Gold Collection Report

Arquivo.pt Page Search Quality Evaluation

2026-03-03

1. Summary 2
2. Queries and Corpus Selection 3
    2.1 Query Selection Rationale 3
    2.2 Query Type Classification 3
    2.3 Corpus 3
3. Annotation Platform (anota.arquivo.pt) 4
    3.1 Architecture and Deployment 4
    3.2 Annotation Workflow 4
    3.2 Annotation Guide 4
    3.3 External Access 5
4. Annotation Results and Analysis 5
    4.1 Overall Statistics 5
    4.2 Annotator Analysis 6
        4.2.1 Annotator Outliers 6
    4.3 Query-Level Analysis 6
        4.3.1 Queries with Lowest Agreement 6
        4.3.2 Queries with Lowest Relevance 7
5. Search Quality Assessment 7
    5.1 Solr Field Weight Tuning 7
    5.2 Nutchwax vs. Solr Comparison 8
7. Conclusion and Future Work 9
    7.1 Summary of Achievements 9
    7.2 Future Work 9
Annex A: Field Weight Experiment Results 10
Annex B: Complete Per-Query Summary 11
Annex C: Per annotator summary 13
Annex D: Relevance Metrics Summary 14

1. Summary

This document describes the creation of a gold collection for Arquivo.pt, designed to assess the quality of the page search system. The goal of this project was to collect a set of queries based on historical usage patterns, annotate search results for relevance, compare the quality of the old (Nutchwax-based) and new (Solr-based) search systems, and lay the groundwork for continuous search quality monitoring.

The following work was completed:

- Selected a set of 50 queries based on previous work (github.com/arquivo/TestCollection), maintaining compatibility with the existing gold collection while covering navigational, informational, and transactional query types.
- Developed an annotation platform (anota.arquivo.pt) enabling relevance judgement collection from both FCCN staff and external collaborators.
- Collected over 8,000 relevance judgements across 50 queries, with at least 3 annotators per URL for 40 results per query (20 from each search system).
- Published the collected dataset on the TestCollection GitHub repository as an open scientific dataset.
- Evaluated and compared both search systems using standard Information Retrieval metrics, establishing a baseline for future improvements.
- Performed initial ranking function tuning for the Solr system based on earlier relevance judgement data from the previous gold collection.

2. Queries and Corpus Selection

2.1 Query Selection Rationale

A set of 50 queries was selected based on the previous test collection work. This enabled compatibility with the existing test collection[1], and direct comparison with prior evaluation results.

Previous work on the test collection had already identified a diverse set of 50 queries. However, the limited number of annotated results and the diversity of those results made it difficult to fairly evaluate the new Solr-based system, as it surfaces results that were not shown by Nutchwax and therefore had not been classified.

As with the original dataset, some queries have a desired temporal range for their results (indicated by a from: to: in the query text). These time periods are specially relevant for that specific personality or institution. For example, “Santana Lopes” query temporal range refers to the period when he was prime minister of Portugal. More information is available on the original dataset description.

2.2 Query Type Classification

The 50 queries were selected to represent the three main types of web archive search usage as specified in the requirements:

