.2 Roles and Responsibilities

JAS coordinated Content Review activities among the three technical/operational and financial evaluator firms. Prior to the availability of actual applicant data, JAS developed several mock applications as a part of the training materials.

3.3 Exceptions

Differences in scoring were discussed and remediated between the evaluator firms with input from ICANN requested on an as-needed basis.

3.4 Metrics and Reporting

The primary objective was to facilitate calibration and maintain communication; the Content Review program did not generate metrics.

4 Blind Content Inspections

A statistically relevant number of technical/operational and financial evaluations were subject to half-blind Content Inspection reviews performed on a *de novo* basis. A *de novo* review is a complete and independent review performed "from the beginning" by the quality firm simultaneously with – but independently from – the primary evaluator firm. The review is also half-blind; the primary evaluator firm did not know in advance which applications were selected for Content Inspection. The intent of the review was to measure CQ and scoring consistency and accuracy against scoring guidance and training, and to provide an opportunity to quickly detect quality and consistency issues.

4.1 Process and Sampling

Blind Content Inspections were selected via random ordering of the 1917 application IDs receiving Prioritization Draw results. JAS performed the random ordering via computer on 20 Dec 2012. Note that withdrawals reduced the size of the population, requiring limited selection of additional samples to compensate for the aforementioned issues. The first 15% (288) applications in the random ordering were selected for Content Inspection. As additional samples were needed due to withdrawals or other factors requiring de-sampling, applications starting at 289 in the random ordering were selected.

Final metrics for the quality control program were taken on 28 August 2013 at the conclusion of Initial Evaluation work and are as follows:

Total Active Applications (28 Aug 2013)	1768
Applications Sampled	274
Sampled Proportion	15.50%

Table 2: Content Inspection Sampling

4.2 Metrics

The blind Content Inspections produced the following quantitative metrics:

- **Consistency Rating (per question).** This is the simple numeric pairwise comparison between the primary and QC review final scores on a per question basis. A pairwise comparison of 0 indicates that the primary and QC review final scores are identical whereas a pairwise comparison of +1 or -1 indicates the final scores differ. Instances of non-objection were de-sampled (see below).

  For the purpose of QC, no distinction is made between passing scores with score = 1 and score > 1. Any score greater than or equal to 1 will be considered a 1 for the purpose of QC – for both the primary firm score and the QC firm score. For example, a score of 2 is equal to a score of 1 and to a score of 3 – all were transformed to a score of 1 prior to calculation of the consistency rating. This transformation is necessary to align the QC program with the pass/fail design of Initial Evaluation as described in the Applicant Guidebook.

- **Consistency Rating (per application).** This is a proportional measure of consistency of final (pass/fail) dispositions for a given application. The quality evaluator firm maintained the option to deem an application "non-objection" meaning that for reasons related to maintaining the

integrity of the half-blind selection, not enough information was available to score the application but the quality evaluator firm did not find sufficient cause to disagree with the primary firm's pass/fail disposition.

4.3 Roles and Responsibilities

JAS was the quality evaluator firm. If an application was selected for Content Inspection where JAS was the Primary Review Firm (due to conflict with both primary evaluator firms), the application was de-sampled for quality control purposes and the next application in the random ordering that had not already been released was selected.

JAS' small number of primary evaluations were therefore ineligible for Content Inspection; however, as JAS was a party to each and every consistency rating metric, evaluation of JAS' performance as compared to the other firms was evident and obvious.

4.4 Exceptions

Differences in scoring appear in the consistency rating; exceptions were brought to ICANN's attention as soon as they were discovered for discussion with the evaluator firms as necessary.

4.5 Results

Content Inspections generated metrics on a horizontal basis (per question across applications) and on a per-application basis. Content Inspection samples were taken before and after the Outreach phase. Outreach was an ICANN process that in limited situations allowed the applicant to provide missing information that may have stemmed from an oversight.

Shown below are statistics describing the Content Inspection samples taken prior to Outreach; following Outreach, all primary and Content Inspection evaluations were in agreement (consistency rating = 0). Small variances in the sample size in the table below occurred because in certain limited circumstances the quality firm asserted "non-objection" discrepancies as described above and those individual questions were de-sampled for statistical purposes.

