NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002088

# NTCIR-18 SUSHI Pilot Task Overview

Tokinori Suzuki
University of Tsukuba
Tsukuba, Japan
tokinori@cs.tsukuba.ac.jp

Douglas W. Oard
Shashank Bhardwaj
University of Maryland
College Park, MD, USA
oard@umd.edu
sbhardw7@umd.edu

Emi Ishita
Yoichi Tomiura
Kyushu University, Fukuoka, Japan
ishita.emi.982@m.kyushu-u.ac.jp
tom@inf.kyushu-u.ac.jp

## Abstract

This paper describes the NTCIR-18 SUSHI Pilot Task. The task included two subtasks: folder search and archival reference detection. Details are presented for each subtask on the design of the test collection, the runs submitted by participating teams, and the evaluation results for those submitted runs.

## Keywords

Archival search, Inference, Text classification

## 1 Introduction

This paper describes the NTCIR-18 Searching Unseen Sources for Historical Information (SUSHI) pilot task. The broad goal of SUSHI is to support the development and evaluation of new technology to improve access to the vast quantities of undigitized materials that are held by archival institutions worldwide. The NTCIR-18 SUSHI pilot task includes two subtasks: (1) Subtask A for folder ranking, and (2) Subtask B for Archival Reference Detection. We address each in turn.

## 2 Subtask A: Folder Ranking

The test collection for Subtask A consists of a set of 31,681 PDF documents from the United States (U.S.) National Archives and Records Administration (NARA). These documents were originally created by the U.S. State Department in the 1960's and early 1970's, and they were subsequently digitized as part of a collaborative project between NARA and the Brown University Library. At NARA, each document was stored in some folder, and each folder was stored in some box. There are a total of 1,336 folders, and a total of 126 boxes.

To examine documents that have not yet been digitized, a user must travel to NARA, request that a specific box be brought to them in NARA's public reading room from that box's shelf location in NARA's closed stacks to them in the public reading room, and then examine some or all of the documents in some or all of the folders in that box. Just ordering and examining a box can easily take half a day. The goal of SUSHI Subtask A is to build systems that can, after being given a text query, tell a user which box to order, and which folder in that box to examine. Participating systems do this by producing a ranked list of folders that they expect will contain documents that the searcher would want to see.

Because our focus is on searching for undigitized content, in SUSHI Subtask A we simulate a condition in which most of the documents in the collection are not yet digitized. We do this by selecting a small sample that systems can treat as digitized, and then requiring that systems treat all other documents in the collection as undigitized. For the NTCIR-18 SUSHI Pilot Task, that small sample contains 5 documents from each of the 126 boxes, for a total of 630 digitized documents.

We start by providing details on the test collection, followed by details on how we encode the results of the the sampling process for use by participating systems (we call this encoding an Experiment Control File). We then describe of the evaluation measures used in the task, along with some remarks on the intended use of each measure. With that as background, we then present the details of the dry run task, which served both to help uncover any unspecified details in Subtask A, and which had the side effect of also producing training data that participants in Subtask A could used to perform formative evaluation during system development. We follow this by a description of our topic development process for use in the final official evaluation, after which we briefly describe the runs that were received from participating systems (these are referred to at NTCIR as "formal" system runs, to distinguish them from the earlier dry run submissions). We then describe the process by which we created relevance judgments, and we follow that with a summary of the results achieved by each participating system.

### 2.1 The Subtask A Collection

Fundamentally the Subtask A collection contains four kinds of features that a system might use:

**Folder Labels.** Each folder has a label, which generally includes a State Department Subject-Numeric Code (SNC) and a date. The date indicates the month and year (and sometimes the day) on which the folder first began collecting documents. Digital metadata is available for each folder label in the collection, and a translation table for converting SNC to text is included with the collection.

**Document Images.** Although document images exist for almost every document in the collection, systems may use only the images for the small sample of those documents specified in the Experiment Control File (see Section 2.2, below). Optical Character Recognition (OCR) has been performed on the document images, almost always with ABBYY FineReader, which was among the best available OCR systems at the time the document images were created. The document images and OCR are packaged together in PDF files (this form of PDF is referred to by Adobe as "searchable" PDF, because the OCR can be used to find the location of specific strings in the document images.

**Document Metadata.** Although digital metadata exists for almost every document in the collection, systems may use document metadata only for the documents whose use is

allowed by the Experiment Control File (i.e., the same documents for which the use of document images is allowed). There is quite a wide range of document metadata available, but the most notable document metadata is a document title. This title was created by archivists, and the title is not necessarily present in the document. Most documents also have date metadata (reflecting the document's creation date), and some documents have an SNC.

**Physical Location.** The documents are stored in folders, the folders are stored in boxes, and the boxes are stored in on shelves. These relationships are recorded as digital metadata. Systems may use the folder-box relations for all folders in all boxes, but they may only use the document-folder relation for documents whose use is allowed by the Experiment Control File. Storage relationships for boxes are encoded using sequences of box numbers, so as a made up example, box N1221 precedes box N1222, and box N1545 is quite distant from box N1221. There is no sequence information available in digital form for documents within a folder or for folders within a box.

Additional details on the collection can be found in the Subtask A guidelines on the NTCIR-18 SUSHI Pilot Task website.[1] Earlier versions of this collection have been used in published work [4, 8]. For the NTCIR-18 SUSHI pilot task, some additional PDF files and some additional metadata have been added that were not used in that earlier work.

## 2.2 Experiment Control Files

A typical ranked retrieval evaluation at NTCIR, CLEF or TREC would provide participating teams with a set of topic descriptions, each of which includes a Title field (a short web-like query), a Description field (a sentence or so that expresses the information need succinctly), and a Narrative field (a short paragraph that provides further details as the basis for relevance assessment) [6]. For Subtask A, we also need to specify which documents (i.e., document images and document metadata) are in the small sample that systems can make use of (we call these "training documents").

Because training a system can be much more expensive than searching with that system, we expected that participating teams would prefer to train a system once and then use that trained system to generate ranked lists for several topics. We therefore packaged one set of training data with several topics; we call such a package an "experiment set." Because the results might depend to some degree on which documents were in the training set, a full Subtask A run includes several experiment sets, each with a different set of training documents. For the NTCIR-18 SUSHI Pilot Task, topics were not duplicated across experiment sets.

A full experiment is thus a set of experiment sets. An "Experiment Control File" (ECF) is used to specify that set of experiment sets. The ECF is distributed as a JSON object. Systems are expected to read an ECF and then write a set of ranked lists, which we call a run, in a standard format (we used the same standard format for runs as in TREC). Additional details on the format of an ECF can be found in the Subtask A guidelines on the SUSHI website.

[1] https://sites.google.com/view/ntcir-sushi-task/

## 2.3 Evaluation Measures

We report four types of evaluation measures for each run. All measures are reported principally as averages across the full topic set (i.e., across every topic in the ECF, regardless of which experiment set each topic is in) on the ranked list of folders for each topic.[2]

**nDCG@5.** normalized Discounted Cumulative Gain at cutoff 5 (nDCG@5) is the primary evaluation measure for Subtask A. Its design reflects an assumption that a searcher has time to examine no more that five folders, and that the searcher prefers to see the most highly relevant folders first when working down from the top of the ranked list. Only nDCG@5 makes use of graded relevance; all other measures use binary relevance.

**MAP.** Mean Average Precision (MAP) has no cutoff. Its design reflects an assumption that the searcher wishes to find some (unknown) number of relevant folders, and that they prefer to achieve this while looking at the smallest number of folders when working down from the top of the ranked list.

**MRR.** Mean Reciprocal Rank (MRR) also has no cutoff. Its design reflects an assumption that the searcher wishes to find any one relevant folder, and that they prefer that the first relevant folder be as close to the top of the ranked list as is possible.

**S@1.** Success at 1 (S@1) assumes that the searcher will examine only the top-ranked folder. If it is relevant they will be fully satisfied. If not, they will be completely dissatisfied.

In addition to computing these measures on folders, we also compute the same measures on ranked lists of boxes. To do so, we map every Folder ID in a run to the Box ID of the box containing that folder, and then we remove duplicates from the top down in each ranked list in that run. We then evaluate using box qrels rather than box qrels.

All of our computations for evaluation are performed using the python version Terrier's implementation of the trec_eval evaluation script, with Terrier's default settings for the nDCG@5 discount rate.

## 2.4 The Dry Run

For the dry run we didn't yet have any topics developed, so instead we used document titles (from the document metadata) as substitutes for the short "Title" queries. This idea has been tried twice before with smaller collections from the same source [4, 8], in a known-item retrieval setting (i.e., the item from which the title had been taken was known to be relevant, and all other documents are assumed not to be relevant). We improved over that prior practice in three ways. First, we considered only relatively short titles of 2, 3, 4, or 5 words; in prior work the length had not been limited, and there were some quite long titles. Second, we asked two assessors (the last two authors of this paper) to examine the titles and select 200 (50 of each length) that looked to them like queries that someone might realistically issue. They had to look at about 400 titles to find 200 good queries. Third, we treated not just the original document from which the title had been obtained as relevant, but any other document in the collection with the same title (in its document

[2] For reasons of convention, two of the measures include "mean" in their name, but all are actually normally reported as means across the topic set. The only exception is nDCG@5, which we also report on a per-topic basis.

metadata). Earlier work had used only MRR and S@1 as evaluation measures, but once multiple relevant documents were possible all four of our evaluation measures became potentially insightful.

