From Shared Tasks to Benchmark Datasets: Lessons from NTCIR's 26-Year Journey

Part 2: Participants, Experiments, and Key Findings

Makoto P. Kato — National Institute of Informatics / University of Tsukuba

Makoto P. Kato

- Associate Professor at the University of Tsukuba and National Institute of Informatics (NII), Japan
- Research interests: **information retrieval**, information access, evaluation methodologies, user behavior analysis
- Co-organizer of multiple NTCIR tasks, including INTENT, 1CLICK, IMine, MobileClick, OpenLiveQ, and Data Search
- Active participant in the NTCIR community
  - PC Co-chair for NTCIR-12 to NTCIR-15
  - General Co-chair for NTCIR-16 to NTCIR-19
- Also served as:
  - SIGIR 2023 PC Co-chair
  - SIGIR-AP 2026 PC Co-chair

What this hour covers

1. **Breadth of NTCIR tasks** — what we have evaluated, and where NTCIR is distinctive
2. **Spotlight on tasks I co-organized** — INTENT, 1CLICK, OpenLiveQ, DataSearch
3. **Lessons learned**
4. **NTCIR vs TREC vs CLEF** — a General Chair perspective
5. **Running an NTCIR task** — a Program Committee Co-chair perspective

Bridge from Part 1

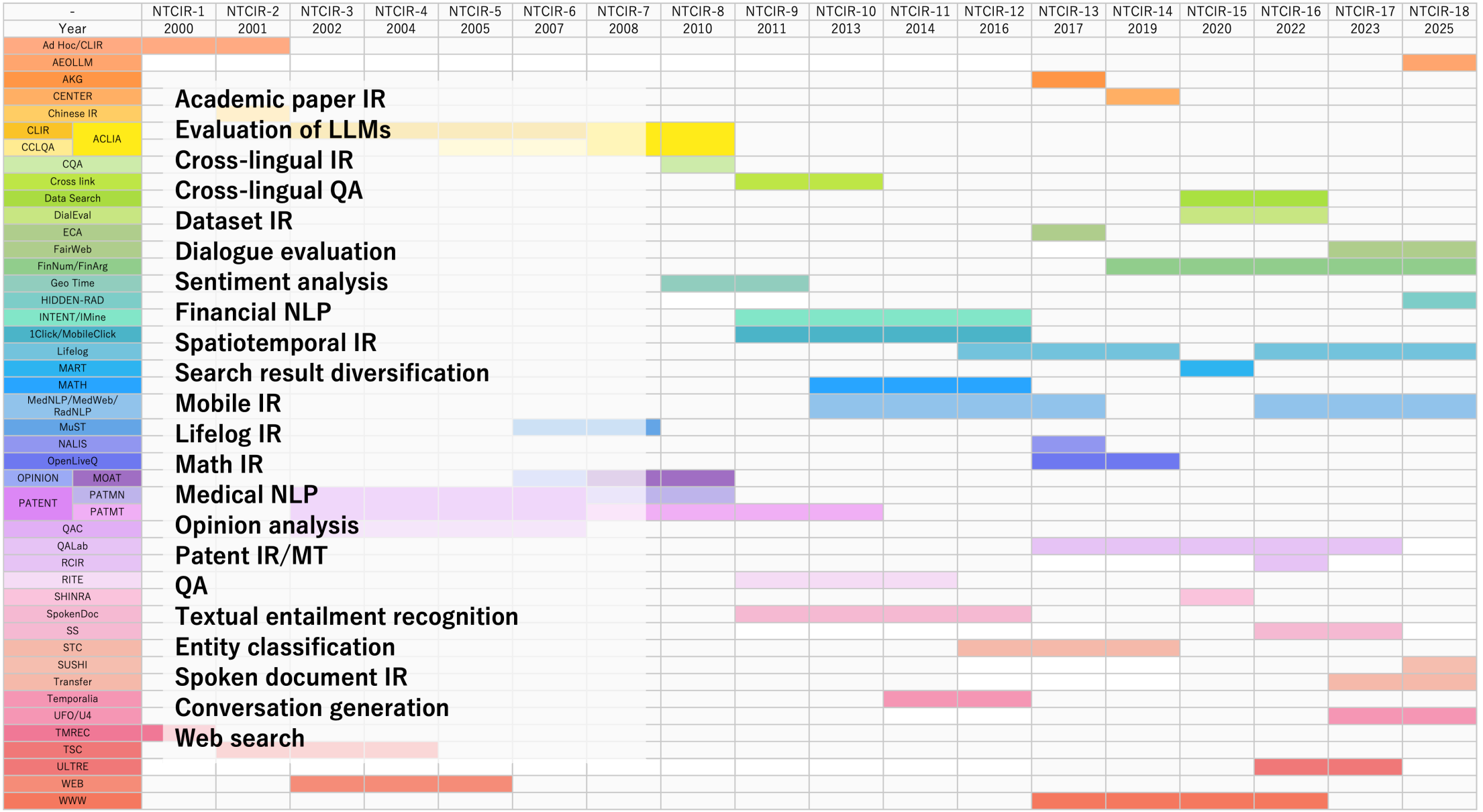
Part 1 covered the *what* and *how* of NTCIR. **Part 2 is about what the community learned** — more than 25 years, from dozens of tasks, and hundreds of teams.

> The most durable contribution of an evaluation campaign is rarely a single winning system. It is the **tasks, test collections, and metrics** that outlive every submission.

Sakai, Oard, Kando (eds.). 2021. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer.

1. Breadth of NTCIR Tasks

A two-decade portfolio of NTCIR tasks



Examples of NTCIR tasks (1/2)

- **Cross-Lingual IR**: Evaluates document retrieval across multiple languages.
- **CLQA**: Finds answers to questions across different languages.
- **Patent Retrieval**: Searches prior art in patent document collections.
- **Patent Machine Translation**: Evaluates specialized machine translation for patent documents.
- **Web Search**: Evaluates Web document retrieval and ranking quality.
- **GeoTime**: Handles search involving geographic and temporal information.
- **Temporalia**: Handles search queries with temporal intent.
- **Math**: Evaluates retrieval of formulas and mathematical documents.
- **Spoken Document Retrieval**: Searches spoken documents using speech recognition results.

Examples of NTCIR tasks (2/2)

- **Recipe Search**: Handles search and recommendation of cooking recipes.
- **Lifelog**: Searches personal lifelog images and records.
- **Data Search**: Evaluates discovery and retrieval of datasets.
- **Dialogue Evaluation (STC, DialEval)**: Evaluates the quality of dialogue system responses.
- **MedNLP**: Handles medical text processing and diagnostic support.
- **FinArg / FinNum**: Analyzes numerical expressions and arguments in financial documents.
- **AEOLLM**: Evaluates automatically generated reports by large language models.
- **FairWeb**: Evaluates fairness in Web search results.
- **Tip-of-the-Tongue**: Searches for targets from vague or partial memories.

What is distinctive about NTCIR's portfolio

- **Asian-language IR** — many tasks focusing on multiple languages including Japanese, Chinese, Korean, and English
- **First, or among the first**, to evaluate:
  - **Patent search and patent MT** — NTCIR-3 (2002)
  - **Math formula retrieval** — NTCIR-10 (2013)
  - **Lifelog retrieval** — NTCIR-12 (2016)
  - **Mobile / one-click access** — NTCIR-9 (2011)
  - **Dataset search** — NTCIR-15 (2020)

Mark Sanderson, Foreword to Sakai/Oard/Kando 2021: "NTCIR was the first to explore patent search, first to incorporate life logging data, and first to examine retrieval of mathematical formulas."