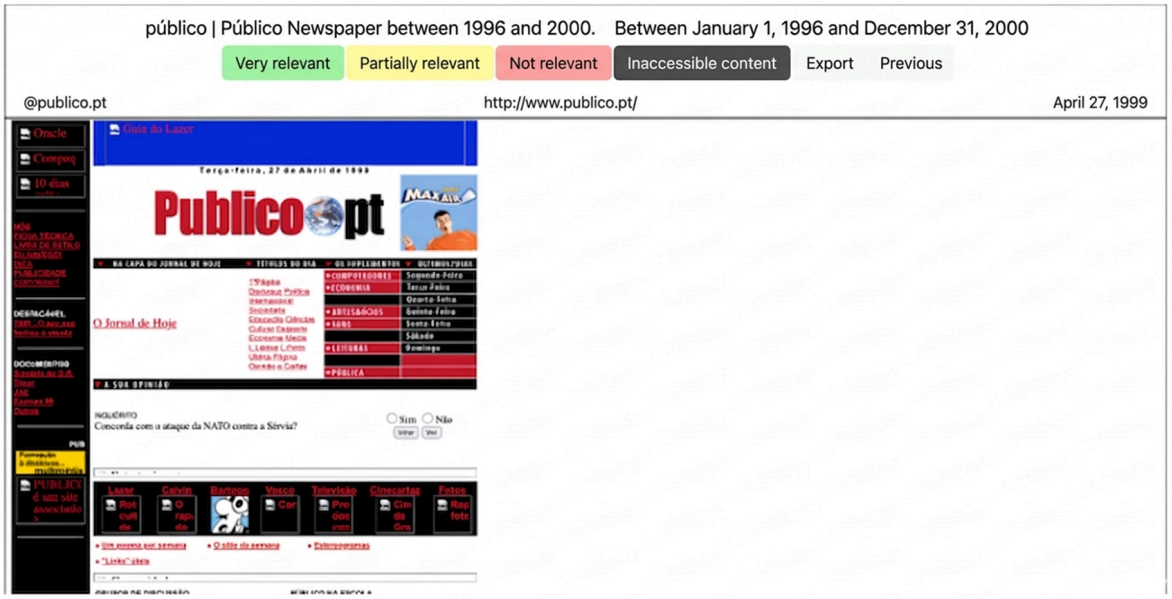
- Navigational (29 queries): Users seeking a specific known website or entity (e.g., “feup”, “google”, “olx”, “telepizza”, “volkswagen”, “moviflor”, “giatsi isel”, “icat projectos”, “net surf”, “lpcpcc”).
- Informational (17 queries): Users seeking information about a topic, person, or event (e.g., “benfica”, “cavaco silva”, “fernando pessoa”, “expo 98”).
- Transactional (4 queries): Users seeking to perform a specific action or retrieve specific content (e.g., “horários soflusa”, “restaurantes público”, “bolsa emprego sapo”, ”serviços consulado angola”).

2.3 Corpus

The evaluation corpus was the PWA9609 collection, which includes documents archived from 1996 to 2009 and has been used in multiple prior research projects. The selection of this limited corpus was driven by practical constraints: the annotation platform required two parallel search systems (Solr and Nutchwax) running side-by-side against the same data, and the additional computational capacity needed made it unfeasible to include more recent collections.

[1] https://github.com/arquivo/TestCollection

3. Annotation Platform (anota.arquivo.pt)



3.1 Architecture and Deployment

The annotation platform was deployed as a web application accessible at anota.arquivo.pt[2]. The platform runs two search backends in parallel, Solr and Nutchwax, both pointing to the same PWA9609 corpus. This side-by-side deployment enables fair comparison of search results from both systems against identical queries and data.

3.2 Annotation Workflow

The platform was designed for efficient annotation. For each query, search results are shown one at a time in their original form as a full webpage, PDF, or office document. The annotator selects a relevance level, and the page automatically advances to the next result. For each query, the system shows 20 results from Solr and 20 from Nutchwax, enabling a total of 40 annotated URLs per query. At the end of a session, results can be exported for analysis.

3.2 Annotation Guide

An annotation guide was created and published[3], to ensure consistent relevance criteria across annotators.

Annotators followed a structured workflow using the anota.arquivo.pt platform. For each assigned query, the annotator is presented with an initial screen showing the query and its relevant time range. Results are then displayed one at a time, with the original archived page (webpage, PDF, or document) rendered in an iframe below an annotation toolbar. The annotator examines the page content and selects a relevance level; the system then

[2] not available after the annotation period

[3] https://github.com/arquivo/pwa-technologies/wiki/Guia-de-anota%C3%A7%C3%A3o-de-resultados-de-pesquisa-%28pt%29

automatically advances to the next result. A "Previous" button allows corrections. After completing the evaluation, annotators export their judgements as a JSON file and submit it via a Google Form.

Relevance was assessed on a 4-level scale:

- Highly relevant - the page provides primary information about the searched topic, with specific concepts and precise answers (e.g., searching "Lisboa" returns detailed information about the Portuguese capital, its monuments, and city data).
- Partially relevant - the page contains some related information but is incomplete or tangential (e.g., an article about Portuguese cities that mentions Lisbon among other topics, or a Lisbon university page).
- Not relevant - the page has no significant connection to the search intent, even if it contains similar terms (e.g., results about a person with the surname "Lisboa" or a restaurant named "Lisboa" in another country).
- Inaccessible content - error pages, empty pages, or redirects pointing to non-archived pages.