We then created an ECF with two experiment sets, each with 100 topics. We replicated the document title in all three topic fields, just so that code that used any topic field(s) could be tested. We selected five documents from every box as the training documents in each experiment set. This approach had been used previously with random selection for experiments in which searching for boxes was the goal [4, 8], but because our goal was to evaluate systems that would rank folders rather than boxes we chose to select one training document from each of the five largest folders in a box.[3] We were careful not to select as training documents any of the 100 seed documents from which the topics in that same experiment set had been constructed, but otherwise the selections within a folder were random.

We later learned that consistently selecting folders from the largest five folders in each box had two undesirable side effects. First, the folders from which the training documents had been selected were (by construction) identical in both experiment set. This resulted in less diversity than we would have expected to see in practice, so when we created ECF for the final evaluation (which had human-created topics, see Section 2.5 below) we handled this differently. Second, our choice of the largest folders risked biasing systems towards larger folders. That might seem like a useful heuristic (since folders with more documents would have a higher chance of having a relevant document, all else equal), but we had explicitly prohibited systems from using full-collection document metadata (because in practice document metadata is typically only created when documents are digitized, and we were simulating most of the collection being undigitized). So this essentially leaked some potentially useful, but prohibited, data.

We distributed the ECF for the Dry Run test collection on the SUSHI website, along with python code to create a Terrier baseline run for that collection. Our intent in sharing that code was to simplify the task of managing the rather complex ECF and metadata files in ways that complied with all requirements of Subtask A. A participating team could then simply rip out the Terrier code and add their own code for indexing and search and they would have a dry run submission. The code we distributed also included code and qrels for scoring the resulting runs by all four measures for either folders or boxes.

Dry Run submissions were due on August 31, 2024. No submissions were received, but all three participating teams in the final evaluation did make use of the dry run test collection to tune their systems for that final evaluation.

## 2.5 Topic Development

Topic development was performed by seven people, including the first four authors of this paper. Topic developers used an early version of the relevance assessment system (see Section 2.7, below) to explore the content of the collection and to iteratively refine a topic to have not too few and not too many relevant documents (i.e., so-called "Goldilocks" topics). We sought to avoid topics with very few relevant documents because our nDCG@5 and MAP evaluation measures exhibit higher quantization noise with very few relevant documents, thus making it harder to measure statistically significant differences. We sought to avoid topics with very many relevant documents because our use of search-guided assessment, without pooled assessment, was based on an assumption that almost all of the relevant documents could be found by assessors, and the cost in assessor time to find a very large number of relevant documents would simply be too high for our time and budget constraints. Other types of topics are, of course, also important in practice, but they would not be compatible with our approach to evaluation.

```
Title: Marijuana consumption
Description: I want to find documents about marijuana
consumption in Brazil.
Narrative: Marijuana, known also as "cannabis," "pot,"
and "weed," is a psychoactive drug made from the cannabis
plant that is used for medical and recreational purposes.
Any relevant documents would describe the creation,
distribution, and consumption of marijuana in Brazil,
social attitudes regarding its use, and legal restriction
and/or prohibition of its trade and consumption.
```

**Figure 1: An example Subtask A topic (Topic 9).**

Assessors created Title, Description and Narrative fields for each topic, some which were then edited for clarity by the second author of this paper. Figure 1 shows an example. Overall, our assessment is that the Description fields are less different from the Title fields than is typical for earlier NTCIR, CLEF and TREC evaluations, but that the Narrative fields are considerably richer than is typical at NTCIR, CLEF and TREC.

We somewhat overgenerated topics and then removed a few (generally for having too many relevant documents), ultimately settling on a topic set with 45 topics. We then created an ECF with three experiment sets, each of which included a disjoint set of 15 of our 45 topics. We then selected training documents for each experiment set, again with five documents per box, but this time by randomly selecting a folder from a box and then randomly selecting a document from a folder. We did this random selection without replacement for both folders[4] and documents, so as to maximize the coverage of the training set. We did not exclude relevant documents from the training set because relevance judgments were performed after the final ECF was distributed to participating teams, so it is possible (although rather unlikely) that a training document might be relevant to some topic in its experiment set.

## 2.6 Submissions and Baseline Runs

We received a total of 37 Subtask A runs from 3 participating teams [2, 5, 9]. All of the received runs were automatic, meaning that there was no human intervention (such as interactive relevance feedback, or improvements to handle certain types of topics) after the creator of the system that created the runs first examined the topics. Table 1 shows the submission statistics by participating team. We note that both the Kyushu University team [9] and the

[3] If there were fewer than five folders in a box, we used a round-robin strategy to rotate between the folders from which we selected training documents.

[4] Unless there were fewer than 5 folders in a box, in which case we again did round-robin selection, but this time in a randomized initial order

**Table 1: Subtask A runs.**

| Organization | Run Name Prefix | Runs |
|---|---|---|
| Kyushu University (Japan) | QshuNLP | 1 |
| University of Maryland (USA) | UMCP | 25 |
| University of Tsukuba (Japan) | KASYS | 8 |
| Organizers' Baselines | TerrierBaseline | 3 |

University of Maryland team [5] included task organizers. Their systems were designed and run using only information that was available to all participants, but as organizers they did make consequential choices (such as how the samples in the final Experiment Control File were drawn), and they were thus aware of the rationale and the details of such choices.

The organizers also contributed 3 baseline runs. These three baseline runs were created by using the BM-25 implementation in Terrier to rank only the training documents in an experiment set for each topic in that experiment set, based on an index containing the document title, OCR text and folder label (the folder label would be the same for any training documents that were in the same folder). The code used to create this run was the same as the baseline system that had been made available to participants prior to the dry run. That code included folder label mappings to convert Subject-Numeric Codes from some of the folder labels (NARA folder labels, and short Brown folder labels that contained fewer than 20 characters) to text. The only change made to that code was to create three result sets, one for Title queries (T), one for Title+Description queries (TD), and one for Title+Description+Narrative queries (TDN). Rather than returning document ID's, Terrier was configured to return the folder ID for the folder containing a document. Duplicate folder ID's were then removed, working down from the top of the list, and the resulting ranked list of folders submitted as a baseline result.

## 2.7 Relevance Assessment

Three assessors created relevance judgments for the 45 topics. All three were graduate students (one Ph.D. student, two Masters students) enrolled in a library science degree program. Two are native speakers of English, the third is fluent in English. They used a search-guided relevance assessment process [1] in which they first judged the documents that had been identified during topic development as worthy of examination, and then they performed several searches using queries of their own design that were based on their understanding of the topics from the full topic description. The goal of this process was to find as many documents as possible. No use was made of system runs in prioritizing documents for relevance assessment (i.e., pooled assessment was not used) because systems submitted ranked lists of folders, whereas assessors created relevance judgments for individual documents.

When possible, relevance assessment was performed by the person who had done topic development; this happened for 22 of the 45 topics.[5] After training over Zoom, assessors worked remotely, using a web browser to connect to the relevance assessment system shown in Figure 2, which was created by the first and third authors

[5]Topics 4, 6, 7, 8, 9, 10, 12, 13, 18, 20, 22, 23, 24, 25, 28, 33, 34, 35, 36, 41, 42, 43.

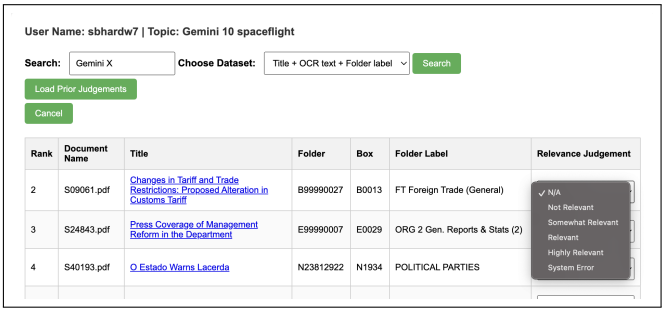


**Figure 2: Relevance assessment system.**

of this paper. That system allowed the assessors to issue queries, select which of three metadata or content field(s) to search, see a ranked list of document titles (along with the label of the folder containing that document), and select and view individual PDF documents. Within-document search for any term was also available. The documents that had been identified as worthy of examination during topic development could also be displayed as an unranked list using the "Load Prior Judgments" button.

Assessors recorded graded relevance judgments as Highly Relevant, Relevant, Somewhat Relevant or Not Relevant. They did this using per-document pull-down lists, as shown for the first document on the right side of Figure 2. Assessors were asked to find all of the documents that were relevant to any degree to the topic being judged, and to also record judgments of Not Relevant for a reasonable number of documents (which we interpreted as being about 50 documents, although the number of not relevant judgments varies considerably by topic). In our evaluation we treat all unjudged documents as Not Relevant, so we do not use these explicit Not Relevant judgments in our evaluation. They were collected because we expect that they might be useful in the future for contrastive learning or for calculating evaluation measures that treat documents that were judged as Not Relevant differently from unjudged documents.

We performed three mappings to get to the "qrels" files that we used to compute system scores. First, we assigned an integer score (a qrel value) to each relevance judgment as follows: highly relevant=3, relevant=1, somewhat relevant=0, not relevant=0. We then assigned each folder that contained one or more judged documents the maximum numeric qrel value for any document in that folder. We then assigned each box the maximum numeric qrel value for any folder in that box. Three of our evaluation measures require binary relevance judgments; for those measures our evaluation code implicitly maps qrel values of 3 to 1, and leaves other qrel values unchanged.