In summary, prior to the Outreach phase there were six individual application question/response instances (1 technical/operational and 5 financial) where a bona-fide scoring discrepancy existed that would have impacted the final disposition of the application (moving an application from a pass to a fail or vice versa). To highlight root causes, for purposes of this analysis and presentation, a single scoring issue that cascaded into multiple scoring discrepancies has been reduced to the single root cause and the cascading discrepancies are not reflected here. For example, a discrepancy in financial cost calculations may cascade into a discrepancy in the question 50 Continuation of Operations (COI) Instrument calculation; the former is indicative of a root cause quality issue whereas the latter is not.

Applications containing a question that received a zero score following the Clarifying Question phase proceeded to the Outreach phase. All of the per-question discrepancies below were resolved during Outreach; following Outreach, all primary and Content Inspection evaluations were in agreement and every question selected for Content Inspection received a passing (non-zero) score.

Question #	n where consistency rating = 0 (Consistent)	n where consistency rating != 0 (Not Consistent)	Standard Deviation of Consistency Rating for the Population
24	261	0	0.000
25	256	0	0.000
26	261	0	0.000
27	260	0	0.000
28	261	0	0.000
29	261	0	0.000
30	261	0	0.000
31	261	0	0.000
32	260	1	0.024
33	260	0	0.000
34	261	0	0.000
35	261	0	0.000
36	261	0	0.000
37	261	0	0.000
38	261	0	0.000
39	261	0	0.000
40	261	0	0.000
41	261	0	0.000
42	261	0	0.000
43	260	0	0.000
44	N/A – Optional	N/A – Optional	N/A – Optional
45	258	2	0.037
46	261	1	0.000
47	261	0	0.000
48	261	0	0.000
49	261	0	0.000
50	256	2	0.041

Table 3: Per-Question Consistency Rating

An application must have no individually failing questions (score=0) and reach a minimum score threshold in both technical/operational and financial questions in order to pass evaluation. As an application with all passing individual questions may still fail due to insufficient total points, consistency was also analyzed on a per-application basis to capture this aspect.

In summary, prior to the Outreach phase there were five (5) applications where a bona-fide scoring discrepancy existed that would have impacted the final disposition of the application (moving an application from a pass to a fail or vice versa).

Note that this analysis is considering an application as a whole whereas the previous analysis is considering all question/response instances. In the former, there were six (6) question/response instances where the consistency rating was not zero; in the later, there were five (5) whole applications where the final disposition was not consistent pre-Outreach. All inconsistencies were resolved Post Outreach.

Application Status	n	%
Consistent Pre-Outreach	261	95.26%
Not Consistent Pre-Outreach	5	1.82%
No Objection	8	2.92%
Consistent Post Outreach	274	100.00%

Table 4: Per-Application Consistency Rating

Analyzing the five (5) instances where there was a scoring discrepancy prior to Outreach on a per-evaluator firm basis revealed balanced data (note that aliases are used to identify evaluator firms):

Status	n
Evaluator Firm Alpha consistency rating as compared to quality firm is > 0 (Evaluator Firm Alpha scored **higher** than quality firm)	1
Evaluator Firm Alpha consistency rating as compared to quality firm is < 0 (Evaluator Firm Alpha scored **lower** than quality firm)	2
Evaluator Firm Bravo consistency rating as compared to quality firm is > 0 (Evaluator Firm Bravo scored **higher** than quality firm)	0
Evaluator Firm Bravo consistency rating as compared to quality firm is < 0 (Evaluator Firm Bravo scored **lower** than quality firm)	2

Table 5: Per Evaluator Firm Analysis of Application Discrepancies

4.6 Analysis and Discussion

Given the overall scale, scope, and challenge of Initial Evaluation, evaluation was remarkably consistent. Several points are worth noting:

- Evaluator firms spent considerable effort in training and calibration, and clearly it proved effective. The Applicant Guidebook describes Initial Evaluation as a pass/fail exercise (as long as the minimum point requirements are met, there is no benefit in receiving additional points and no penalty in receiving fewer points). As such, during initial training and calibration, evaluator firms focused on "zero/non-zero" issues/scoring to gain confidence that pass/fail alignment

would be high. As a result, pass/fail consistency was very high but raw numeric scoring – which included the additional points – was less consistent. Analysis of the additional point system beyond the minimum pass/fail thresholds was not a part of the design of the quality program.