When in doubt between two relevance levels, annotators were instructed to choose the lower one, promoting conservative judgements.

3.3 External Access

The platform was made accessible to external annotators through a web portal, requiring no software installation. Both FCCN internal staff and external collaborators were invited to participate. However, as discussed in the annotation results section, external participation remained minimal.

4. Annotation Results and Analysis

4.1 Overall Statistics

The annotation campaign collected a total of 8,656 relevance ratings across 50 queries, with the following summary statistics:

Metric	Value
Total relevance ratings	8,656
Average relevance score	0.699
Average agreement (% matching majority)	0.777
Average binary agreement	0.867
Total URLs with single rating	2,400
Total URLs with multiple ratings	2,226
Total URLs with 3+ ratings	2,137
Queries with >35 URLs with 3+ ratings	50

4.2 Annotator Analysis

The annotation effort was heavily concentrated among a small number of dedicated annotators. The top 5 contributors (Persons 2, 3, 8, 5, and 13) produced 76% of all ratings (6,593 of 8,656). The 13 core annotators with 100 or more ratings contributed 94.5% of all judgements. Seven annotators contributed fewer than 10 ratings each and are not statistically meaningful. Annex C contains the per-annotator statistics, including number of ratings given, average rating and average inter-annotator agreement (both standard and binary).

Inter-annotator agreement is measured by comparing each annotator's relevance rating against the majority vote for that document. For every (query, URL) pair rated by at least two people, the most frequent score is identified as the majority; ties are broken by taking the highest score. Agreement is then the fraction of ratings that match this majority, averaged across all multiply-rated documents. This is computed both at the query level and per annotator. Inaccessible results are excluded from agreement considerations.

Two variants are reported: (standard) **agreement** operates on the original 3-point scale (0 = not relevant, 1 = relevant, 2 = highly relevant), while **binary agreement** first collapses scores to 0/1 (any score ≥ 1 becomes 1) before applying the same majority-vote logic. Only results within rank 40 are considered, and documents with a single rater are excluded from agreement calculations

4.2.1 Annotator Outliers

Several annotators showed notable deviations from the group average:

- Low agreement: Person 21 (agreement=0.310, 29 ratings) and Person 22 (agreement=0.386, 114 ratings) had significantly lower agreement than the average of 0.777. Person 22 is particularly notable given their larger contribution volume. Person 4 (agreement=0.555, 211 ratings) also showed below-average agreement as a moderately active annotator. These annotators may have applied different relevance criteria or found certain query types ambiguous.
- Lenient raters: Person 11 (avg_rel=1.492, 264 ratings) and Person 25 (avg_rel=1.800, 10 ratings) rated significantly higher than the overall weighted average of 0.790, suggesting more generous relevance criteria.
- Strict raters: Person 9 (avg_rel=0.258, 89 ratings) and Person 32 (avg_rel=0.357, 14 ratings) were notably stricter than average.

Despite these individual variations, the use of majority voting across 3 or more annotators per URL mitigates the impact of individual biases on the final relevance judgements.

4.3 Query-Level Analysis

The following sections contain a summary of the lowest agreement and relevance queries.

4.3.1 Queries with Lowest Agreement

The following queries showed the lowest inter-annotator agreement, indicating that these topics are inherently more ambiguous or difficult to judge for relevance:

Query	Agreement	Bin. Agreement	Avg Relevance
bolsa emprego sapo	0.592	0.637	0.249
cavalo lusitano	0.600	0.783	0.738
giatsi isel	0.648	0.741	0.130
benfica	0.652	0.821	0.998
volkswagen	0.670	0.694	0.626
cavaco silva	0.672	0.881	1.176
dns	0.677	0.777	0.455
pedro abrunhosa	0.688	0.935	1.139
público	0.689	0.744	0.997
arquidiocese évora paróquias oeste	0.702	0.856	0.774

Low-agreement queries tend to be multi-word or ambiguous terms where results may partially match the query intent. For example, “bolsa emprego sapo” (agreement=0.592) combines a job search concept with a specific portal name, creating disagreement about whether tangentially related employment pages are relevant. Similarly, “cavalo lusitano” (0.600) and “giatsi isel” (0.648) are domain-specific terms where annotators may lack sufficient background to judge relevance consistently.