The net result of this process was three qrels files, one each for documents, folders, and boxes.[6] All three sets are incomplete (because assessors may have assessed no document in some box or some folder). On average per topic there were 75.4 documents

[6]The story here is slightly simplified, in that the documents qrels file actually contains the unmapped original judgments on a scale of 1 (not relevant) to 4 (highly relevant). This unmapped version was created to support future use of the collection; we do not compute evaluation measures using the document qrels file.

judged, 37.0 folders with qrels, and 25.9 boxes with qrels.[7] Our evaluation code treats things that are not in a qrels file as if they had a qrel value of 0. The contributions of the three assessors to the final judgment file were well balanced, with Assessor A1 making the judgments for 18 topics in the final official qrels file,[8] A2 making the judgments for 15 topics,[9] and A3 making the judgments for 12 topics.[10]

Nine topics were subsequently assessed by a second assessor.[11] We stratified our sample for this dual assessment process to include three topics for each assessor pairing. Using an example from Table 2, we can see that Assessor 1's judgments for topic 23 would have resulted in 7 folders with qrel values of 1 or 3, and using Assessor 3's judgments also would have resulted in 7 folders with qrel values of 1 or 3. However, only 5 of these judgments of 1 or 3 were for the same folder, and only 4 of those five folders had exactly the same qrel value (i.e., either both 1 or both 3). Looking at the full set of 9 dual assessed topics, we see folder-level agreement patterns that seem to us to be well within the usual range of document-level) agreement that is typically seen in NTCIR, CLEF and TREC evaluations [10]. Moreover, we see no systematic differences that might cause us any concern regarding assessor training or diligence. The observed differences could, for example, easily result from differences in assessor decisions regarding issues that are not fully specified in the topic description. Moreover, looking in detail at cases in which one assessor's judgments produced a folder qrel of 3 and the other's produced a folder qrel of 1, we found no pattern in which assessor was more strict or more generous, indicating that the relevance assessments of individual assessors did not seem to be markedly different. We thus believe the resulting folder qrels to be useful for our intended purpose of evaluating folder ranking.[12]

Our initial goal for doing dual assessment had been to characterize the degree of agreement on relevance between annotators, but we found that we were also able to use the results of that dual annotation to detect cases in which different choices made by different annotators might have caused one to find a relevant document that the other had missed. In this way, we could get some indication of how comprehensive the search process had been. As the "Both Judged" column in Table 2 shows, 8 of the 9 folders that one assessor's judgments gave a 1 or a 3 to were judged by the other assessor. Looking at the fraction of the Union that was Both Judged over all 9 topics leads us to conclude that it was rare for one Assessor to find a relevant document that the other assessor had missed.[13] Hence, we believe that the relevance judgments are sufficiently complete for our intended purpose of evaluating folder ranking.

Another way to characterize the completeness of the relevance judgments would be to calculate the fraction of the top-5 folders for each run for which a relevance judgment for at least one document is available. If that "pool completeness" number were high, then we would know that the assessors had seen and judged at least one document in most of the highly ranked folders that had contributed to the nDCG@5 computation. If pool completeness were low, we might expect that guiding the assessor to judge at least some documents from highly ranked but unjudged folders could have been useful. We have not yet performed that analysis, but we expect to do so before the NTCIR conference and we will report the results of that analysis there.

**Table 2: Subtask A folder qrels positive agreement. All counts except Exact Agree are for binarized qrels. Number of positive judgments for an assessor in italics indicates those are used in the test collection; underline indicates the assessor was the topic developer.**

| Topic | A1 | A2 | A3 | Union | Both Judged | Binary Agree | Exact Agree |
|---|---|---|---|---|---|---|---|
| 1 | *13* | | 8 | 16 | 11 | 5 | 2 |
| 5 | 36 | | *26* | 39 | 35 | 23 | 17 |
| 14 | *2* | 2 | | 2 | 2 | 2 | 1 |
| 15 | | *7* | 30 | 32 | 17 | 5 | 2 |
| 23 | 7 | | *7* | 9 | 8 | 5 | 4 |
| 29 | *12* | 15 | | 21 | 18 | 6 | 4 |
| 36 | | *9* | 25 | 26 | 19 | 8 | 3 |
| 38 | *8* | 4 | | 8 | 7 | 4 | 3 |
| 39 | | *2* | 2 | 2 | 2 | 2 | 1 |

## 2.8 Results

Table 3 summarizes the folder ranking results for the four evaluation measures, and Table 4 does the same for box ranking. Each table is sorted in decreasing order of nDCG@5 for each query type (TDN, TD, T, or D). Note, however, that with 45 topics the confidence intervals on nDCG@5 are rather wide, so at least the top half of the runs for each query type would not be statistically distinguishable using a population test. Paired tests (pairing on topic) could be used to better distinguish systems, but we have not yet performed that analysis.

Comparing across query types using Folder nDCG@5, we see that among the best runs TDN>TD>T, although the measured differences are smaller than the confidence intervals so this pattern is at best suggestive. As Table 5 shows, participating systems found some topics to be easier than others. The topics in that table are sorted from easiest to hardest, as measured by the average across all participating systems of the per-topic nDCG@5. As can be seen, most systems were able to place some relevant folder somewhere in the top five ranks for 21 of the 45 topics. For another 13 topics, at least one system managed to get at least relevant folder somewhere in the top five.[14] No system found any relevant folder for any of the remaining 11 topics.

[7] These counts include folders or boxes with any qrel value, including 0.
[8] Assessor A1 judged topics 1, 2, 3, 11, 14, 16, 19, 26, 27, 29, 30, 31, 32, 37, 38, 40, 44, 45.
[9] Assessor A2 judged topics 4, 6, 7, 8, 15, 17, 18, 20, 21, 22, 33, 34, 35, 36, 39.
[10] Assessor A3 judged topics 5, 9, 10, 12, 13, 23, 24, 25, 28, 41, 42, 43.
[11] The official qrels used in the evaluation were built from the first set of relevance judgments that were completed for each of the dual-assessed topics.
[12] We note, however, that our analysis of assessor agreement was conducted only on folder qrels, and that additional difference in detail may be present in the document-level relevance judgments from which those folder qrels were constructed.
[13] More precisely, we conclude that it is rare for this to happen in a way that changes the folder qrels after the two mappings that are needed to create folder qrels.
[14] However, for two of those topics (Topics 12 and 40) we can't reject the hypothesis that the single system that found a relevant folder did so by blind luck. Making an (unjustified) independence assumption, with 37 systems, each with 5 top ranks, random selection among 1,336 folders would yield a 14% chance of getting any one folder into some system's top five.

**Table 3: Subtask A folder ranking results, with 95% confidence intervals.**

| Run | Query | nDCG@5 | ± | MAP | ± | MRR | ± | S@1 | ± |
|---|---|---|---|---|---|---|---|---|---|
| UMCP-TDN-TOFS-L2-CF | TDN | 0.229 | 0.08 | 0.165 | 0.06 | 0.490 | 0.13 | 0.400 | 0.14 |
| UMCP-TDN-TOFS-L2 | TDN | 0.214 | 0.08 | 0.143 | 0.06 | 0.462 | 0.14 | 0.400 | 0.14 |
| UMCP-TDN-TOFS-U2-CF | TDN | 0.210 | 0.07 | 0.159 | 0.06 | 0.446 | 0.12 | 0.333 | 0.14 |
| KASYS-1 | TDN | 0.203 | 0.08 | 0.129 | 0.06 | 0.417 | 0.13 | 0.356 | 0.14 |
| TerrierBaseline-TDN | TDN | 0.203 | 0.08 | 0.132 | 0.06 | 0.417 | 0.13 | 0.333 | 0.14 |
| UMCP-TDN-O-B | TDN | 0.200 | 0.08 | 0.131 | 0.07 | 0.425 | 0.14 | 0.378 | 0.14 |
| UMCP-TDN-S-B | TDN | 0.188 | 0.08 | 0.130 | 0.07 | 0.419 | 0.14 | 0.400 | 0.14 |
| UMCP-TDN-T-B | TDN | 0.183 | 0.08 | 0.097 | 0.06 | 0.339 | 0.13 | 0.289 | 0.13 |
| UMCP-ColBERT-TDN | TDN | 0.173 | 0.08 | 0.103 | 0.06 | 0.313 | 0.12 | 0.244 | 0.13 |
| KASYS-5 | TDN | 0.131 | 0.07 | 0.094 | 0.06 | 0.272 | 0.11 | 0.178 | 0.11 |
| KASYS-4 | TDN | 0.120 | 0.07 | 0.102 | 0.06 | 0.264 | 0.11 | 0.178 | 0.11 |
| KASYS-8 | TDN | 0.120 | 0.07 | 0.102 | 0.06 | 0.264 | 0.11 | 0.178 | 0.11 |
| KASYS-3 | TDN | 0.081 | 0.06 | 0.082 | 0.05 | 0.165 | 0.08 | 0.067 | 0.07 |
| KASYS-7 | TDN | 0.081 | 0.06 | 0.082 | 0.05 | 0.165 | 0.08 | 0.067 | 0.07 |
| KASYS-2 | TDN | 0.059 | 0.05 | 0.073 | 0.05 | 0.140 | 0.07 | 0.067 | 0.07 |
| KASYS-6 | TDN | 0.059 | 0.05 | 0.073 | 0.05 | 0.140 | 0.07 | 0.067 | 0.07 |
| UMCP-TDN-F-B | TDN | 0.059 | 0.04 | 0.036 | 0.02 | 0.173 | 0.09 | 0.089 | 0.08 |
| UMCP-TD-TOFS-L2-CF | TD | 0.228 | 0.08 | 0.151 | 0.06 | 0.434 | 0.13 | 0.333 | 0.14 |
| UMCP-TD-TOFS-L2 | TD | 0.213 | 0.08 | 0.129 | 0.06 | 0.412 | 0.13 | 0.333 | 0.14 |
| UMCP-TD-TOFS-U2-CF | TD | 0.212 | 0.08 | 0.153 | 0.06 | 0.428 | 0.13 | 0.333 | 0.14 |
| TerrierBaseline-TD | TD | 0.197 | 0.08 | 0.122 | 0.06 | 0.384 | 0.13 | 0.289 | 0.13 |
| UMCP-ColBERT-TD | TD | 0.183 | 0.08 | 0.116 | 0.06 | 0.347 | 0.11 | 0.222 | 0.12 |
| UMCP-TD-O-B | TD | 0.183 | 0.08 | 0.117 | 0.06 | 0.378 | 0.13 | 0.311 | 0.14 |
| UMCP-TD-S-B | TD | 0.180 | 0.08 | 0.119 | 0.07 | 0.377 | 0.14 | 0.333 | 0.14 |
| UMCP-TD-T-B | TD | 0.160 | 0.08 | 0.088 | 0.06 | 0.297 | 0.12 | 0.222 | 0.12 |
| UMCP-TD-F-B | TD | 0.048 | 0.03 | 0.021 | 0.02 | 0.135 | 0.09 | 0.093 | 0.09 |
| UMCP-T-TOFS-L2-CF | T | 0.226 | 0.08 | 0.150 | 0.06 | 0.455 | 0.13 | 0.378 | 0.14 |
| UMCP-T-TOFS-L2 | T | 0.211 | 0.08 | 0.128 | 0.06 | 0.416 | 0.14 | 0.356 | 0.14 |
| UMCP-T-TOFS-U2-CF | T | 0.210 | 0.08 | 0.152 | 0.06 | 0.432 | 0.13 | 0.333 | 0.14 |
| TerrierBaseline-T | T | 0.204 | 0.08 | 0.125 | 0.06 | 0.41 | 0.13 | 0.333 | 0.14 |
| UMCP-T-O-B | T | 0.191 | 0.08 | 0.119 | 0.06 | 0.398 | 0.14 | 0.356 | 0.14 |
| UMCP-ColBERT-T | T | 0.185 | 0.08 | 0.113 | 0.06 | 0.377 | 0.13 | 0.311 | 0.14 |
| UMCP-T-S-B | T | 0.180 | 0.08 | 0.116 | 0.07 | 0.384 | 0.14 | 0.364 | 0.14 |
| UMCP-T-T-B | T | 0.178 | 0.09 | 0.099 | 0.07 | 0.337 | 0.14 | 0.282 | 0.14 |
| QshuNLP-GPT-1 | T | 0.126 | 0.06 | 0.105 | 0.06 | 0.267 | 0.11 | 0.200 | 0.12 |
| UMCP-T-F-B | T | 0.072 | 0.05 | 0.031 | 0.02 | 0.213 | 0.14 | 0.179 | 0.14 |
| UMCP-ColBERT-D | D | 0.163 | 0.07 | 0.096 | 0.05 | 0.327 | 0.12 | 0.267 | 0.13 |