- Consistency of CQs was desirable but not always possible. Variance in internal firm processes and other factors reduced the overall consistency of CQs. However, pass/fail application disposition remained high despite variance in CQs. A contributing factor is that a significant proportion of CQ inconsistencies were related to additional points components of questions (criteria required to receive a score of two (2) or three (3) on a question).
- Consistency issues are highly concentrated in very few questions, particularly financial questions 45 and 50. Anyone familiar with the application process will recognize these questions and not be at all surprised with this finding. The fact that these questions were the subject of the majority of post-AGB ICANN guidance – both to applicants and evaluators – underscores the localized difficulties present in these two questions. Discrepancies that surfaced in questions 45 and 50 tended to be systemic issues (symptoms of unanticipated scenarios and/or broader lack of clarity) whereas the discrepancies that surfaced in other questions tended to be isolated and unusual corner cases.
- Numerous subjective terms (such as “adequate,” “commensurate,” “comprehensive,” “highly developed,” and similar terms) appear frequently in the Applicant Guidebook. Evaluator firms and ICANN spent significant effort defining these terms crisply and calibrating for the purpose of consistent evaluation. While the results show that this effort was largely successful, additional definition of subjective terms in future revisions of the Applicant Guidebook would be of value.
- The Applicant Guidebook did not recognize the concept of a Registry Service Provider nor did it contemplate an applicant describing a registry being run as a cost center with limited or no revenue. Ambiguity surrounding these concepts was the root cause of several calibration discussions and scoring discrepancies. Overt recognition of these concepts in future revisions of the Applicant Guidebook would be of value.

5 Blind Procedural Inspections

Work performed by technical/operational, financial, string similarity, and geographic name panels/providers was subject to a Procedural Inspection on a statistically relevant randomly selected sample of applications. The intent of the Procedural Inspection was to provide assurance that the application was fully processed, and that all panel providers completed (and provided evidence of completing) all the steps required of them as documented in the Applicant Guidebook and individual SOWs. A team of JAS personnel conducted the Procedural Inspections.

Each of the five panel types had a "procedural checklist" which was developed by ICANN and the panel providers in advance. Multiple firms performing the same function (e.g. financial review) used the same procedural checklist. The procedural checklist was the basis on which the Procedural Inspections were conducted.

5.1 Process and Sampling

Blind Procedural Inspections were selected via random ordering of the 1917 application IDs receiving Prioritization Draw results. The first 35% (671) applications in the random ordering were selected for Procedural Inspection; if additional samples were needed due to withdrawals, selection of an application where the applicant is conflicted with both primary evaluator firms, or other factor requiring de-sampling, applications starting at 672 in the random ordering were selected. Each selected application was subjected to a Procedural Inspection for all panel types. Note that the random ordering generated for Procedural Inspections was different – and independent – from the random ordering generated for Content Inspections.

Procedural Inspections were conducted on final work products after final scoring was submitted to ICANN.

Final metrics for the quality control program were taken on 28 August 2013 and are as follows:

Total Active Applications (28 Aug 2013)	1768
Applications Sampled	639
Sampled Proportion	36.14%
Compliance Rate	99.84%

Table 6: Procedural Inspection Sampling

As the String Similarity panel operated on unique strings, a separate random ordering and selection were performed for these Procedural Inspections. Content Inspection metrics for String Similarity are as follows:

Unique Strings (28 Aug 2013)	1388
Applications Sampled	490
Sampled Proportion	35.30%
Compliance Rate	100.00%

Table 7: String Similarity Procedural Inspection Sampling

5.2 Metrics

Each Procedural Inspection reviewed the primary evaluation as a whole and generated one metric per application. The resulting metric is an assessment of the fidelity with which the primary evaluation followed the agreed-upon Procedural Checklist for the specific application. The metric is one of: *Compliant (C); Minor Discrepancy (MD); Significant Discrepancy (SD).*

5.3 Roles and Responsibilities

JAS was the quality evaluator firm. If an application was selected for Procedural Inspection where JAS was the Primary Review Firm (due to conflict with both primary evaluator firms), the application was de-sampled for quality control purposes and the next application in the random ordering that had not already been released was selected.