2. Spotlight: NTCIR Tasks I Co-Organized

Six tasks across NTCIR-9 to NTCIR-16

- NTCIR-9, 10 **INTENT** — search intent diversification (subtopic mining + diversified ranking)
- NTCIR-9, 10 **1CLICK** — direct, immediate textual answer to a query
- NTCIR-11, 12 **IMine** — unified intent mining; vertical-aware diversification
- NTCIR-11, 12 **MobileClick** — two-layered summary for mobile zero-click access
- NTCIR-13, 14 **OpenLiveQ** — online evaluation against real Yahoo! Chiebukuro users
- NTCIR-15, 16 **Data Search** — first IR campaign for dataset retrieval (e-Stat / Data.gov)

These are the parts of NTCIR I can answer in depth

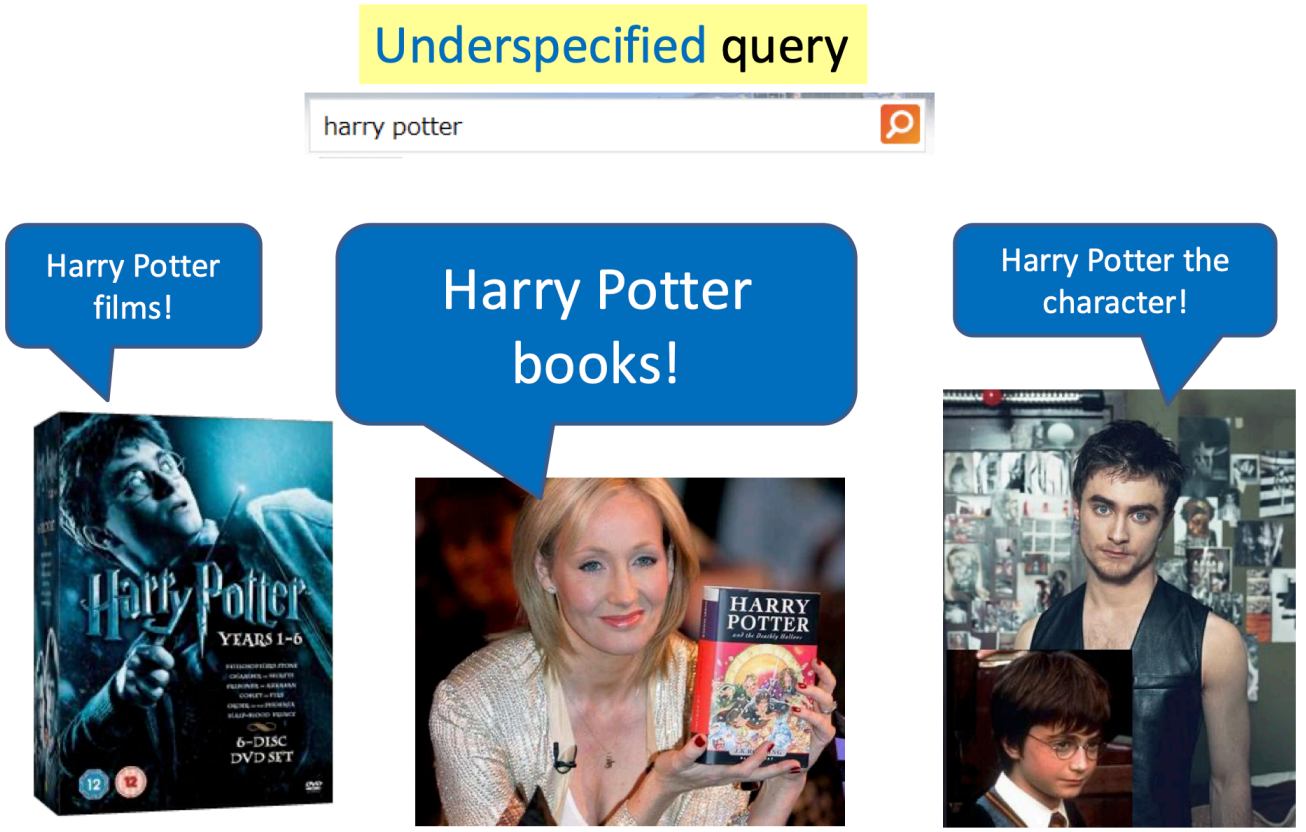
INTENT

INTENT — Ambiguous queries

A query *"apple"* is an **ambiguous** query that can be interpreted in many ways:

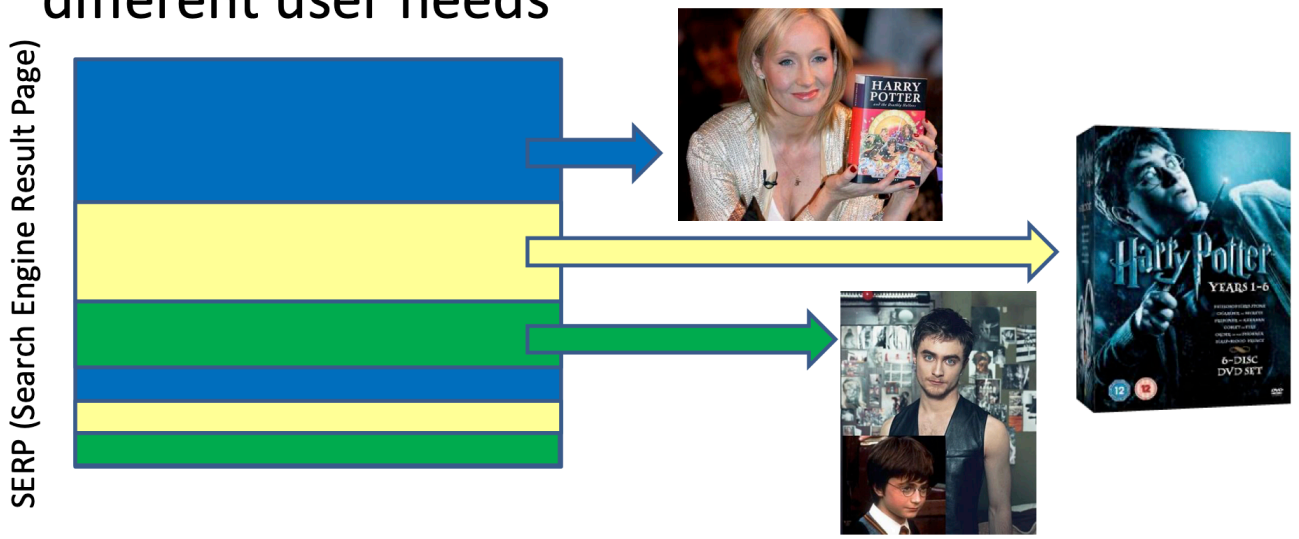
INTENT — Underspecified queries

A query *"harry potter"* is an **underspecified** query:



Why Search Result Diversification?

- If query *q* has multiple intents *i* and we have no knowledge of the user, we should cover different *i*'s in the first SERP to accommodate different user needs



INTENT — Subtopic Mining + Document Ranking

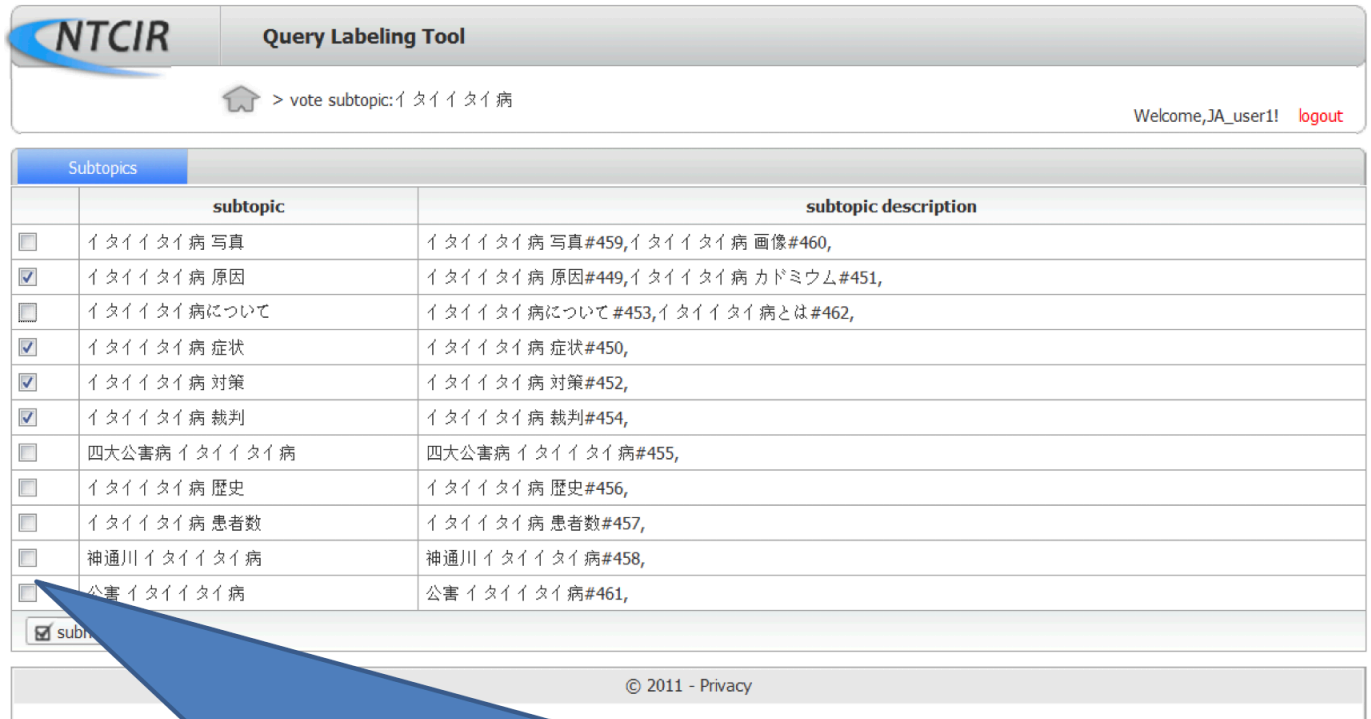
Systems must:

1. **Subtopic Mining (SM)** — return a ranked list of subtopic strings covering the query's intents
2. **Document Ranking (DR)** — return a diversified ranked list satisfying multiple intents

- **Data**
  - Bilingual task (Chinese and Japanese)
  - Intents were created by manual clustering of submitted subtopic strings
  - Documents were judged for relevance to each intent on a 3-point scale (irrelevant, relevant, highly relevant)

Song et al., Overview of NTCIR-9 INTENT (2011); Sakai et al., Overview of NTCIR-10 INTENT-2 (2013).