Counts for Highly Relevant (qrel=3) and Relevant (qrel=1) folders are shown at the top of Table 5, and corresponding counts for boxes are shown at the bottom of that table. Notably, every topic has at least one Highly Relevant or Relevant folder (and thus at least one Highly Relevant or Relevant box) that could have been found by participating systems. We see no pattern in those counts that would explain topic difficulty. For example, we Topics 16 and 28 each have just one Highly Relevant folder, and most systems were perfect (i.e., ranking that folder first) on both of those topics. However, topic 26 also has just one Highly Relevant folder, and Topic 43 has just one Relevant Folder, and but no system ranked the single Highly Relevant or Relevant folder in the top 5 for either of those topics. The existence of relatively large numbers relevant folders is similarly uninformative, as can be seen by comparing topic 44, where systems did fairly well with topic 13 where systems did quite badly. From this we conclude that it is not just the number of relevant folders that determines whether a topic is easy or hard, at least when the determination of what is easy and what is hard is made using our 37 participating systems.

We also checked for an assessor effect, but found no clear pattern. For example, the top 22 of the 45 topics (when ordered by per-topic nDCG@5) included about half from each assessor (A1: 8/18, A2: 10/15, A3: 4/12); all of those numbers are within two of half, and a chi-squared test can not reject the possibility that the observed variation in those proportions results from chance.

So what is it that makes a topic easy or hard? One obvious possibility is that the training set of five digitized documents per box might be more useful for some topics than for others. A second possibility might be that the easier or harder topics could share some characteristic that we have not yet looked at (such as the

**Table 4: Subtask A box ranking results, with 95% confidence intervals.**

| Run | Query | nDCG@5 | ± | MAP | ± | MRR | ± | S@1 | ± |
|---|---|---|---|---|---|---|---|---|---|
| UMCP-TDN-TOFS-L2-CF | TDN | 0.306 | 0.08 | 0.300 | 0.07 | 0.603 | 0.12 | 0.489 | 0.15 |
| UMCP-TDN-TOFS-U2-CF | TDN | 0.298 | 0.08 | 0.291 | 0.07 | 0.536 | 0.12 | 0.378 | 0.14 |
| UMCP-TDN-TOFS-L2 | TDN | 0.297 | 0.09 | 0.270 | 0.07 | 0.572 | 0.13 | 0.467 | 0.15 |
| TerrierBaseline-TDN | TDN | 0.266 | 0.08 | 0.250 | 0.07 | 0.526 | 0.13 | 0.422 | 0.15 |
| KASYS-1 | TDN | 0.261 | 0.08 | 0.228 | 0.06 | 0.500 | 0.12 | 0.378 | 0.14 |
| UMCP-TDN-O-B | TDN | 0.253 | 0.09 | 0.240 | 0.07 | 0.520 | 0.13 | 0.444 | 0.15 |
| UMCP-ColBERT-TDN | TDN | 0.248 | 0.09 | 0.217 | 0.08 | 0.430 | 0.11 | 0.289 | 0.13 |
| UMCP-TDN-S-B | TDN | 0.243 | 0.09 | 0.239 | 0.07 | 0.503 | 0.13 | 0.444 | 0.15 |
| UMCP-TDN-T-B | TDN | 0.237 | 0.09 | 0.188 | 0.07 | 0.474 | 0.13 | 0.378 | 0.14 |
| KASYS-5 | TDN | 0.226 | 0.08 | 0.198 | 0.06 | 0.372 | 0.10 | 0.200 | 0.12 |
| KASYS-4 | TDN | 0.199 | 0.08 | 0.224 | 0.07 | 0.369 | 0.11 | 0.222 | 0.12 |
| KASYS-8 | TDN | 0.199 | 0.08 | 0.224 | 0.07 | 0.369 | 0.11 | 0.222 | 0.12 |
| UMCP-TDN-F-B | TDN | 0.148 | 0.06 | 0.142 | 0.05 | 0.313 | 0.10 | 0.156 | 0.11 |
| KASYS-3 | TDN | 0.113 | 0.06 | 0.171 | 0.05 | 0.220 | 0.07 | 0.067 | 0.07 |
| KASYS-7 | TDN | 0.113 | 0.06 | 0.171 | 0.05 | 0.220 | 0.07 | 0.067 | 0.07 |
| KASYS-2 | TDN | 0.072 | 0.05 | 0.159 | 0.05 | 0.200 | 0.07 | 0.067 | 0.07 |
| KASYS-6 | TDN | 0.072 | 0.05 | 0.159 | 0.05 | 0.200 | 0.07 | 0.067 | 0.07 |
| UMCP-TD-TOFS-L2-CF | TD | 0.308 | 0.09 | 0.281 | 0.07 | 0.546 | 0.12 | 0.422 | 0.15 |
| UMCP-TD-TOFS-U2-CF | TD | 0.283 | 0.08 | 0.270 | 0.07 | 0.514 | 0.12 | 0.378 | 0.14 |
| UMCP-ColBERT-TD | TD | 0.261 | 0.08 | 0.219 | 0.07 | 0.503 | 0.12 | 0.378 | 0.14 |
| UMCP-TD-TOFS-L2 | TD | 0.255 | 0.09 | 0.243 | 0.07 | 0.526 | 0.13 | 0.422 | 0.15 |
| UMCP-TD-S-B | TD | 0.239 | 0.09 | 0.214 | 0.07 | 0.469 | 0.13 | 0.400 | 0.14 |
| TerrierBaseline-TD | TD | 0.233 | 0.08 | 0.218 | 0.07 | 0.473 | 0.12 | 0.356 | 0.14 |
| UMCP-TD-O-B | TD | 0.211 | 0.08 | 0.203 | 0.07 | 0.465 | 0.13 | 0.378 | 0.14 |
| UMCP-TD-T-B | TD | 0.207 | 0.08 | 0.156 | 0.07 | 0.415 | 0.12 | 0.311 | 0.14 |
| UMCP-TD-F-B | TD | 0.109 | 0.06 | 0.078 | 0.05 | 0.213 | 0.10 | 0.116 | 0.10 |
| UMCP-T-TOFS-L2-CF | T | 0.287 | 0.08 | 0.265 | 0.07 | 0.548 | 0.13 | 0.467 | 0.15 |
| UMCP-T-TOFS-U2-CF | T | 0.279 | 0.08 | 0.261 | 0.07 | 0.512 | 0.12 | 0.378 | 0.14 |
| QshuNLP-GPT-1 | T | 0.262 | 0.09 | 0.233 | 0.08 | 0.440 | 0.12 | 0.333 | 0.14 |
| UMCP-T-TOFS-L2 | T | 0.254 | 0.09 | 0.208 | 0.07 | 0.504 | 0.13 | 0.422 | 0.15 |
| UMCP-ColBERT-T | T | 0.249 | 0.08 | 0.211 | 0.07 | 0.485 | 0.12 | 0.356 | 0.14 |
| TerrierBaseline-T | T | 0.240 | 0.08 | 0.192 | 0.07 | 0.490 | 0.13 | 0.400 | 0.14 |
| UMCP-T-S-B | T | 0.235 | 0.09 | 0.172 | 0.07 | 0.463 | 0.14 | 0.409 | 0.15 |
| UMCP-T-T-B | T | 0.232 | 0.09 | 0.157 | 0.08 | 0.460 | 0.15 | 0.410 | 0.16 |
| UMCP-T-O-B | T | 0.223 | 0.08 | 0.175 | 0.07 | 0.481 | 0.14 | 0.422 | 0.15 |
| UMCP-T-F-B | T | 0.140 | 0.08 | 0.086 | 0.05 | 0.278 | 0.14 | 0.179 | 0.14 |
| UMCP-ColBERT-D | D | 0.235 | 0.08 | 0.185 | 0.06 | 0.486 | 0.13 | 0.378 | 0.14 |

use of highly distinguishing proper names). Further analysis might suggest other possibilities as well.