5.4 Exceptions

Exceptions were brought to ICANN's attention as soon as they were discovered for discussion with the evaluator firms as necessary.

5.5 Results

Procedural Inspections generated metrics on a per-evaluator firm basis for each evaluation type. One sample was taken after the primary evaluator firm submitted final results for an application that was selected for Procedural Inspection.

Evaluation Type	Evaluator Firm (alias)	n Compliant	n Minor Discrepancy	n Significant Discrepancy
Technical/Operational	Charlie	329	1	0
Technical/Operational	Delta	309	0	0
Financial	Charlie	329	1	0
Financial	Delta	309	0	0
Geographic	Echo	399	0	0
Geographic	Foxtrot	240	0	0
DNS Stability	Golf	639	0	0
Registry Services	Lima	639	0	0
String Similarity[2]	Oscar	490	0	0

Table 8: Per Evaluator Firm Analysis of Procedural Inspections

5.6 Analysis and Discussion

Each evaluation vendor's adherence to agreed-upon evaluation procedures was a critical success factor for the program. Procedural Inspection results show that this adherence did indeed occur.

[2] Note that String Similarity Procedural Inspections were performed on 490 evaluations based on applications for 1388 unique strings.

6 Analytical System Review

ICANN received in excess of 1900 applications, largely comprised of unstructured text and attachments. Many latent similarities existed between the applications due to common applicants, consultants, and service providers. Analytical tools were developed to achieve three objectives:

- Provide confidence that all similar applications received similar final (pass/fail) dispositions;
- Help identify potential CQ inconsistencies that could lead to a discrepancy in final disposition;
- Improve the quality of CQs by programmatically checking application and Applicant Guidebook citations.

While the previously described quality procedures applied to a sample of applications, analytical techniques were performed on all applications and CQs.

The analytical system allowed the evaluator firms, quality firm, and ICANN to visually review connections between similar applications, the CQs generated for those applications, the responses to those CQs from applicants, and the final score on an ongoing basis. While complete and absolute consistency through all of those steps would be a desirable – albeit Quixotic – outcome, in reality, analytics allowed discrepancies to be identified and reviewed for impact. Potentially problematic discrepancies were identified and rectified.

6.1 Process

Financial and technical/operator evaluator firms interacted with the analytical system at three points in time:

1. Following submission of CQs to ICANN's application management system (but prior to their transmission to the applicant);
2. Prior to submitting final scores to ICANN; and
3. Following submission of final scores to ICANN.

Following submission of CQs to ICANN's application management system, the analytical system programmatically matched quotes and citations appearing in the CQs to the relevant application and the Applicant Guidebook. Matches were confirmed and potential mismatches were flagged for manual verification. This step reduced the occurrence of misquotes and copy/paste errors given that thousands of similar CQs were generated. This was an especially important error mode to control, given that oft-quoted portions of the applications were confidential. Additionally, the analytical system compared the CQs for the submitted application to the CQs generated for similar applications and flagged discrepancies for manual verification.

Following submission of final scores to ICANN's application management system, the analytical system compared the scores of the submitted application to the scores of similar applications previously submitted. Potential discrepancies were flagged for manual verification.

Finally, at the completion of Initial Evaluation, JAS performed an analytical review of all applications that completed Initial Evaluation successfully vs. those that were referred to Extended Evaluation.