Intent voting interface



10 assessor's votes used for estimating intent probabilities

Per-intent graded relevance assessment interface

INTENT Test Collection

Table 7: Statistics of the INTENT-1 topics and intents.

		Subtopic Mining	Document Ranking
Chinese	topics	100	100
	intents	917	917
	subtopics	20,354	–
	unique rel docs	–	23,571
Japanese	topics	100	100
	intents	1,091	1,091
	subtopics	4,103	–
	unique rel docs	–	19,841

Table 2. NTCIR-9 INTENT Chinese and Japanese Document Collections.

	Chinese	Japanese
Collection name	SogouT	ClueWeb-JA
#Pages	138 million	67.3 million
Size and Storage Medium	7z files on a 500GB hard disk	tarred/gzipped files on a 500GB hard disk

Table 6: Statistics of the INTENT-2 topics and intents. Those for the revised Subtopic Mining data are shown in parentheses.

		Subtopic Mining	Document Ranking
English	topics	50	–
	intents	392	–
	subtopic strings	4,157 (5,410)	–
Chinese	topics	98	97
	nav topics	23	22
	amb/faceted topics	23/52	23/52
	shared topics	21	21
	reused topics	19	19
	intents	616	615
	nav intents	–	125
	inf intents	–	490
	subtopic strings	6,251 (6,253)	–
	unique rel docs	–	9,295
Japanese	topics	100	95
	nav topics	33	28
	amb/faceted topics	27/40	27/40
	shared topics	21	21
	reused topics	33	33
	intents	587	582
	nav intents	–	259
	inf intents	–	323
	subtopic strings	2,979 (2,989)	–
	unique rel docs	–	5,085

INTENT — Evaluation Metric

- $\text{D}\sharp\text{-nDCG} = \gamma\text{I-rec} + (1-\gamma)\text{D-nDCG}$
  - Official measure of NTCIR-9 INTENT
  - Combines *intent recall* with D-nDCG
    - Intent recall: fraction of subtopics covered by the ranked list
    - D-nDCG: nDCG with a global gain, $\sum_i P(i|q)g_i(d)$, where
      - $P(i|q)$: probability of intent i for a given query q
      - $g_i(d)$: gain (relevance score) of a document d in terms of intent i
- **D♯-nDCG family went on** to be reused by IMine, Temporalia — i.e. the metric outlived the task

> The technical legacy of INTENT is not a winning system — it is a **family of diversification metrics** still used a decade later.

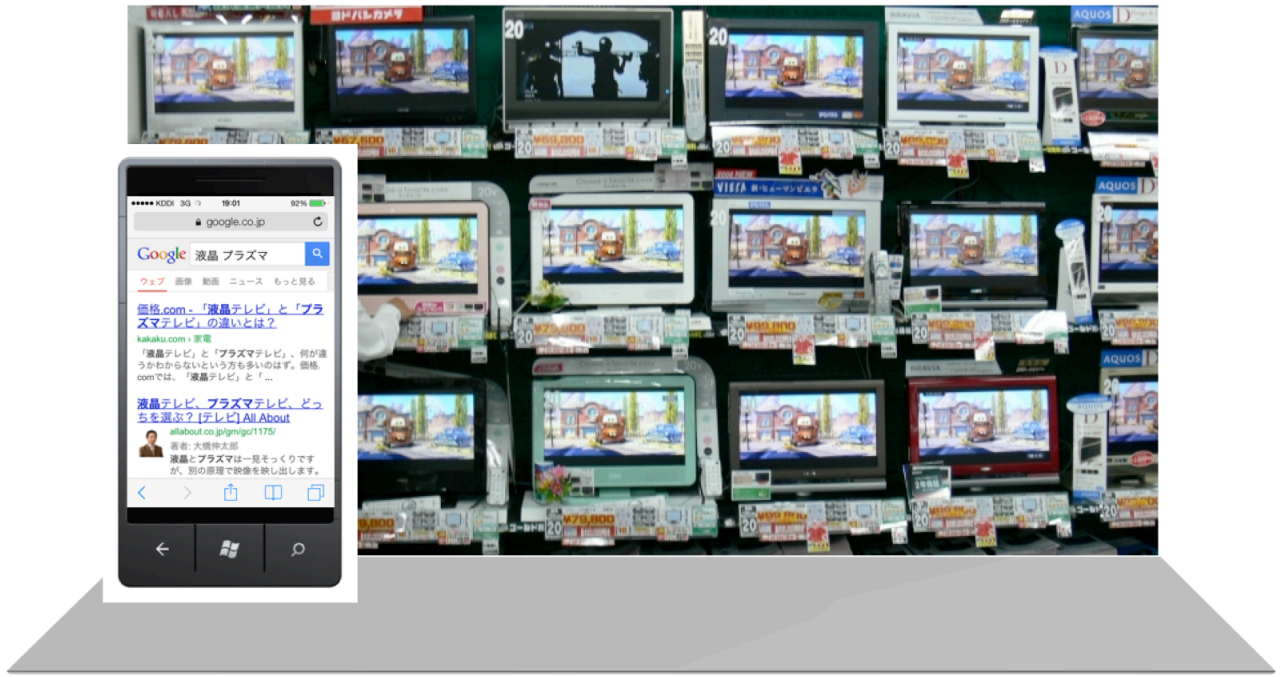
Sakai. 2021. *Graded Relevance*, § 1.3.6 (in the NTCIR book).

1CLICK

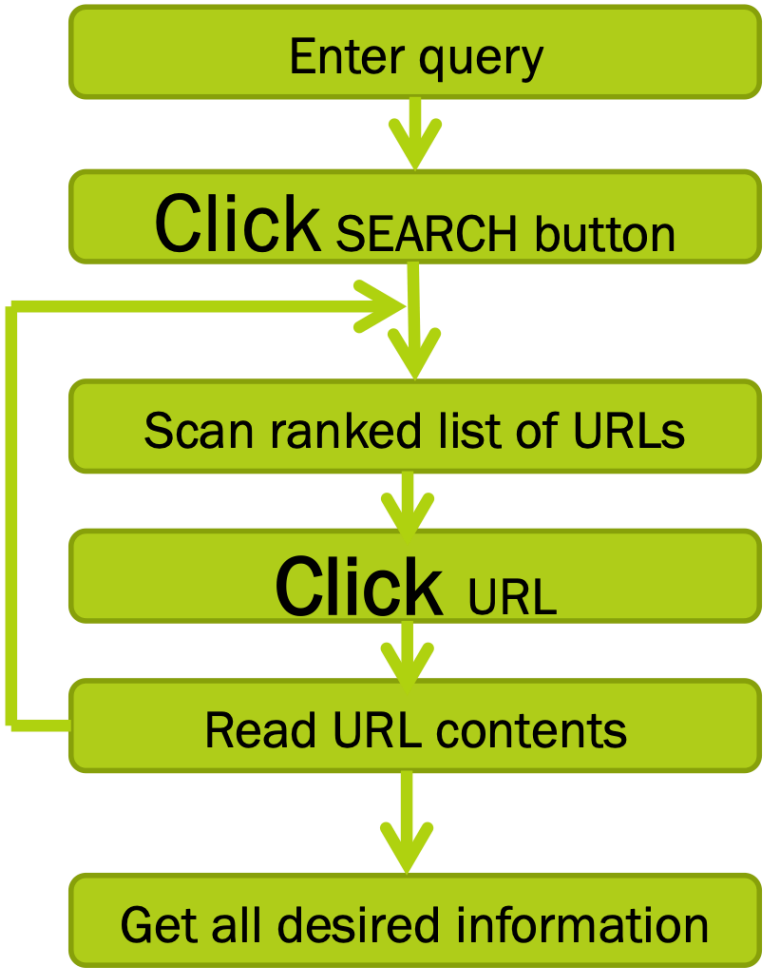
1CLICK (NTCIR-9, NTCIR-10)

Suppose that ...

- Finding answers for a question
  “what’s the difference between PDP and LCD?”