**Table 5: Folder ranking, per-topic results, nDCG@5. Bottom two rows show counts for Box qrels.**

| | 16 | 28 | 36 | 2 | 19 | 44 | 23 | 15 | 5 | 20 | 35 | 37 | 11 | 1 | 18 | 22 | 21 | 7 | 4 | 24 | 32 | 39 | 33 | 27 | 38 | 42 | 45 | 41 | 29 | 40 | 34 | 9 | 12 | 6 | 3 | 8 | 10 | 13 | 14 | 17 | 25 | 26 | 30 | 31 | 43 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Highly Relevant | 1 | 1 | 4 | 5 | 6 | 2 | 6 | 4 | 15 | 7 | 3 | 9 | 10 | 6 | 3 | 1 | 1 | 1 | 7 | 2 | 2 | 1 | 3 | 4 | | 3 | 5 | 3 | 8 | 2 | 3 | 2 | 2 | 3 | 3 | 1 | 2 | 13 | 1 | 2 | 1 | 1 | 1 | 3 | | 3.62 |
| Number of Relevant | | | 5 | 5 | 9 | 16 | 1 | 3 | 11 | 3 | 4 | 5 | | 7 | 2 | 5 | 2 | 4 | 3 | 3 | 4 | 1 | | 5 | 7 | | 2 | | 4 | 1 | 6 | 6 | 1 | 8 | 2 | 1 | 10 | 1 | 1 | 1 | 8 | 6 | 1 | | 3.64 |
| UMCP-TDN-TOFS-L2-CF | 1 | 1 | 0.54 | 0.56 | 0.55 | 0.56 | 0.55 | 0.61 | 0.47 | 0.34 | 0.42 | 0.47 | 0.35 | 0.44 | 0.35 | 0.20 | 0.24 | 0.09 | 0.28 | | 0.16 | | 0.20 | 0.08 | 0.17 | | 0.17 | 0.30 | | | 0.05 | 0.16 | | | | | | | | | | | | | | 0.22 |
| UMCP-TD-TOFS-L2-CF | 1 | 1 | 0.54 | 0.56 | 0.55 | 0.56 | 0.55 | 0.53 | 0.51 | 0.47 | 0.42 | 0.49 | 0.55 | 0.32 | 0.47 | 0.13 | 0.24 | 0.13 | 0.28 | 0.19 | | | 0.23 | 0.08 | 0.32 | | 0.17 | | | | | | | | | | | | | | | | | | | 0.22 |
| UMCP-T-TOFS-L2-CF | 1 | 1 | 0.54 | 0.61 | 0.55 | 0.48 | 0.55 | 0.37 | 0.55 | 0.55 | 0.42 | 0.49 | 0.49 | 0.44 | 0.39 | 0.20 | 0.24 | 0.13 | 0.28 | | 0.22 | | 0.23 | 0.08 | 0.17 | | 0.17 | | | | | | | | | | | | | | | | | | | 0.22 |
| UMCP-TDN-TOFS-L2 | 1 | 1 | 0.72 | 0.61 | 0.55 | 0.56 | 0.55 | 0.61 | 0.55 | 0.34 | 0.42 | 0.55 | 0.38 | 0.41 | 0.30 | 0.20 | 0.24 | 0.13 | 0.11 | | 0.16 | | | 0.08 | | | | | | | | 0.16 | | | | | | | | | | | | | | 0.21 |
| UMCP-TD-TOFS-L2 | 1 | 1 | 0.72 | 0.6 | 0.55 | 0.56 | 0.55 | 0.56 | 0.55 | 0.34 | 0.42 | 0.55 | 0.55 | 0.33 | 0.26 | 0.10 | 0.24 | 0.21 | 0.11 | | 0.08 | | | 0.08 | 0.21 | | | | | | | | | | | | | | | | | | | | | 0.21 |
| UMCP-TD-TOFS-U2-CF | 1 | 1 | 0.54 | 0.56 | 0.55 | 0.55 | 0.55 | 0.37 | 0.32 | 0.55 | 0.42 | 0.47 | 0.34 | 0.26 | 0.43 | 0.20 | 0.24 | 0.2 | 0.22 | | 0.16 | | 0.23 | 0.08 | 0.17 | | 0.15 | | | | | | | | | | | | | | | | | | | 0.21 |
| UMCP-T-TOFS-L2 | 1 | 1 | 0.58 | 0.61 | 0.55 | 0.55 | 0.55 | 0.37 | 0.55 | 0.55 | 0.42 | 0.55 | 0.55 | 0.41 | 0.26 | 0.13 | 0.24 | 0.21 | 0.11 | | 0.23 | | | 0.08 | | | | | | | | | | | | | | | | | | | | | | 0.21 |
| UMCP-TDN-TOFS-U2-CF | 1 | 1 | 0.54 | 0.56 | 0.55 | 0.48 | 0.55 | 0.52 | 0.28 | 0.21 | 0.42 | 0.34 | 0.34 | 0.40 | 0.30 | 0.20 | 0.24 | 0.2 | 0.17 | | 0.16 | | 0.23 | 0.08 | 0.17 | | 0.17 | 0.3 | | | 0.06 | | | | | | | | | | | | | | | 0.21 |
| UMCP-T-TOFS-U2-CF | 1 | 1 | 0.54 | 0.61 | 0.55 | 0.55 | 0.55 | 0.37 | 0.34 | 0.55 | 0.42 | 0.47 | 0.13 | 0.31 | 0.30 | 0.20 | 0.24 | 0.20 | 0.24 | | 0.22 | | 0.23 | 0.08 | 0.17 | | 0.17 | | | | | | | | | | | | | | | | | | | 0.21 |
| TerrierBaseline-T | 1 | 1 | 0.72 | 0.61 | 0.55 | 0.55 | 0.55 | 0.37 | 0.51 | 0.55 | 0.42 | 0.51 | 0.17 | 0.36 | 0.09 | 0.20 | 0.24 | 0.13 | 0.07 | | 0.16 | | | 0.08 | | | | | 0.34 | | | | | | | | | | | | | | | | | 0.20 |
| KASYS-1 | 1 | 1 | 0.7 | 0.68 | 0.34 | 0.58 | 0.55 | 0.56 | 0.21 | 0.55 | 0.42 | 0.34 | 0.34 | 0.56 | 0.16 | 0.10 | 0.24 | 0.1 | 0.11 | | 0.26 | 0.12 | | 0.08 | 0.13 | | | | | | | | | | | | | | | | | | | | 0.20 |
| TerrierBaseline-TDN | 1 | 1 | 0.72 | 0.58 | 0.55 | 0.48 | 0.55 | 0.23 | 0.55 | 0.55 | 0.42 | 0.51 | 0.17 | 0.21 | 0.14 | 0.20 | 0.24 | 0.23 | 0.07 | 0.48 | 0.16 | | | 0.05 | | | | | | | | | | | | | | | | | | | | | | 0.20 |
| UMCP-TDN-O-B | 1 | 1 | 0.72 | 0.58 | 0.49 | 0.48 | 0.55 | 0.37 | 0.55 | 0.55 | 0.42 | 0.47 | 0.21 | | 0.14 | 0.20 | 0.24 | 0.33 | 0.06 | 0.48 | 0.16 | | | | | | | | | | | | | | | | | | | | | | | | | 0.20 |
| TerrierBaseline-TD | 1 | 1 | 0.72 | 0.61 | 0.51 | 0.48 | 0.55 | 0.37 | 0.38 | 0.55 | 0.42 | 0.51 | 0.30 | 0.19 | 0.25 | 0.20 | 0.24 | 0.13 | 0.07 | | 0.16 | | | 0.08 | | | | | 0.13 | | | | | | | | | | | | | | | | | 0.19 |
| UMCP-T-O-B | 1 | 1 | 0.58 | 0.61 | 0.55 | 0.55 | 0.55 | 0.37 | 0.55 | 0.55 | 0.42 | 0.47 | | 0.13 | 0.07 | 0.20 | 0.24 | 0.2 | 0.06 | | 0.16 | | | | | | | | 0.34 | | | | | | | | | | | | | | | | | 0.19 |
| UMCP-TDN-S-B | 1 | 1 | 0.70 | 0.58 | 0.61 | 0.48 | 0.55 | 0.52 | 0.55 | 0.34 | 0.42 | 0.41 | 0.15 | 0.11 | 0.14 | 0.20 | 0.24 | 0.33 | 0.11 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.18 |
| UMCP-ColBERT-T | 1 | 1 | 0.72 | 0.61 | 0.55 | 0.48 | 0.55 | 0.53 | 0.51 | 0.34 | 0.42 | 0.17 | 0.17 | 0.19 | 0.18 | 0.20 | | 0.13 | 0.11 | 0.21 | 0.16 | | | 0.08 | | | | | | | | | | | | | | | | | | | | | | 0.18 |
| UMCP-TDN-T-B | 1 | 1 | 0.86 | 0.70 | 0.55 | 0.58 | 0.34 | 0.61 | 0.34 | | 0.42 | 0.34 | 0.49 | 0.15 | 0.21 | 0.13 | | 0.13 | 0.05 | | | | | | 0.34 | | | | | | | | | | | | | | | | | | | | 0.18 |
| UMCP-ColBERT-TD | 1 | 1 | 0.72 | 0.50 | 0.55 | 0.55 | 0.55 | 0.37 | 0.35 | 0.32 | 0.42 | 0.15 | 0.34 | 0.05 | 0.26 | 0.20 | | 0.13 | | 0.30 | | 0.17 | | 0.08 | | | 0.17 | | | | | | | 0.05 | | | | | | | | | | 0.18 |
| UMCP-TD-O-B | 1 | 1 | 0.72 | 0.58 | 0.38 | 0.48 | 0.55 | 0.37 | 0.55 | 0.55 | 0.42 | 0.34 | 0.13 | 0.05 | 0.09 | 0.20 | 0.24 | 0.2 | 0.06 | | 0.16 | | | | | | | | 0.13 | | | | | | | | | | | | | | | | | 0.18 |
| UMCP-TD-S-B | 1 | 1 | 0.70 | 0.58 | 0.60 | 0.48 | 0.55 | 0.52 | 0.55 | 0.21 | 0.42 | 0.4 | 0.21 | | 0.09 | 0.20 | 0.24 | 0.20 | 0.11 | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.17 |
| UMCP-T-S-B | 1 | 1 | 0.69 | 0.58 | 0.60 | 0.48 | 0.55 | 0.37 | 0.51 | 0.34 | 0.42 | 0.53 | | | 0.09 | 0.20 | 0.24 | 0.2 | 0.11 | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.17 |
| UMCP-ColBERT-TDN | 1 | 1 | 0.56 | 0.5 | 0.47 | 0.48 | 0.55 | 0.19 | 0.32 | 0.3 | 0.42 | 0.15 | 0.34 | | 0.18 | 0.20 | | | | 0.24 | | | | | | | 0.17 | 0.47 | | | 0.25 | | | | | | | | | | | | | | | 0.17 |
| UMCP-ColBERT-D | 0.39 | 1 | 0.72 | 0.46 | 0.51 | 0.56 | 0.51 | 0.53 | 0.51 | 0.32 | 0.42 | | 0.34 | 0.07 | | 0.20 | | | | 0.48 | | 0.28 | | 0.06 | | | | | | | | | | | | | | | | | | | | | | 0.16 |
| UMCP-TD-T-B | 1 | 1 | 0.42 | 0.64 | 0.51 | 0.47 | 0.34 | 0.61 | 0.34 | | 0.42 | 0.21 | 0.49 | 0.13 | 0.26 | | | 0.13 | 0.07 | | | | | | 0.17 | | | | | | | | | | | | | | | | | | | | 0.16 |
| UMCP-T-T-B | 1 | 1 | 0.61 | 0.34 | 0.55 | 0.39 | 0.34 | 0.61 | 0.13 | | 0.42 | 0.34 | 0.51 | 0.13 | 0.26 | 0.09 | | 0.13 | 0.11 | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.15 |
| KASYS-5 | 1 | 1 | 0.37 | 0.15 | 0.49 | 0.48 | 0.21 | 0.42 | | 0.34 | 0.18 | 0.15 | 0.3 | 0.39 | | | | 0.09 | 0.11 | | 0.08 | | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.13 |
| QshuNLP-GPT-1 | | | | 0.43 | 0.11 | | | 0.37 | 0.34 | | | | 0.49 | 0.76 | | | 0.24 | 0.23 | 0.13 | | 0.06 | 0.32 | 0.77 | 0.08 | | | | | | 0.81 | 0.26 | | 0.22 | 0.05 | | | | | | | | | | 0.12 |
| KASYS-4 | 1 | 1 | 0.37 | 0.15 | 0.17 | | 0.34 | 0.38 | 0.17 | 0.34 | 0.21 | 0.17 | | 0.46 | | 0.13 | | 0.08 | 0.05 | | | 0.28 | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.12 |
| KASYS-8 | 1 | 1 | 0.37 | 0.15 | 0.17 | | 0.34 | 0.38 | 0.17 | 0.34 | 0.21 | 0.17 | | 0.46 | | 0.13 | | 0.08 | 0.05 | | | 0.28 | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.12 |
| KASYS-3 | 1 | 0.50 | 0.35 | 0.49 | | | 0.17 | | | | | | | 0.07 | | 0.13 | | | | 0.19 | | 0.28 | | | | 0.47 | | | | | | | | | | | | | | | | | | | 0.08 |
| KASYS-7 | 1 | 0.50 | 0.35 | 0.49 | | | 0.17 | | | | | | | 0.07 | | 0.13 | | | | 0.19 | | 0.28 | | | | 0.47 | | | | | | | | | | | | | | | | | | | 0.08 |
| KASYS-2 | 1 | 0.50 | | 0.17 | | | | | | | | | | 0.06 | | 0.20 | | | | 0.21 | | 0.28 | | | | 0.23 | | | | | | | | | | | | | | | | | | | 0.05 |
| KASYS-6 | 1 | 0.5 | | 0.17 | | | | | | | | | | 0.06 | | 0.20 | | | | 0.21 | | 0.28 | | | | 0.23 | | | | | | | | | | | | | | | | | | | 0.05 |
| UMCP-TDN-F-B | | | 0.41 | | 0.17 | 0.37 | | 0.23 | | 0.34 | | 0.13 | 0.21 | 0.47 | | | | | | | | 0.17 | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.05 |
| UMCP-TD-F-B | | | | | 0.17 | 0.37 | | 0.23 | 0.13 | 0.34 | | | 0.21 | 0.50 | | | | | | | | | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.04 |
| UMCP-T-F-B | | | | | 0.34 | 0.37 | | | 0.13 | 0.34 | | | 0.21 | 0.50 | | | | | | | | | | 0.12 | | | | | | | | | | | | | | | | | | | | | | 0.04 |
| Topic Mean | 0.88 | 0.84 | 0.52 | 0.47 | 0.42 | 0.41 | 0.40 | 0.37 | 0.34 | 0.32 | 0.31 | 0.29 | 0.26 | 0.26 | 0.15 | 0.15 | 0.13 | 0.13 | 0.09 | 0.09 | 0.08 | 0.07 | 0.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | 0.00 | | | | | | | | | | | | 0.16 |
| Box Highly Relevant | 1 | 1 | 4 | 5 | 4 | 2 | 5 | 4 | 12 | 5 | 2 | 8 | 7 | 5 | 2 | 1 | 1 | 1 | 5 | 2 | 2 | 1 | 2 | 4 | | 3 | 5 | 1 | 8 | 2 | 3 | 2 | 2 | 3 | 3 | 1 | 2 | 5 | 1 | 2 | 1 | 1 | 1 | 3 | | 3.00 |
| Box Relevant | | | 2 | 4 | 7 | 9 | 1 | 3 | 9 | 2 | 4 | 3 | | 6 | 2 | 5 | 2 | 2 | | 3 | 2 | 1 | | 5 | 6 | | 2 | | 4 | 1 | 3 | 5 | 1 | 8 | 2 | 1 | 9 | 1 | 1 | 1 | | 7 | 6 | 1 | 2.91 |

