Quepid testing for Elastic Search launch

FEB to NOV 2023

The better the question. The better the answer.
The better the world works.

Building a better working world

Relevance Tuning – Input Types

For Top queries

1 Expert Judgements

Gold Standards:

- Query/doc pairings per domain expert(s)
- Pre-tune search engine

2 Targeted Testing

Quepid and A/B Comparison tool:

- Evaluate results
- Non-binary scoring NDCG@10 and SQM

For All queries

3 User Testing

Search log analysis

- Subjective ratings
- Less structured
- More hands on / BAU

Relevance Testing Stages

Preparatory activities through pre launch user testing

Pre-migration / End of May

Gold Std Content Quepid Set up

- KM Excellence to provide:
  - Gold content for top 25 keywords by: **End of May**
  - Quepid populated for user testing with query types:
  - Content Category
  - Gold Standard Content
  - Service Line/Industry Sector
  - Topics
- Hybrid Testing for query types :
  - Names
  - Navigational

Baseline: 1 week/12 JUN

Validations

- Follow field mapping from SP to Elastic to AEM to Elastic
- Verify correct fields in use for:
  - Relevance Tuning
  - UX display
- DSL Query templates working as expected
- Index validation – before/after quantity comparisons, etc
- Baseline relevance scores in Quepid by query type
- Initial tuning to Gold Standard Content

Round 2: 1 week/19 JUN

Friends & Family

- Solicit user search ratings in Quepid
  - Ranking scores
  - By Query Types
- Relevance Tuning based on User Feedback
- Consultant advice enacted
- People Results Normalization
- Best Bets Integration
- Apply query boosts
  - Match Phrase
  - Endorsed / Flagship
  - Gaussian Decay

Round 3: 1 week /26 JUN

KMs & Sounding Board

- General User Testing in Production
  - Available metrics
  - Subjective ratings
- User feedback on People / Wayfinding queries
- Relevance Tuning based on User Feedback
- Consultant advice enacted
- Fields adjustments
- Re-indexing if needed

Search Relevance Tuning – Iterative & Cyclical Process

Use Cases by Category
- Methods, Credentials, Policies, etc.
- Data Fields & Settings

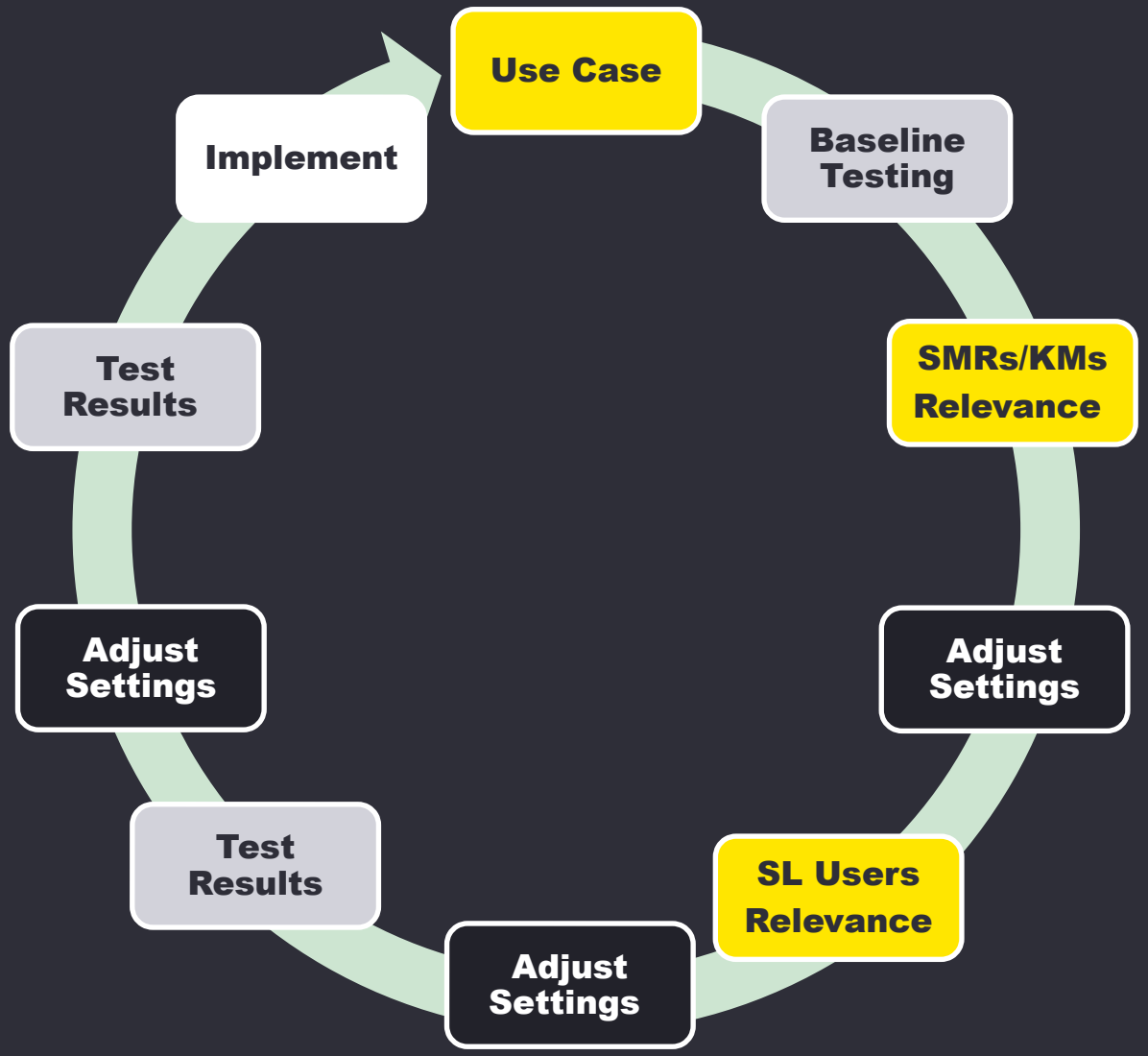
Baseline Testing
- Discover User Interface and EYSIGHT
- Initial hypothesis for best result relevance