In the "ten-blue-link" paradigm



More than one clicks needed before being satisfied

One Click Access

Enter query

Click SEARCH button

Get all desired information

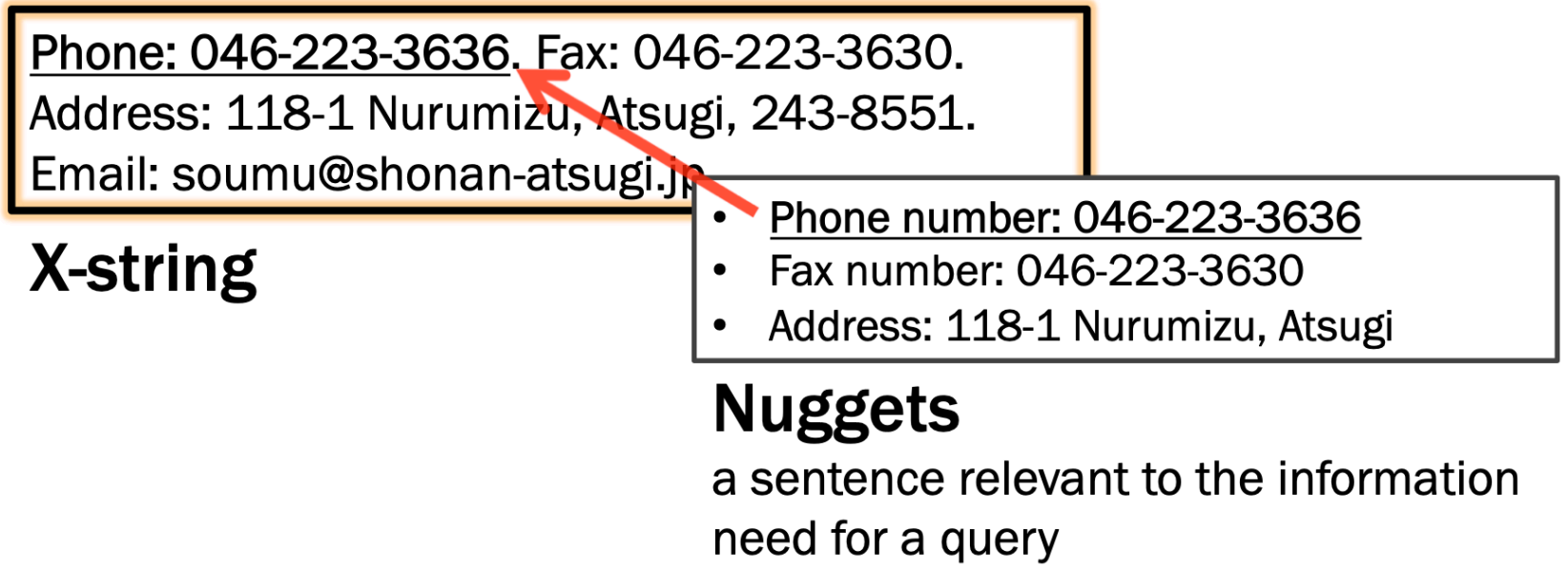
The system outputs *X-string*

Task:
Given a search query, return a single textual output (*X-string*)

Go beyond the "ten-blue-link" paradigm, and tackle *information* retrieval rather than document retrieval

Evaluation of 1CLICK Systems

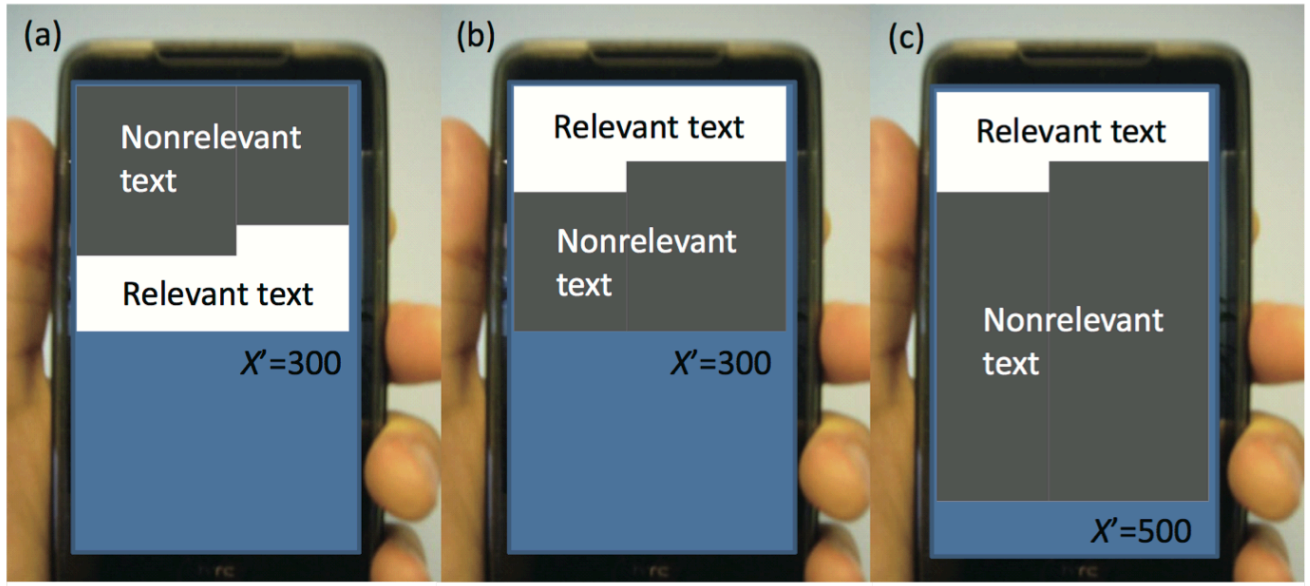
- Manual/automatic matching between the **X-string** and **nuggets**



- Systems are required to present

more important information earlier

Evaluation Metrics for 1CLICK



- Unlike nugget precision/recall, S-measure (position-aware weighted recall) says (a)<(b).
  T-measure (a kind of precision) says (b)>(c).
  S# (official evaluation metric) combines S and T

1CLICK — direct, immediate information access

I believe that this task, which started in 2011, is precursor to the direct answer function of current search engines.

Instead of a ranked list of URLs, return a **single text** that answers the query directly.

> *Given a query and an output window of size **X** characters, return an X-string that:*
>
> ***1.** presents important nuggets first, and*
> ***2.** minimises the amount of text the user has to read.*

- **NTCIR-9**: 60 Japanese queries, 4 query types (CELEBRITY / LOCATION / DEFINITION / QA); 2,839 nuggets
- **NTCIR-10**: 8 query types, English + Japanese; 3927 nuggets for 100 Japanese jqueries

Sakai, Kato, Song. 2011. "Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access." CIKM 2011.

1CLICK — S-measure

S-measure: discount the value of a nugget by *how far the user has to read* to find it. (or a position-aware version of BLEU or ROUGE)

$$S = \frac{1}{\mathcal{N}} \sum_{u \in M} w(u) \cdot \max\left(0, 1 - \frac{\text{offset}(u)}{L}\right)$$

- M: a set of nuggets for a query (useful information pieces)
- $w(u)$: weight of a nugget u
- offset(u): position of u
- L: a **patience parameter** (e.g. 500 chars ≈ one minute of reading at 500 chars/min for Japanese)
- **Findings**: simple Wikipedia snippet baselines were surprisingly hard to beat for celebrity queries; participants were better for facility, definition, and QA queries

Sakai, Kato, Song 2011 (CIKM); Kato et al., Overview of NTCIR-10 1CLICK-2, 2013.