4.3.2 Queries with Lowest Relevance

Query	Avg Relevance	Agreement	Bin. Agreement
olx	-0.167	0.792	0.867
notícias de leiria from:1996 to:2003	0.031	0.931	0.938
timeout arte	0.051	0.848	0.873
instituto da energia	0.077	0.948	0.956
teatro dona maria segunda	0.081	0.913	0.928
net surf	0.111	0.974	0.983
instituto ricardo jorge from:2007 to:2009	0.116	0.848	0.879
giatsi isel	0.130	0.648	0.741
misericórdia de fafe from:2000 to:2009	0.158	0.892	0.917
restaurantes público	0.235	0.812	0.812

Notable patterns: “olx” has a negative average relevance (–0.167), indicating annotators actively marked results as non accessible. This is likely due to issues collecting OLX content. Several queries with very low relevance but high agreement (e.g., “notícias de leiria” at 0.031 relevance with 0.931 agreement) represent cases where annotators consistently agreed that the returned results were mostly not relevant. This is likely due to a lack of coverage during the crawling stage and not something to be tied to the search systems.

5. Search Quality Assessment

5.1 Solr Field Weight Tuning

To improve the Solr ranking function, a series of experiments were conducted using relevance judgements from the previous gold collection to evaluate different field weight configurations. The Solr ranking function combines scores from multiple document fields (title, URL tokens, internal anchor text, external anchor text, and content).

The current Solr configuration uses the following field weights:

```
title^5.0 urlTokens^2.0 inlinkAnchorsInternal^2.0 inlinkAnchorsExternal^2.0
content^1.0
```

Key findings from the field weight experiments:

- Individual field performance: When used alone, external anchor text (inlinkAnchorsExternal) is by far the strongest individual signal (MAP=0.171), followed by URL tokens (MAP=0.044) and title (MAP=0.028). Content alone is the weakest (MAP=0.009).
- Neither Solr configuration approaches Nutchwax performance (MAP=0.326), indicating that the gap is not solely due to field weighting but likely also involves differences in tokenization, scoring algorithms, and other system-level factors.

These tuning experiments used the previous gold collection's relevance judgements and to the initial set of candidates for the Solr baseline. Further tuning using the new, more comprehensive relevance judgements collected in this project is planned as future work.

5.2 Nutchwax vs. Solr Comparison

The newly collected relevance judgements were used to compare the quality of both search systems using standard information retrieval metrics. Results show that the Nutchwax system outperforms the current Solr system, particularly at top precision ranks. Annex B: Relevance Metrics Summary contains a more detailed description of the metrics and results:

Metric	Nutchwax	Solr
Mean Average Precision	0.3582	0.3147
Precision@1	0.78	0.48
Precision@5	0.62	0.54
Precision@10	0.56	0.51
Precision@100	0.10	0.09
nDCG@10	0.55	0.46

The most striking difference is at P@1 (first result relevancy), where Nutchwax achieves 0.78 compared to Solr's 0.48, a 63% relative advantage. This gap narrows at higher cutoff points but remains consistent across all metrics, indicating that Nutchwax delivers higher-quality results overall on this dataset.

7. Conclusion and Future Work

7.1 Summary of Achievements

This project successfully created a novel open scientific dataset for measuring the quality of the Arquivo.pt page search system. Over 8,000 relevance judgements were collected for 50 queries from 33 annotators, covering navigational, informational, and transactional query types. The data has been made available on GitHub for the research community.

The evaluation revealed that the current Solr-based system significantly underperforms the legacy Nutchwax system, particularly in first-result precision (P@1: 0.48 vs 0.78). Initial field weight tuning shows that external anchor text is the strongest ranking signal, but even optimized field combinations do not close the gap with Nutchwax.

7.2 Future Work

Several areas require further investigation:

- Ranking function optimization: Use the new relevance judgements to further tune Solr field weights and explore alternative scoring models (e.g., BM25 parameter tuning, learning-to-rank approaches).
- System-level analysis: Investigate the root causes of the Nutchwax–Solr performance gap beyond field weighting, including tokenization differences, scoring algorithms, and document processing pipelines.