## 2.9 Making Sense of the Results

Putting all of these results together, we can explain the behavior of these search systems in this way. Imagine that a searcher comes to an archive that has digitized five documents from every box and that has one of the best systems that was submitted to this evaluation (UMCP-T-TOFS-L2-CF), and that they type a short query into that system. The system will then recommend a few boxes to them. If they request only the system's highest-ranked box, we can see from Table 4 that they have a 47% chance that they will find something that they think is relevant somewhere in that box. That system can also suggest which folder in that box to look through. If they choose to look through only the system's highest-ranked folder in its highest-ranked box, we can see from Table 3 that their chance of finding something relevant goes down a bit, from 47% to 42%. However, if they choose to look at the system's five highest-ranked folders (which might require requesting as many as five different boxes) they can do a bit better – their chances of finding something relevant goes up from 42% to 53% (this is based on 24 of 45 topics having a non-zero nDCG@5 in Table 5).

Now imagine instead that this archive has not yet installed a system like the ones that submitted runs for this evaluation. The searcher will instead need to search the metadata records of the archive to find what they are looking for. In such a case the searcher might also solicit assistance from an archivist, who would be familiar with how the collection is organized, what metadata exists for the collection, how that metadata can be interpreted, and how the existing metadata-based search systems at that archive work. We can not fully simulate the effect of that expert assistance using the evaluation results that we have, but one system that we know searched only folder metadata (UMCP-T-F-B) recommended a box that contained at least one relevant document just 18% of the time. on the bright side, its recommendation of which folder to look at in that box was also correct 18% of the time. On balance, it does appear that systems of the type that have submitted their results to this evaluation might well be of some value to some searchers, both for suggesting additional boxes to examine (beyond those they might be able to find in other ways) and for suggesting which folders in those boxes might be given the highest priority for examination if the searcher's time is limited.