OpenLiveQ

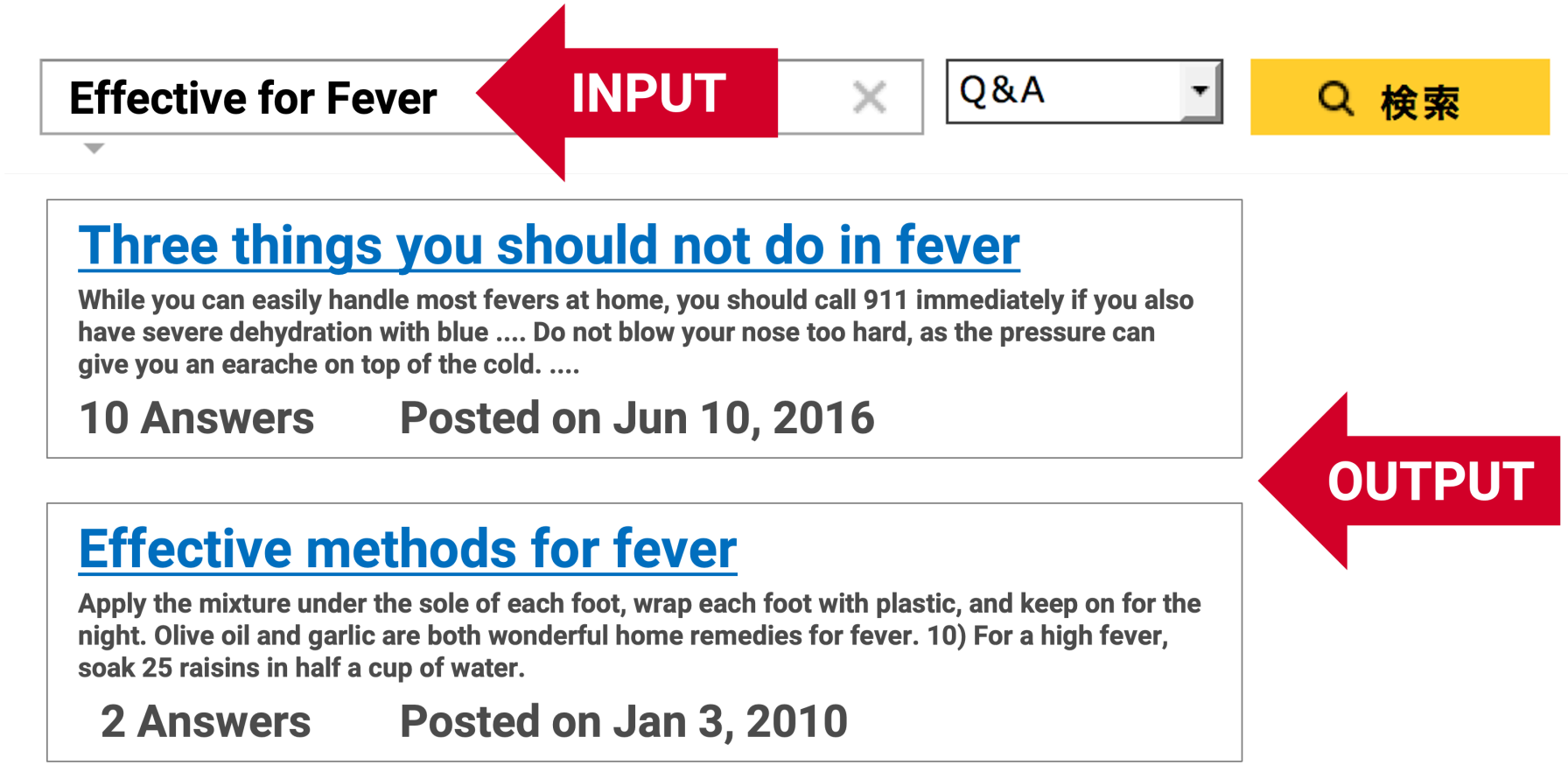
Performance evaluated by REAL users

Improve the REAL performance of question retrieval systems in a production environment

Yahoo! Chiebukuro
(a CQA service of Yahoo! Japan)

OpenLiveQ — Task

- **Given a query, return a ranked list of questions**
  - **Must satisfy many REAL users in Yahoo! Chiebukuro (a CQA service)**



OpenLiveQ — Online Evaluation

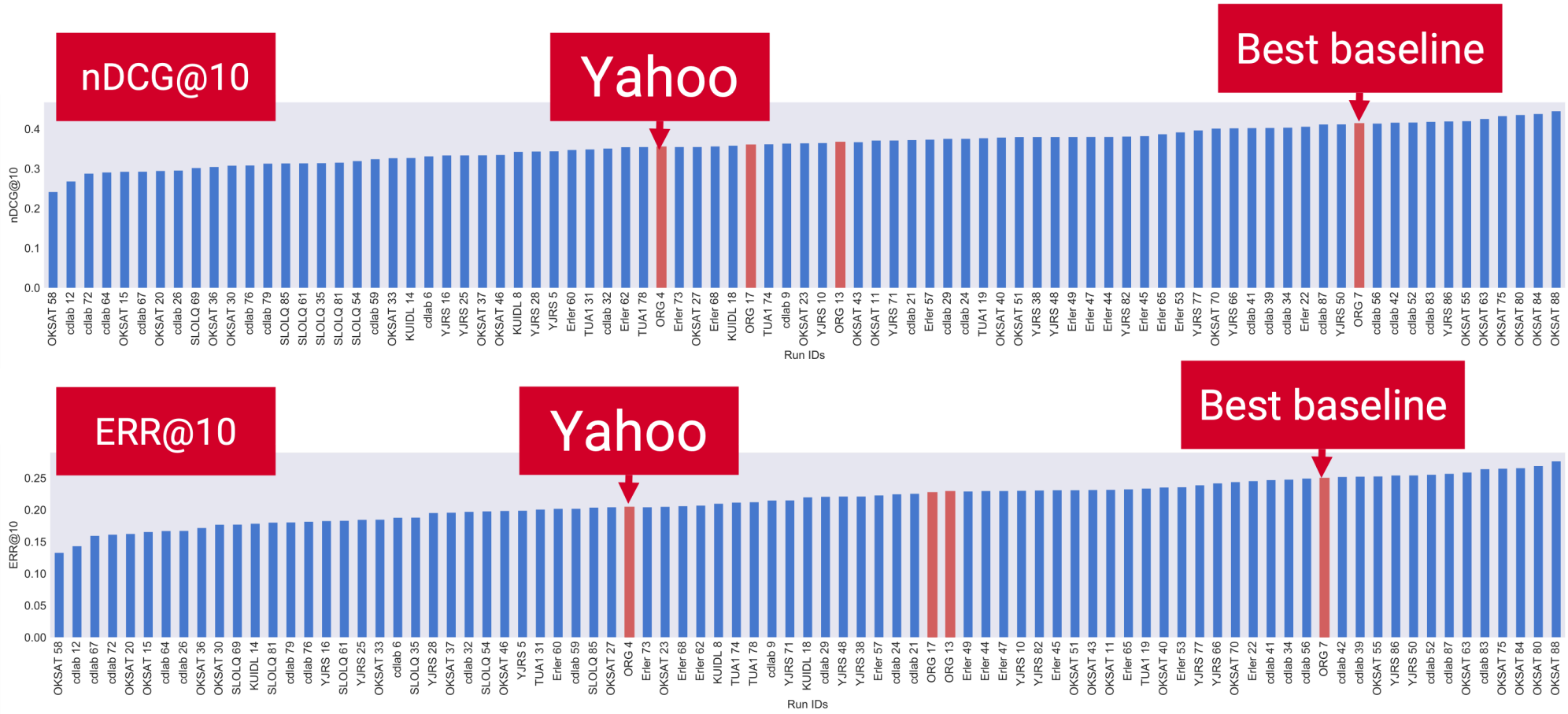
Ranked lists of questions from participants' systems are INTERLEAVED, presented to real users, and evaluated by their clicks

OpenLiveQ — Offline and Online Evaluation

- Participants were given training data (queries + questions + clicks) and test queries (no clicks)
  - 1,000 queries sampled from Yahoo! Chiebukuro (Japanese Yahoo! Answers)
  - ~1M questions
  - 3-month click data
- Participants were asked to submit runs for the test queries, which were then evaluated both offline (using manual relevance labels) and online (against real users)
- **Online evaluation**
  - Multileaved comparison (Optimized Multileaving[1]) was used
    - We found it the best in our experiments[2]
- We intended to see a strong correlation between offline and online evaluation