SMR / Service Line User Testing
- Search Quality Metrics (SQM) ratings in Quepid
- Check between experienced knowledge & general impressions

Adjust DSL query settings
- Based upon user testing feedback and metrics
- Metadata field inclusions/boost adjustments as needed
- Or determine alternate query treatment required

Test Results – Repeat
- A vs. B Comparison tool - EYSIGHT
- Test result sets between multiple configurations

Implement
- Move changes to Production
- Monitor metrics for continued success
- Identify when new research/settings required



Relevance Tuning Priorities: Beta thru Release 2

Priority	Relevance challenge to address:	Resolution Options:
1	**Balancing result type ranking**– Credentials not interspersed. Documents, proposals, solutions, policies, methods.	Boosting with content stratification layers of content: Recency ,Endorsed, Flagship status
2	**People Results in All vertical** – People results NOT in top 96 Due to Elastic ranking scores increasing from DSL query boosts. Name based searches ok; expertise keywords 0 people results	Ranking score normalization for People data coming from SharePoint index. Expertise as separate search experience.
3	**Federation Logic** – Round Robin approach as short term solution? SP score normalization may not be sufficient.	Code obsurced; can NOT make further adjustments to DSL query that would impact interleaving.
4	**Recency decay** – pipeline stage code vs. coding into Domain Specific Language (DSL) query for Elastic	Future DSL query has stronger boost on newer content based on Gaussian decay for content age
5	**Phrase match** – evaluate strength of exact phrase match within federated results	Weigh benefits of higher recall vs. tighter precision for upcoming DSL query enhancements
6	**Extractive Summarization / UI** - extract is surfacing hits on single words vs. phrase. DSL scores on exact phrase but not apparent in UI	User impact TBD. Hit highlighting, extractive summarization snippet NOT under our code control.
7	**Partial name query enhancement** - Analyzers, N-grams	Pre req: People data indexed in Elastic

Discover Search Success Rates: Legacy thru Release 3

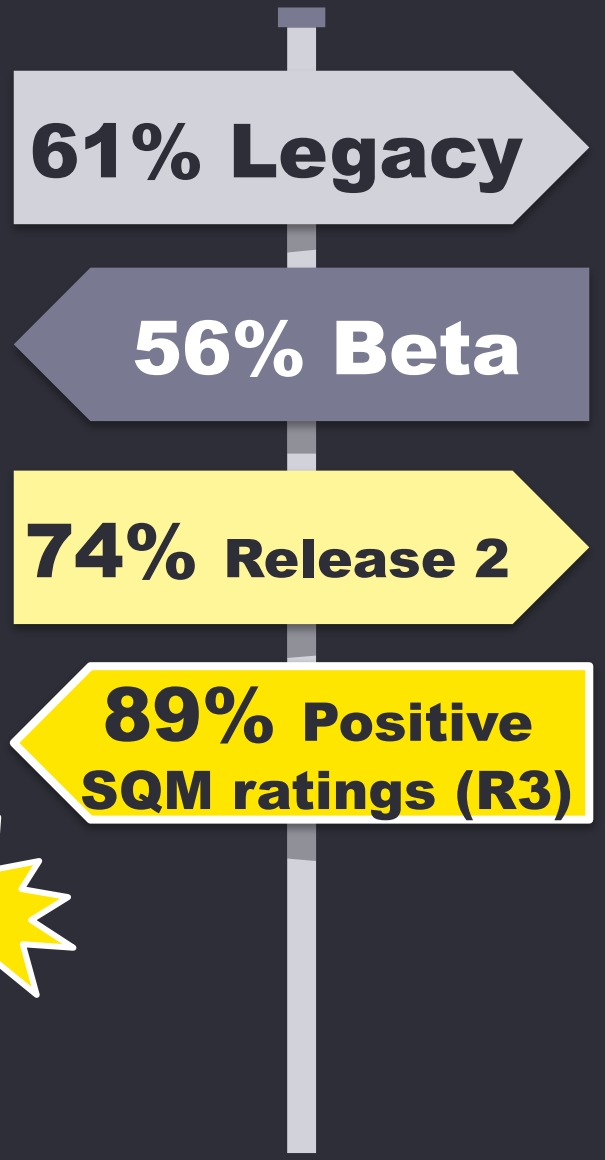
Relevence Testing response overview

56% BETA SUCCESS

Initial search relevance settings 1st attempt with content in Elastic index Improvement needed in: result set chronology, phrase matching Synonym/Acronoyms .

89% RELEASE 3 (R3)

Early results report **89%** of queries score a **Search Quality Metric (SQM)** of 2 (Good) or 3 (Perfect) results for R3 DSL Query in **Quepid**.



61% LEGACY SUCCESS

Users reported accurate & relevant results at a **61% success rate** (May 2023 study)

74% RELEASE 2 SUCCESS

Users reported R2 Discover Search was **Significantly better, Somewhat better or equal to legacy search.**
User relevance testing reported **81%** of queries scored a: **Search Quality Metric (SQM)** of 2 (Good) or 3 (Perfect) rating in Quepid