# 3 Subtask B: Archival Reference Detection

In Subtask B, we defined two tasks—Archival Reference Detection and Archival Reference Boundary Detection—focusing on footnotes and endnotes found in academic papers. The Archival Reference Detection task is, given the text of a footnote or endnote (in isolation), detect whether that text contains one or more references to the location of specific documents (or other information objects) in some archival repository. Second, Archival Reference Boundary Detection is the task of identifying the start and end character positions of those archival references. The following are examples of footnotes or endnotes containing archival references.

- Roosevelt to Secretary of War, June 3, 1939, Roosevelt Papers, O.F. 268, Box 10; unsigned memorandum, Jan. 6, 1940, ibid., Box 11.
- Wheeler, D., and R. García-Herrera, 2008: Ships' logbooks in climatological research: Reflections and prospects. Ann. New York Acad. Sci., 1146, 1-15, doi:10.1196/annals.1446.006. Several archive sources have been used in the preparation of this paper, including the following: Logbook of HMS Richmond. The U.K. National Archives. ADM/51/3949

For Subtask B, we crawled papers from the field of History using Semantic Scholar, and then used GROBID [3] to extract the text of footnotes or endnotes in those papers to develop a test collection. The annotation of whether a footnote or endnote contained one or more archival references was made after the submission of runs by participating teams, using sampling strategy. For the formal run of Subtask B, a total of 7 submissions were received from two teams, Kyushu University (QshuNLP) [9] and the University of Maryland (UMCP) [5]. The breakdown of the runs for each task is shown in Table 6.

This section explains the test collection for Subtask B, the annotation process for footnotes and endnotes, and the results of the submitted runs.

**Table 6: Number of submissions for Subtask B.**

| Team | Reference Detection | Boundary Detection |
|---|---|---|
| QshuNLP | 2 | 1 |
| UMCP | 5 | 0 |

## 3.1 Test Collection

*Stratified Sampling.* In Subtask B, two teams, QshuNLP and UMCP, participated. We received seven runs, in which (after deduplication) at least one run classified a total of 9,394 footnotes or endnotes as archival references. Among these, only one run (QshuNLP-B2) included results for the Archival Reference Boundary Detection Task. We conducted relevance judgments in January 2025 with a single assessor, a PhD student in Library Science.

Due to the assessor constraints, we applied stratified sampling to select footnotes or endnotes for annotation. We defined 5 strata based on agreement between the submitted runs, as follows.

- Stratum 1 consists of texts (i.e., footnotes or endnotes) classified as archival references in all seven runs.
- Strata 2, 3, and 4 include texts classified as archival references in any five or six runs, any three or four runs, and any one or two runs, respectively.
- Stratum 5 consists of texts that were not classified as archival references in any submitted run.

The statistics for the stratified sampling are summarized in Table 7. Since we expect more actual archival references in texts with greater system agreement, we allocated more samples to strata with higher agreement. For Stratum 1, which is very small, we sampled at 100%. The sampling rates are about 57% for strata 2 and 3; it is lower for stratum 4 at about 44%. We very sparsely sampled texts from stratum 5, with a 0.2% sampling rate, in order to get a rough estimate for the number of archival references that all systems had missed. The overall positive sampling rate for each system's positive classifications are shown in Table 8; all are between 41% and 46%. These sampling rates result in relatively tight confidence intervals on precision, as seen in Table 9.

**Table 7: Stratified sampling overview.**

| Stratum | Runs | Footnotes | Samples | Sampling rate |
|---|---|---|---|---|
| 1 | 7 | 4 | 4 | 100% |
| 2 | 5, 6 | 3,145 | 1,800 | 57.2% |
| 3 | 3, 4 | 2,596 | 1,500 | 57.8% |
| 4 | 1, 2 | 3,604 | 1,600 | 44.4% |
| 5 | 0 | 865,106 | 1,500 | 0.2 % |

*Annotation.* We hired as an annotator the Ph.D. student who served as an assessor for Subtask A, who has experience conducting research on archival institutions. Before the annotator began the annotation, we created an annotation guide and provided instruction using that guide. In that instruction, we introduced the goals of Subtask B and then outlined the annotation rules and the annotator's work.

In the annotation guide, we defined the annotation rule as follows: Footnotes or endnotes that the annotator can recognize as expected to be useful for finding the location of a specific information object in an archival repository should be marked as an archival reference. We provided a detailed explanation of three specific criteria for applying this rule.

- An information object is a physical or digital container whose main purpose is to convey information. Examples include documents, films, photographs, or audio recordings. Physical objects that principally serve some other purpose (e.g., statues, jewelry, biological specimens, or artwork) were not considered information objects, even if they incidentally include information (e.g., as inscriptions on statues or signatures on paintings).
- An archival repository is a collection whose purpose is, at least in part, to collect certain types of rare and often unique information objects, but only in cases in which that collection has been created by an institution that had curatorial intent when creating that collection.
- A location in an archival repository could be general (e.g., the name of an archival repository) or specific (e.g., Fonds, Box or Folder). A location need not be complete to be useful because it is reasonable to assume that such specific locations would have been contextualized within the document from which the footnote or endnote had been extracted.

We drew two samples of 2,687 and 2,869 footnotes and endnotes. We did this because we were uncertain about whether there would be sufficient time to annotate both batches. We chose the size of the second batch based on the estimated annotation time for the first batch.

The annotation was performed using a spreadsheet, where each row contained the text of the footnote or endnote to be annotated, along with a "Decision" column and a "More" column in which the annotation was to be made. If the annotator determined that the text was an archival reference, they would enter 1 in the "Decision" column, otherwise, they would enter 0. If, however, they were unable to make a confident decision, they could enter u (for "unknown"). If the footnote or endnote being annotated contained additional contained text that was not a part of an archival reference, they would enter "x" in the "More" column. Clarification questions were sometimes sent to the organizers by email.

After completing each batch, we reviewed the annotations and provided feedback. For example, 205 footnotes and endnotes that were initially marked as 1 in the first batch were later corrected by the annotator to 0 (or to unknown) after discussion with the organizers. The most common reason for making that change was the annotator's not having initially understood that only footnotes or endnotes that included information about the location of some specific item could be an archival reference according to our definition.

Ultimately, a total of 5,376 footnotes and endnotes were annotated, of which 2,728 were marked by the annotator as archival references.

After completing annotations for the Archival Reference Detection task (Batch 1 and Batch 2), we conducted annotation for the Archival Reference Boundary Detection task with the same assessor. We randomly selected 230 footnotes or endnotes from Batch 1 that had been marked by the annotator as containing archival references (115 of them were checked in the "More" column, and the rest had not such annotation), as well as another 50 footnotes or endnotes that been marked as not containing any archival references (with 0 in the "Decision" column). As with the annotations for the Archival Reference Detection task, before beginning the annotation, we created an annotation guide and provided instruction to the annotator.

In this instruction, we first shared our broad goal for the Archival Reference Boundary Detection (which was to use archival references found in the literature to help scholars find specific items that they would like to see). Extracting archival references from footnotes or endnotes was a first step that could later be followed by identifying specific information such as the name of the archival repository or the box number containing an item. Then, we provided an explanation of the annotation task, which was to copy each archival reference from the text of a complete footnote or endnote into a separate field that would contain only a single minimal complete archival reference.

We prepared a spreadsheet where each row corresponds to the text of a footnote or endnote to be annotated. The columns included the "Archival Reference Span", which records the start and end positions of the first reference in the text, as well as individual columns for each archival reference, such as "Archival Reference 1", "Archival Reference 2", and so on. The annotation process was conducted by copying a single minimal contiguous text span for an archival reference from a footnote or endnote into each corresponding cell. Later, we used substring matching to calculate the start and end character positions for each archival reference in the original footnote or endnote.

This annotation process found archival reference spans in 228 footnotes or endnotes that could be used for evaluation of the Boundary Detection task. Among those 228, 48 contained two or more more archival references, with a mean in those 228 of 1.6 and a maximum of 31.

**Table 8: Per-system overall positive sampling rate.**

| System | System detections | Sampling rate |
|---|---|---|
| UMCP-SGDC | 4,464 | 44.4% |
| UMCP-GPT-SGDC | 7,219 | 44.9% |
| UMCP-GPT-SGDC-U-2-30 | 7,804 | 44.3% |
| UMCP-SGDC-U-3-75 | 5,060 | 46.4% |
| UMCP-GPTauto-SGDC | 4,857 | 44.8% |
| QshuNLP-B1 | 231 | 43.3% |
| QshuNLP-B2 | 272 | 40.8% |

## 3.2 Evaluation Measures

Because archival reference detection is a binary classification task with highly skewed class prevalence, we use precision, recall, and $F_1$ to characterize classification effectiveness. Because stratified sampling was used, these values were computed from estimates rather than directly from measured values.

We did this by first estimating the true number of archival references in each stratum. Since we know the sampling rate for each stratum, we estimate the true number in that stratum by multiplying the number found in that stratum by the annotator by the multiplicative inverse of the sampling rate for that stratum. For example, if 100 archival references had been found by the annotator in a stratum that had been sampled at 50%, we would estimate that that there were really likely to be about 200 archival references in that stratum. To get an estimate for the total number of archival references in the collection, we then summed those per-stratum estimates. We estimated the number of correct positive system decisions (i.e., system detections) for each stratum and for the full collection in the same way.