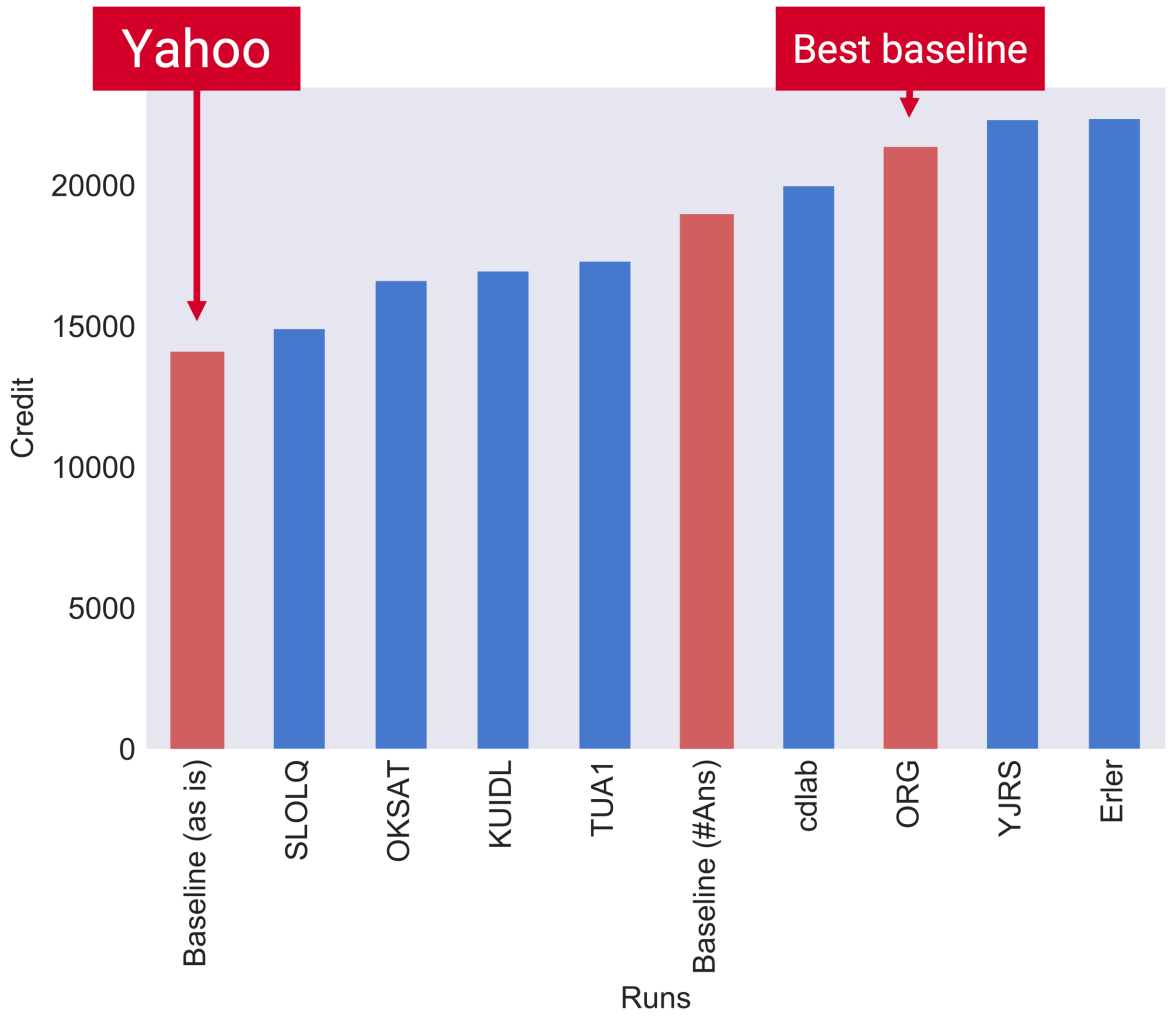
1. Schuth et al.: Multileaved comparisons for fast online evaluation, CIKM 2014. 2. Manabe et al. A Comparative Live Evaluation of Multileaving Methods on a Commercial cQA Search, SIGIR 2017

OpenLiveQ — Offline Evaluation



Offline evaluation showed that the top 3 teams (OAKSAT, YJRS, cdlab) were significantly better than the Yahoo! production baseline.

OpenLiveQ — Online Evaluation



Online evaluation showed a slightly different results: YJRS and Erler performed well but the others did not.

OpenLiveQ — Offline vs Online Evaluation

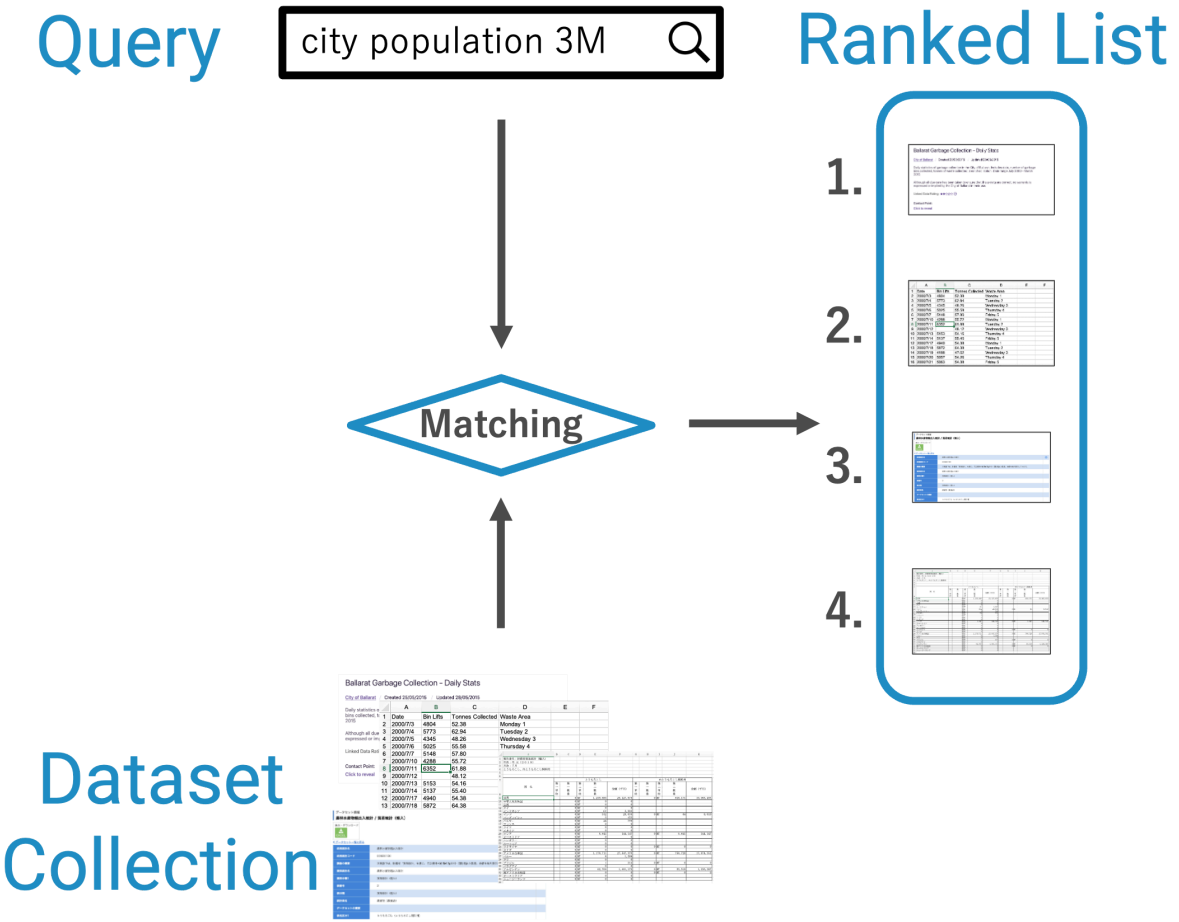
- **The offline and online evaluation results were not strongly correlated**
  - This suggests that offline evaluation may not always predict online performance: IR evaluation needs both offline and online methods to get a complete picture of system performance.
- **Advantages of offline evaluation:**
  - Controlled environment, reproducibility, and detailed analysis
  - Allows for rapid iteration and comparison of different approaches without the variability of real user interactions
- **Advantages of online evaluation:**
  - Real user interactions, capturing user satisfaction and engagement, and evaluating system performance in a live setting
  - Many services want to maximize user satisfaction, which is best measured through online evaluation

Data Search

Data Search — the first IR campaign for datasets

Setting: ad-hoc retrieval over **statistical / governmental datasets**, not Web documents.

Information need	Query
I am looking for evidences of domestic self-sufficiency rate of salt	domestic self salt rate
Are there many people who can't drive large trailers?	people can't drive large trailers
How many people have a second house?	many people second house



- **English dataset collection**
  - Data.gov https://www.data.gov/
- **Japanese dataset collection**
  - e-Stat https://www.e-stat.go.jp/

Kato, Ohshima, Liu, Chen. 2020. Overview of the NTCIR-15 Data Search Task.

Data Search — the first dataset search test collection

English	**Documents (or *datasets*)**	46,615
	Training queries	192
	Test queries	58
	Relevance judgments for training queries	8,248
	Relevance judgments for test queries	6,550

Japanese	**Documents (or *datasets*)**	1,338,402
	Training queries	192
	Test queries	72
	Relevance judgments for training queries	7,754
	Relevance judgments for test queries	4,035

DATA.GOV

DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONTACT

Data.gov users! We welcome your suggestions for improving Data.gov and federal open data.

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

For information regarding the Coronavirus/COVID-19, please visit Coronavirus.gov.

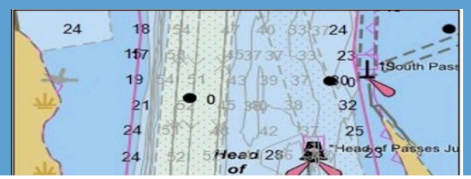
GET STARTED

SEARCH OVER 335,221 DATASETS

Federal Student Loan Program Data

HIGHLIGHTS

Rivers of Data – Inland Electronic Navigation Charts



Nautical charts provide critical information to mariners in support of safe navigation. Historically these charts have been printed and distributed on paper, but modern communications systems allow for electronic charts that are able to be updated as ne information becomes available. The National Oceanic and Atmospheric Administrati (NOAA) Office of Coast Survey produces charts for coastal and Great Lakes areas, an

255,226 datasets found

Lottery Powerball Winning Numbers: Beginning 2010 1892 recent views

State of New York — Go to http://on.ny.gov/1GpWiHD on the New York Lottery website for past Powerball results and payouts.