Annex A: Field Weight Experiment Results

The following table shows the complete results of the field weight experiments conducted using the previous gold collection's relevance judgements. Each column represents a different Solr field weight configuration, with the Nutchwax baseline included for comparison.

Metric	title^5 url^2 anchInt^2 anchExt^2 cont^1	content only	anchInt only	title only	urlTokens only	anchExt only	anchExt^5 url^3 title^3 anchInt^1 cont^1	Nutchwax
MAP	0.030	0.009	0.011	0.028	0.044	0.171	0.167	0.326
MRR	0.054	0.034	0.028	0.053	0.072	0.244	0.223	0.592
P@5	0.020	0.008	0.008	0.020	0.028	0.122	0.116	0.240
P@10	0.014	0.006	0.006	0.014	0.022	0.086	0.082	0.140

Annex B: Complete Per-Query Summary

Query	Type	Ratings	Agreement	Bin. Agree.	Avg Rel.	URLs 3+ ratings
público	Nav	390	0.689	0.744	0.997	108
expo 98 from:1998 to:1998	Info	234	0.736	0.936	0.897	43
benfica	Info	467	0.652	0.821	0.998	77
cavaco silva	Info	489	0.672	0.881	1.176	74
universidade de lisboa	Nav	255	0.772	0.835	1.075	60
josé saramago	Info	196	0.742	0.938	1.270	39
câmara municipal do marco de canaveses	Nav	162	0.738	0.847	0.494	40
primeiro ministro from:2005 to:2009	Info	172	0.817	0.845	0.448	40
festival da canção	Info	160	0.790	0.956	0.906	40
finanças from:2007 to:2009	Info	165	0.772	0.833	0.739	40
federação de kickboxing	Nav	121	0.738	0.788	0.612	40
volkswagen	Nav	163	0.670	0.694	0.626	40
telepizza	Nav	144	0.785	0.821	0.708	40
hospital barreiro from:1996 to:2007	Nav	179	0.771	0.854	0.436	40
fernando pessoa	Info	142	0.788	0.921	1.310	40
google	Nav	140	0.740	0.917	0.843	40
adelaide ferreira from:1996 to:2005	Info	140	0.827	0.904	0.550	40
porto aveiro	Nav	140	0.831	0.871	0.250	40
porto de lisboa	Nav	140	0.902	0.965	0.321	40
embaixada do brasil	Nav	136	0.784	0.874	0.963	39
teatro dona maria segunda	Nav	161	0.913	0.928	0.081	40
notícias de leiria from:1996 to:2003	Nav	127	0.931	0.938	0.031	40
horários soflusa from:2003 to:2006	Trans	160	0.806	0.844	0.356	40
restaurantes público	Trans	149	0.812	0.812	0.235	40
icat projectos	Nav	160	0.806	0.894	0.606	40
centro Jacques Delors from:1996 to:1999	Nav	189	0.750	0.880	1.333	40

Santana Lopes from:2004 to:2005	Info	173	0.730	0.878	0.792	40
assembleia da república	Info	120	0.775	0.858	0.692	40
feup from:1996 to:2000	Nav	114	0.912	0.956	0.798	38
pedro abrunhosa	Info	122	0.688	0.935	1.139	40
marinha	Nav	120	0.950	0.983	0.600	40
rock in rio	Nav	119	0.771	0.921	1.160	40
cavalo lusitano	Info	126	0.600	0.783	0.738	40
moviflor	Nav	138	0.713	0.744	0.478	40
giatsi isel	Nav	108	0.648	0.741	0.130	36
instituto ricardo jorge from:2007 to:2009	Nav	155	0.848	0.879	0.116	40
dns	Info	143	0.677	0.777	0.455	40
olx	Nav	120	0.792	0.867	-0.167	40
ordem dos engenheiros from:2007 to:2010	Nav	128	0.756	0.900	1.148	40
net surf from:1996 to:2001	Nav	117	0.974	0.983	0.111	39
escola profissional chaves	Info	121	0.802	0.917	0.785	40
misericórdia de fafe from:2000 to:2009	Info	120	0.892	0.917	0.158	40
bolsa emprego sapo	Trans	173	0.592	0.637	0.249	40
timeout arte	Info	138	0.848	0.873	0.051	40
serviços consulado angola	Trans	120	0.800	0.850	0.575	40
embaixada da estónia	Nav	125	0.898	0.906	0.288	40
lpcpcc from:1996 to:2003	Nav	81	0.901	0.938	1.074	27
arquidiocese évora paróquias oeste	Nav	133	0.702	0.856	0.774	40
rádio douro sul from:1997 to:1997	Nav	152	0.776	0.928	0.842	38
instituto da energia	Nav	130	0.948	0.956	0.077	40