We assigned each stratum a sampling rate, denoted $s_1, s_2, s_3, s_4, s_5$. For stratum $i$, we define the observed number of true positives in the stratum as $TP_i$, false positives as $FP_i$, and false negatives as $FN_i$. The estimated actual number of true positives, false positives and false negatives in stratum $i$ is then $\hat{TP}_i = TP_i/s_i$, $\hat{FP}_i = FP_i/s_i$ and $\hat{FN}_i = FN_i/s_i$, respectively. With these preliminaries, we then estimate Precision and Recall as follows.

$$\text{Precision} = \frac{\sum_{i=1}^{5} \hat{TP}_i}{\sum_{i=1}^{5} (\hat{TP}_i + \hat{FP}_i)},$$

$$\text{Recall} = \frac{\sum_{i=1}^{5} \hat{TP}_i}{\sum_{i=1}^{5} (\hat{TP}_i + \hat{FN}_i)}.$$

We then estimated $F_1$ as the harmonic mean of our estimates for Precision and Recall. We also calculated the 95% confidence intervals for Precision, Recall and $F_1$. For $F_1$, we estimated the confidence interval using bootstrap resampling, with 1,000 iterations.

For the Archival Reference Boundary Detection task, we evaluated the results using a character-based Jaccard coefficient. This value is computed as the number of characters in the intersection between the system's output and the annotators ground truth, divided by the number of characters in the union, as follows:

$$Jaccard(G, S) = \frac{|G \cap S|}{|G \cup S|},$$

where $G$ denotes the span for archival reference in the ground truth annotation and $S$ denotes the span for that in a system's submission.

## 3.3 Dry Run

We held the dry run for Subtask B by distributing the dry-run test collection on July 1, 2024, and setting the submission deadline for August 31, 2024. As SUSHI is a pilot task, with Subtask B being run for the first time, the purpose of the dry run was to identify and resolve any potential issues in our task organization. Specifically, we wanted to verify whether the guidelines provided sufficient information for participants, whether our evaluation program correctly processed the submission format, and whether the overall process functioned without complications. The dry run collection could also serve as training data for participating teams who wished to use it in that way.

With those objectives in mind, we distributed 1,868 fully annotated footnotes and endnotes as the dry run test collection. We also distributed two python programs with this data, one that implemented a term-matching baseline for archival reference detection (in which footnotes or endnotes containing terms such as "archive" or "box" would be classified as archival references), and one to perform evaluation. Both were limited to the Archival Reference Detection; no dry run was performed for Archival Reference Boundary Detection task because at the time the dry run was conducted we had no annotated ground truth for Archival Reference Boundary Detection.

The dry run test collection was based on previously annotated data for the Archival Reference Detection task [7]. The collection included 671 positive examples and 1,165 negative examples. Details on the annotation process and the term-matching baseline method can be found in Suzuki et al. [7].

We received no submissions to the Subtask B dry run, but we did use the dry run for its intended purpose of verifying the compatibility of our submission formats and evaluation program, and the dry run test collection was then available to participating teams for use in system training or formative evaluation.

## 3.4 Results

Table 9 summarizes the results for the seven runs by each evaluation measure. The runs submitted by UMCP achieved higher precision, with a maximum of 0.774. Recall is, however, rather low for all systems, with none higher than 0.119. The confidence intervals on recall are quite tight (none are larger than ±0.003), but we note that these confidence intervals reflect only sampling error, and not measurement error.

To see the potential effect of measurement error, consider the effect of one incorrect judgment in stratum 5, for which the sampling rate was 0.1734%. At that sampling rate, one incorrect positive judgment by the assessor would increase our estimate for the number of archival references in the collection by 527. The best system by $F_1$, UMCP-GPT-SGDC-U-2-30, found an estimated 5,424 positive examples out of an estimated 45,580 total positive examples in the collection, yielding an estimated Recall of 0.119. If we were to assume that an annotator working for several days to create several thousand annotations might make a net of 5 mistakes in either

**Table 9: Subtask B results.**

| Run | System Yes | Est Correct Yes | Precision | Recall | $F_1$ | Jaccard |
|---|---|---|---|---|---|---|
| UMCP-GPT-SGDC-U-2-30 | 7,804 | 5,424 | 0.695 ± 0.009 | 0.119 ± 0.003 | 0.155 ± 0.006 | – |
| UMCP-GPT-SGDC | 7,219 | 5,061 | 0.701 ± 0.010 | 0.110 ± 0.003 | 0.145 ± 0.005 | – |
| UMCP-SGDC-U-3-75 | 5,060 | 3,916 | 0.774 ± 0.011 | 0.081 ± 0.002 | 0.111 ± 0.004 | – |
| UMCP-GPTauto-SGDC | 4,857 | 3,623 | 0.746 ± 0.011 | 0.074 ± 0.002 | 0.099 ± 0.004 | – |
| UMCP-SGDC | 4,464 | 2,973 | 0.666 ± 0.013 | 0.061 ± 0.002 | 0.082 ± 0.004 | – |
| QshuNLP-B1 | 272 | 103 | 0.379 ± 0.058 | 0.002 ± 0.000 | 0.004 ± 0.001 | – |
| QshuNLP-B2 | 231 | 70 | 0.303 ± 0.049 | 0.002 ± 0.000 | 0.004 ± 0.001 | 0.480 |

direction in stratum 5, then the effect of that measurement error would be ±0.008, which is somewhat larger than the confidence intervals shown in Table 9. We don't know how much of an effect to attribute to measurement error because we have no dual annotation on this collection to ground such an estimate, but for now we would suggest considering the combined confidence interval on recall to be at least as large as the confidence intervals on precision for the best systems (i.e., at least 0.010). But regardless of the specific size of that confidence interval, we can surely say from these results that there is clearly still substantial room for improvement in recall.

Only one run, QshuNLP-B2, performed the Archival Reference Boundary Detection task, achieving a Jaccard coefficient of 0.480. Choosing the first half of each footnote or endnote as the archival reference (a very simple baseline) would have yielded a Jaccard coefficient of 0.512, so there is clearly work still to be done on Archival Reference Boundary Detection. We note that future Archival Reference Boundary Detection systems will have the benefit of training data from the annotations, whereas this year's one system lacked that resource.

## 4 Conclusion and Future Work

This SUSHI pilot task has produced two new test collections, and participating teams have made substantial improvements to the state of the art on both subtasks. The test collection for the Folder Ranking subtask includes more comprehensive metadata than earlier collections, and we have shown how an evaluation design first developed in the context of box ranking can be extended to the more challenging evaluation setting of folder ranking. The best submitted systems for that subtask achieved nDCG@5 scores about 25% better, relative to those achieved by baselines that represent the prior state of the art. In the Archival Reference Detection subtask, the best submitted systems achieved better than 77% precision while finding very many more archival references than did earlier systems, although sampling for estimation of recall clearly indicates that further room for improvement in recall exists.

One puzzle in this first instance of the SUSHI task has been the low participation. It is simply not necessary to run an entire NTCIR task for a single participating team (excluding here the participating teams that included task organizers, who of course don't need an NTCIR task to work with themselves!). Perhaps the existence of the test collection, together with this year's results, will help to interest additional teams in participating in a future evaluation. Expanding the set of subtasks being evaluated to include full-collection search might also be of interest to teams who are interested in OCR search, metadata search, or the combination of the two.

## Acknowledgments

This work has been supported in part by the Japan Society for the Promotion of Science KAKENHI Grant Number 23KK0005 and the National Institute of Informatics Open Collaborative Research 2024 (24S0505). This work would not have been possible without the tireless efforts of our assessors: Maia Johnston, Nicolas Lizarralde, and Widiatmoko Adi Putranto.

## References

[1] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 282–289.

[2] Haruki Fujimaki and Makoto P. Kato. 2025. KASYS at the NTCIR-18 SUSHI Task. In *Proceedings of NTCIR-18*.

[3] Patrice Lopez et al. 2025. GROBID. https://github.com/kermitt2/grobid.

[4] Douglas W. Oard. 2023. Known by the Company It Keeps: Proximity-Based Indexing for Physical Content in Archival Repositories. In *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26-29, 2023, Proceedings (Lecture Notes in Computer Science, Vol. 14241)*. Springer, 17–30. doi:10.1007/978-3-031-43849-3_3

[5] Douglas W. Oard, Shashank Bhardwaj, and Emi Ishita. 2025. Biting into SUSHI: The University of Maryland at NTCIR-18. In *Proceedings of NTCIR-18*.

[6] Tetsuya Sakai, Douglas W Oard, and Noriko Kando. 2021. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Vol. 43. Springer Nature.

[7] Tokinori Suzuki, Douglas W. Oard, Emi Ishita, and Yoichi Tomiura. 2023. Automatically Detecting References from the Scholarly Literature to Records in Archives. In *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration - 25th International Conference on Asia-Pacific Digital Libraries, ICADL 2023, Taipei, Taiwan, December 4-7, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14458)*, Dion Hoe-Lian Goh, Shu-Jiun Chen, and Suppawong Tuarob (Eds.). Springer, 100–107. doi:10.1007/978-981-99-8088-8_9

[8] Tokinori Suzuki, Douglas W Oard, Emi Ishita, and Yoichi Tomiura. 2024. Searching for Physical Documents in Archival Repositories. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2614–2618.

[9] Tokinori Suzuki and Yoichi Tomiura. 2025. QshuNLP's Participation in SUSHI: Systems and Result Analysis. In *Proceedings of NTCIR-18*.

[10] Ellen M Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management* 36, 5 (2000), 697–716.