CSV RDF JSON XML

State

FDIC Failed Bank List 1449 recent views

Federal Deposit Insurance Corporation — The FDIC is often appointed as receiver for failed banks. This list includes banks which have failed since October 1, 2000.

CSV HTML

Federal

Electric Vehicle Population Data 1253 recent views

State of Washington — This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department...

CSV RDF JSON XML

State

National Student Loan Data System 1135 recent views

Department of Education — The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher...

XLS XLS XLS XLS XLS XLS 11 more in dataset

Federal

Alzheimer's Disease and Healthy Aging Data 967 recent views

U.S. Department of Health & Human Services — 2015-2020. This data set contains data from BRFSS.

CSV RDF JSON XML

Federal

Examples of data

Ballarat Garbage Collection - Daily Stats

City of Ballarat / Created 25/05/2015 / Updated 28/05/2015

Daily statistics of garbage collection in the City of Ballarat. Includes date, number of garbage bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015

Although all due care has been taken to ensure that these data are correct, no warranty is expressed or implied by the City of Ballarat in their use.

Linked Data Rating: ★★☆☆☆

Contact Point:

Click to reveal

	A	B	C	D	E	F
1	Date	Bin Lifts	Tonnes Collected	Waste Area		
2	2000/7/3	4804	52.38	Monday 1		
3	2000/7/4	5773	62.94	Tuesday 2		
4	2000/7/5	4345	48.26	Wednesday 3		
5	2000/7/6	5025	55.58	Thursday 4		
6	2000/7/7	5148	57.80	Friday 5		
7	2000/7/10	4288	55.72	Monday 1		
8	2000/7/11	6352	61.88	Tuesday 2		
9	2000/7/12		48.12	Wednesday 3		
10	2000/7/13	5153	54.16	Thursday 4		
11	2000/7/14	5137	55.40	Friday 5		
12	2000/7/17	4940	54.38	Monday 1		
13	2000/7/18	5872	64.38	Tuesday 2		
14	2000/7/19	4188	47.02	Wednesday 3		
15	2000/7/20	5057	54.26	Thursday 4		
16	2000/7/21	5063	54.38	Friday 5		

データセット情報

農林水産物輸出入統計 / 貿易統計（輸入）

表示・ダウンロード

EXCEL

データセット一覧に戻る

政府統計名	農林水産物輸出入統計
政府統計コード	00500100
調査の概要	本統計では、財務省「貿易統計」を基に、主な農林水産物の品目別・国別輸出入数量、金額を毎月提供しています。
提供統計名	農林水産物輸出入統計
提供分類1	貿易統計（輸入）
表番号	2
表分類	貿易統計（輸入）
統計表名	農産物（農産品）
データセットの概要	
表名区分1	とうもろこし（とうもろこし飼料用）

	A	B	C	D	E	F	G	H	I	J	K
1	報告書名：財務省貿易統計（輸入）										
2	年次：令.元（２０１９）										
3	月次：７月										
4	とうもろこし、内とうもろこし飼料用										
5											
6											
7		とうもろこし					内とうもろこし飼料用				
8	国　名	第一単位	第一数量	第二単位	第二数量	金額（千円）	第一単位	第一数量	第二単位	第二数量	金額（千円）
9	世界		0	MT	1,239,883	29,127,675		0	MT	818,171	19,063,495
10	中華人民共和国		0	MT	0	0					
11	台湾		0	MT	0	0					
12	タイ		0	MT	0	0					
13	インドネシア		0	MT	64	3,003					
14	インド		0	MT	532	28,835		0	MT	84	8,610
15	バングラデシュ		0	MT	2	452					
16	ベルギー		0	MT	24	795					
17	フランス		0	MT	0	0					
18	ドイツ		0	MT	0	0					
19	イタリア		0	MT	0	0					
20	ロシア		0	MT	5,841	144,147		0	MT	5,841	144,147
21	オーストリア		0	MT	0	0					
22	ハンガリー		0	MT	0	0					
23	ルーマニア		0	MT	0	0					
24	ウクライナ		0	MT	0	0		0	MT	0	0
25	カナダ		0	MT	0	0					
26	アメリカ合衆国		0	MT	1,170,711	27,547,370		0	MT	756,728	17,674,351
27	ペルー		0	MT	8	1,589					
28	チリ		0	MT	0	0					
29	ブラジル		0	MT	1	312		0	MT	0	0
30	パラグアイ		0	MT	0	0					
31	アルゼンチン		0	MT	62,700	1,401,172		0	MT	55,518	1,236,387
32	南アフリカ共和国		0	MT	0	0		0	MT	0	0
33	オーストラリア		0	MT	0	0					
34	ニュージーランド		0	MT	0	0					
35											

Why is it difficult?

Required to match text with data (tabular data)

What we are given

domestic self salt rate

What we want to retrieve

Metadata

Ballarat Garbage Collection - Daily Stats

City of Ballarat / Created 25/05/2015 / Updated 28/05/2015

Daily statistics of garbage collection in the City of Ballarat. Includes date, number of garbage bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015

Although all due care has been taken to ensure that these data are correct, no warranty is expressed or implied by the City of Ballarat in their use.

Linked Data Rating: ★★☆☆☆

Contact Point:

Click to reveal

Data files

	A	B	C	D	E	F	G	H	I
1	PARID	PROPERTYH	PROPERTYF	PROPERTYA	PROPERTYA	PROPERTYA	PROPERTYA	PROPERTYU	PROPERTYC
2	1075F00108	4720			HIGHPOINT	DR			GIBSONIA
3	0011A00237	0			LOMBARD	ST			PITTSBURG
4	0011J00047	1903			FORBES	AVE			PITTSBURG
5	0113B00029	479			ROOSEVELT	AVE			PITTSBURG
6	0119S00024	5418			CARNEGIE	ST			PITTSBURG
7	0120D00075	6311			BUTLER	ST			PITTSBURG
8	0125M0014	0			KELLY	ST			PITTSBURG
9	0133B00214	419			UNION	ST			HOMESTEA
10	1361R00253	127			MICHAEL	RD			CHESWICK
11	0139P00233	875			KILLARNEY	DR			PITTSBURG
12	0140N00243	393			JONQUIL	PL			PITTSBURG
13	0015K00072	267			BOGGSTON	AVE			PITTSBURG
14	1510N00180	0			MOUNTAIN	RD			GIBSONIA
15	1518K00012	1734			JUNIATA	ST			NATRONA H
16	1518K00020	1738			PLEASANTV	RD			NATRONA H
17	1678C00205	1261			DAVIS	ST			NATRONA H
18	1678S00159	1013			BEACON	ST			NATRONA H
19	1678G00353	1261			DAVIS	ST			NATRONA H
20	0173B00191	7130			WILTSIE	ST			PITTSBURG

As the metadata often lack details,
not all keywords in the query may not match the metadata

Data Search — some impacts of NTCIR Data Search

- **Followed by other data search test collections**
  - Lin et al. "ACORDAR: a test collection for ad hoc content-based (RDF) dataset retrieval." SIGIR 2022.
  - Chen et al. "ACORDAR 2.0: A test collection for ad hoc dataset retrieval with densely pooled datasets and question-style queries." SIGIR 2024.
  - Kolyada et al. "A Test Collection for Dataset Retrieval", ECIR 2025.
- **Included in a larger test collection**
  - Sun et al. "MAIR: A massive benchmark for evaluating instructed retrieval." EMNLP 2024.
- **Re-used for developing a new test collection**
  - Shi et al. "DSEBench: A Test Collection for Explainable Dataset Search with Examples"

4. Lessons Learned

Cross-task lessons from these tasks

1. **Good metrics outlive any single submission.** S-measure (1CLICK) → U-measure → M-measure (MobileClick); D♯-measure → still in use.
2. **Reusable test collections are the durable contribution.** The deep value of running NTCIR-X is the data + baselines + judgments left behind for the next decade.
3. **Strong baselines force ambition.** 1CLICK's Wikipedia baseline and OpenLiveQ's Yahoo production baseline both reframed what "winning" meant.
4. **Online + offline triangulation reveals what either alone cannot.** OpenLiveQ's offline-vs-online divergence was a result *and* a methodological contribution.