Annex C: Per annotator summary

Person	Total Ratings	Avg Relevance	Avg Agreement	Avg Binary Agreement
1	39	1.128	0.897	0.923
2	2,248	0.601	0.766	0.845
3	1,808	0.918	0.820	0.911
4	211	0.777	0.555	0.782
5	952	0.669	0.778	0.832
7	376	0.928	0.896	0.931
8	1,171	0.602	0.770	0.894
9	89	0.258	0.764	0.933
10	48	0.771	0.708	0.792
11	264	1.492	0.742	0.848
12	215	1.247	0.795	0.879
13	414	1.210	0.872	0.928
15	198	1.086	0.722	0.949
16	53	0.925	0.660	0.679
17	118	0.686	0.788	0.949
18	34	0.588	0.676	0.794
20	116	0.759	0.724	0.862
21	29	1.034	0.310	0.414
22	114	0.728	0.386	0.658
24	12	0.500	0.917	1.000
25	10	1.800	0.700	1.000
26	21	0.952	0.762	0.810
27	58	1.121	0.724	0.810
28	38	0.553	0.605	0.763
30	23	1.000	0.696	1.000
32	14	0.357	0.357	0.929

Annex D: Relevance Metrics Summary

Metric	Nutchwax	Solr
num_q	50	50
num_ret	973	967
num_rel	1153	1153
num_rel_ret	499	457
MAP	0.3582	0.3147
MRR (recip_rank)	0.8575	0.6578
P@1	0.78	0.48
P@5	0.62	0.54
P@10	0.56	0.51
P@20	0.50	0.46
P@50	0.20	0.18
P@100	0.10	0.09
nDCG@10	0.55	0.46

The search quality evaluation uses standard information retrieval metrics from the TREC evaluation framework. While many metrics are reported for completeness, two are particularly important for assessing the Arquivo.pt search experience:

- nDCG@10 (Normalized Discounted Cumulative Gain at 10) is the most holistic measure of search quality for this use case. It evaluates the top 10 results while accounting for both relevance degree (highly relevant results are worth more than partially relevant ones) and position (results shown higher are weighted more heavily). This mirrors how real users interact with search - they expect the best results at the top and care about graded quality, not just binary relevant/not-relevant. Nutchwax scores 0.55 vs Solr's 0.46, a 20% relative gap.
- P@1 (Precision at 1) measures whether the very first result is relevant. This is critical because many users only look at the top result, especially for navigational queries. This is where the gap is largest: Nutchwax achieves 0.78 while Solr manages only 0.48 - meaning Solr fails to return a relevant first result for more than half of queries.

Precision based metrics don’t differentiate between partially relevant and highly relevant; both are counted as relevant, while NDCG implicitly considers them by preferring highly relevant documents over partially relevant documents.
Inaccessible results are considered not-relevant.

The remaining metrics provide supporting context:

- MAP (Mean Average Precision) summarizes precision across all recall levels into a single number. It's useful for overall system comparison but less intuitive than the metrics above. Nutchwax leads 0.3582 to 0.3147.
- MRR (Mean Reciprocal Rank) measures how far down the user must scroll to find

the first relevant result. Nutchwax's 0.8575 means the first relevant result is typically at position 1, while Solr's 0.6578 suggests it's often at position 2 or lower.

- P@5, P@10, P@20 measure the fraction of relevant results (both partial and highly relevant) at increasing depth. The Nutchwax advantage is most pronounced at shallow depths (P@5: 0.62 vs 0.54) and gradually narrows as more results are considered (P@20: 0.50 vs 0.46), suggesting the core problem with Solr is ranking order rather than recall.

In short: the Solr system finds similar amounts of relevant content (num_rel_ret: 457 vs 499) but fails to rank it well, particularly at the very top of the results list. Improving top-precision and nDCG@10 should be the primary optimization targets.