NTCIR-15 Pilot Task
Data Search

Makoto P. Kato (University of Tsukuba)
Hiroaki Ohshima (University of Hyogo)
Ying-Hsang Liu (Australian National University)
Hsin-Liang Chen (Missouri University of Science and Technology)

- **The open data movement is now being accelerated by the expectations for open science and citizen science**
  - Each country strongly encourages the open data movement:
    - Data.gov (United States)
    - Data.gov.uk (United Kingdom)
    - Data.gov.au (Australia)
    - e-Stat (Japan)
- **Besides the governmental portals, there are also thousands of data repositories on the Web**

Demand for a better data search engine

(e.g. Google Dataset Search)

Examples of Information Needs for Data Search

How much is the 20-year-old population now?

I am looking for statistics on the smartphone usage in these five years.

Is the population of Japanese in rural areas decreasing?

How many people are working for each job type?

Please let me know the population of Hachioji at daytime and night.

How much cost is required to farm a pig?

Are there any documents on the income of a family with a double income?

How many lives are born and lost per second in Japan?

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											

- **Query understanding for data search**
  - Queries for data search include more geographical, temporal, and numerical keywords than those for Web search (Kacprzak+ 2017)
  - The goal of data search can be diverse, e.g. time series analysis and summarization (Koesten+ 2017)
- **Data understanding for data search**
  - Metadata are not always sufficiently informative
  - Data in Excel, CSV, XML, and PDF formats is potentially used with metadata to enrich the index for data search,
    while interpreting data on the Web is a still challenging problem
- **Retrieval models for data search**
  - Data contains a lot of entities such as locations or products, temporal expressions, and numerical expressions
  - Numerical expressions might require a new model for better rankings

Ad-hoc retrieval for statistical data

- **Subtasks**
  - English and Japanese
- **Input**
  - 100 queries for each of the subtasks
- **Document collection**
  - e-Stats for Japanese (~1.5M)
  - Data.gov for English (~0.2M)
- **Output**
  - Ranked list of data for each query
- **Resources**
  - Additionally, ~100 queries may be provided for training
    - + relevance judgements for baseline rankers

- **Almost the same as that for ordinary ad-hoc retrieval tasks**

- **Relevance assessment**
  - Three assessors will be hired
  - A three point scale: not relevant, partially relevant, highly relevant
- **Evaluation metrics**
  - nDCG
  - ERR
  - Q-measure

Nov 30, 2019	Data collection/training queries release
Mar 30, 2020	Registration due
Jun 30, 2020	Run submission due
Jul - Aug, 2020	Relevance judgement
Aug 31, 2020	Evaluation results release

- **Please join us if you are interested in data search**
  - http://ntcir.datasearch.jp