Evaluation methods born or matured at NTCIR

Measure	First used	Idea
nDCG / Q-measure	NTCIR-3 Web (2003), CLIR-6 (2007)	Graded-relevance ranked retrieval
D#-nDCG, I-rec, DIN-nDCG, P+Q	INTENT (2011–13)	Diversification with intent recall
S-measure, T-measure, S#	1CLICK-1/2 (2011–13)	Position-aware text evaluation
U-measure	Sakai & Dou, SIGIR 2013	Generalised S to ranked lists / sessions
M-measure, H-measure	MobileClick / IMine	Multi-layer / hierarchical / vertical eval

Compiled from Sakai 2021 Tables 1.1, 1.2; Kato 2021, Ch. 11.

5. NTCIR vs TREC vs CLEF

— A General Chair Perspective —

Evaluation Campaigns in the IR Community (not exhaustive)

- **NTCIR** — since **1997**
- Hosted by **NII, Tokyo**
- Anchor: **Asian languages** (J, C, K)
- **Sesquiannual** (18-months cycle)
- **Tasks:** heavy on novel tasks, multilingual, domain-specific NLP tasks

- **TREC** — since **1992**
- Hosted by **NIST, USA**
- Anchor: **English**
- **Annual**
- **Tasks:** heavy on IR. Strongly connected to SIGIR topics

- **CLEF** — since **2000**
- Hosted by European universities
- Anchor: **European languages**
- **Annual**
- **Tasks:** initially focused on cross-lingual IR, later extended to multimedia and multimodal

They are not competitors. They are **geographically and linguistically complementary**.

How NTCIR/TREC/CLEF themes are differentiated

- **Language / region split** — NTCIR: East Asian (NII, Tokyo); TREC: English Web (NIST, USA); CLEF: European multilingual (CLEF Initiative + rotating European host).
- **NTCIR — pilot-core paradigm** — Patent (NTCIR-3), Math (NTCIR-10), Lifelog (NTCIR-12), 1CLICK / MobileClick (NTCIR-9–12), Data Search (NTCIR-15).
- **TREC — large-scale English benchmarks** — Deep Learning, Web, Total Recall, MS MARCO, RAG via NIST infrastructure.
- **CLEF — social & applied IR** — CheckThat!, EXIST, Touché, JOKER, LongEval, LifeCLEF, eHealth — federated lab proposals might enable societally oriented tasks.

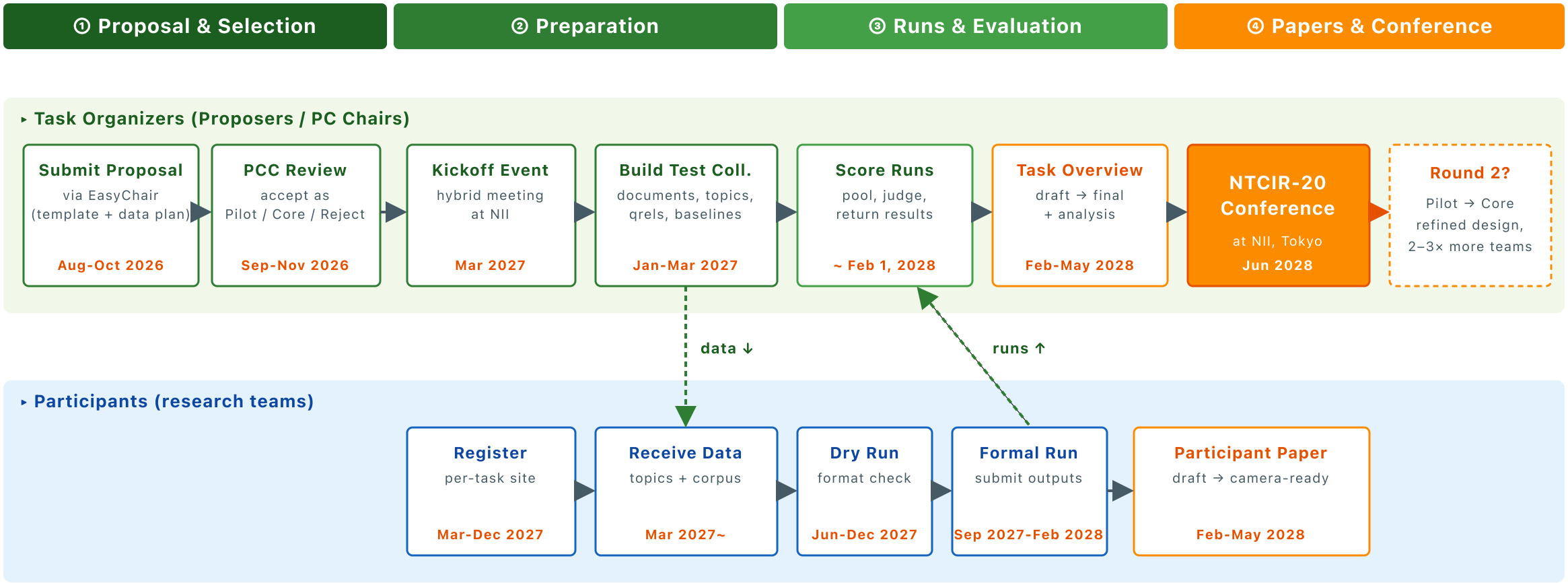
Overlap & Collaboration

- **Different language collections** — same task, different test beds (CLEF + NTCIR + TREC have run parallel CLIR for years)
- **Reproducibility tracks bridge the two — NTCIR-14 CENTRE** (CLEF-NTCIR-TREC REproducibility) explicitly required participants to reproduce NTCIR-13 WWW and TREC 2013 Web Track runs
- **Invited talks at each other's conferences** e.g. NTCIR conferences regularlly featured invited talks by Ian Soboroff (TREC) and Nicola Ferro (CLEF).

6. Running an NTCIR Task

— A PC Chair Perspective —

How a task gets onto NTCIR (e.g., NTCIR-20 (TENTATIVE))



Two rounds of proposals per cycle

- NTCIR-19 (July 2025 → Dec 2026, ~18-month cycle): 14 tasks accepted = 11 initial + 3 additional.
- Proposals are reviewed for clear motivation, clear challenges for participants, workable data plan, defensible evaluation, and likely ≥5 teams.

What makes a strong NTCIR task

Five pillars of a strong NTCIR task proposal

① Motivation

Why does it matter?

What real-world problem or open research question does the task address?

Test question:

Can you state the impact in one sentence?

⚠ Red flag

Solving a problem nobody actually has

② Challenges

Room for improvement?

Strong baselines should NOT solve it. Lots of headroom for participants.

Test question:

Does a state-of-the-art baseline leave a clear gap?

⚠ Red flag

Already solved by off-the-shelf LLM

③ Evaluation

Measurable success?

Is there a clear metric? Can annotaters give labels with high reproducibility?

Test question:

Does the metric correlate with goodness in real-world applications?

⚠ Red flag

Unmeasurable success criterion ("looks good")

④ Data Plan

Survives reality?

Licences cleared. Data annotation capacity.

Test question:

Do you already hold a working corpus + 50 topics?

⚠ Red flag

No working examples

⑤ Community

Will anyone come?

Do you have 5+ teams who would actually submit runs?

Test question:

Can you name 5 PIs who would actually submit runs?

⚠ Red flag

No identifiable community ("we'll advertise widely")

PCC perspectives

✓ All five pillars defensible → **accept as Core**

~ 3–4 pillars clear, 1–2 immature → **accept as Pilot** (smaller, scoped, iterate to round 2)

× Weak motivation · solved by baseline · unmeasurable success · unclear licence · no community → **reject**

Advice for prospective organisers

1. **Build a dataset before you propose.** Even a small prototype makes the proposal concrete and surfaces problems early.
2. **Design with participants in mind.** Overly complex tasks scare teams away — and a task no one joins is a bad outcome for organisers too. Lower the bar with baseline code, starter kits, and a clear tutorial.
3. **Plan for multiple rounds from day one.** NTCIR strongly encourages running a task for **at least two rounds**. Scope round 1 knowing that round 2 is where the task matures.
4. **It does not have to be an IR task.** NTCIR has hosted plenty of NLP tasks — sentence classification, NER, argument mining, and more. Applied NLP almost always reduces to a form of information access.
5. **Involve many people, across many places.** Strong tasks are cross-country and multilingual — in both the organising team and the participants.

Take-home

Take-home messages

1. **Tasks and test collections often outlast systems.** Winning runs fade, but the data, metrics, and baselines left behind tend to be reused for years.
2. **Be open to trying new tasks.** Patent, Math, Lifelog, 1CLICK, Dataset Search — NTCIR has often hosted tasks that later grew into wider research areas.
3. **Metrics and user models deserve careful design,** just as much as the systems they evaluate. Good ones can be reused across tasks.
4. **Multilingual and cross-country collaboration helps,** both in the organising team and among participants.
5. **The community matters.** Tasks come and go, but the people who gather around NTCIR are what keep the field moving forward.