6.2 Analysis and Discussion

The sheer volume and unstructured nature of the application data necessitated an analytical approach. During each weekly application processing cycle, reports were delivered to evaluator firms and ICANN containing the results of the analytical reviews described above. As manual verification confirmed or refuted analytical results, false positives were identified and tuned out to improve future efficacy of the system. Noting that analytical reviews were a backstop measure designed to catch issues that remained undetected relatively late in the application cycle, a low and decreasing number of analytical system exceptions were indicative of high quality work by the evaluator firms. While there was an initial burst of analytical system exceptions, by the end of Initial Evaluation, very few valid analytical exceptions were being identified. This was an indication that the evaluation system was performing adequately and that the internal quality procedures being performed by each firm were effective. This was the desired behavior.

Following the completion of Initial Evaluation, JAS performed an analytical comparison of all applications that completed Initial Evaluation successfully vs. those that were referred to Extended Evaluation and found that the applications that were referred to Extended Evaluation were materially different than the applications that passed Initial Evaluation successfully. As this analysis took the entire population of applications into consideration, this step served as a valuable system-wide double-check on all of the previous sample-oriented quality programs.

Despite acknowledged inconsistencies in CQs and numeric scores (above and beyond the passing thresholds), this last analysis provided a strong indication that – when the process reached completion – all similar applications received passing scores and the applications referred to Extended Evaluation correctly were individual special cases requiring additional clarification.

7 Overall Analysis, Discussion, and Recommendations

The ICANN New gTLD evaluation program resulted in the successful evaluation of over 1900 applications from a full range of global applicants, delivering a demonstrably high level of evaluation consistency while providing ICANN with the practical and commercial benefits of evaluator depth and diversity. Some additional overall comments in closing:

1. The extensive advanced preparation, training, synchronization, and evaluation exercises (pilots) undertaken by the technical/operational and financial evaluator firms were essential and probably the single largest critical success factor. As verified by the positive quality program results, a unified approach to these activities coalesced the team and substantially mitigated the risk of isolation and inconsistent or divergent evaluations.
2. As quality practitioners well know, one value of a proactive quality program is that the mere (visible) existence of such a program helps incent the desired behaviors. In this case, it is highly probable that the existence of a visible and well-publicized proactive quality program properly incented all evaluation panel vendors to be appropriately cognizant of evaluation consistency, accuracy, and process fidelity, and perform accordingly.
3. Although the questions were provided in advance and there was an expectation that applicants would be clear on the material, it was apparent that many applicants, including sophisticated applicants, were confused as to how to respond to the questions. This resulted in two undesirable effects: (a) applicants tended to "over-respond" to the application, adding unnecessary volume and complexity; and (b) there was more effort put into clarification communications (including CQs) than was probably intended in the original vision. While not "providing the answers" there is an opportunity to make the application process more objective and deterministic for both applicants and evaluators. Reducing subjectivity of evaluation will enable improved quality and consistency and reduce costs associated with extensive synchronization activities.
4. The lack of structured application data was an impediment during evaluation; future application rounds should capture data in a more structured format, greatly facilitating evaluation, quality reviews, and subsequent processes like contracting.
5. Several questions, particularly technical/operational questions, have overlapping remits complicating evaluation, quality processes, and unnecessarily creating the appearance of inconsistency. Some topics, such as the use of IDNs, often have material spread throughout several questions. This makes it harder for applicants to "know what to put where" and for evaluators to find the information they're looking for. A highly structured application will help address this issue.
6. Releasing results incrementally opened the opportunity for difficult-to-manage inconsistencies. Future rounds designed for one release of results at the end will make comprehensive consistency and quality checking more effective.
7. The publication of detailed numeric scores confused and undermined the AGB-driven premise that evaluation was pass/fail. Inconsistencies in numeric scores incorrectly sent a message that evaluation was much more inconsistent than the final results and the quality programs assert.

Future application rounds should either publish results as pass/fail only, or re-calibrate the entire process to produce numerically consistent scores.

8. Financial evaluation of questions 45 and 50 exhibited systemic issues that made consistent evaluation difficult. Recognizing applicants that choose to run their registry as a cost center and revising the approach to the problematic question 50 regarding the Continuity of Operations Instrument will go a long way to increase the evaluation consistency of these questions.