ROIS-DS-JOINT
029RP2018

医療アクセス制限研究の属性共起分析による類似オープンデータ順位付けとデータサイエンス応用

神沼英里[1)], 田中博[1)], 山本泰智[2)]

1) 東京医科歯科大学 医療データ科学推進室
2) ライフサイエンス統合データベースセンター

Medical and Drug Discovery Data Science Consortium

https://md-dsc.com/

【AIM : Training participants into ‘data science competent’】

- Familiar with data science of big data analysis and artificial intelligence
- Holding skills to design data science solutions in medical and drug discovery domains.

OBJECTIVES : Constructing a consortium to promote data science education collaborating academic research institutes and companies

Participating Research Institutes

*Tohoku University
*Tokyo University of Science
*Keio University
*National Center for Global Health and Medicine
*National Center for Neurology and Psychiatry
* National Institute of Advanced Industrial Science and Technology
* Japanese Foundation of Cancer Research
*Japan Biological Information Consortium
* Seed Planning Inc.

Representative organization
Tokyo Medical and Dental University

Collaborative Companies
*Mitsubishi Tanabe Pharma
*DENKA Company Ltd.
*Chugai Pharmaceutical Co., Ltd.
*CAC Croit Corporation.
*CAC Corporation.
*Japan Tobacco Inc.
*Asahi Kasei Pharma Corporation
*Otsuka Pharmaceutical Co., Ltd.
*Ajinomoto Co., Inc.
*Takeda Pharmaceutical Co., Ltd.
*NEC Corporation
*Taiho Pharmaceutical Co., Ltd.
*Fujitsu Ltd.
*Kyowa Hakko Kirin Co., Ltd.
*First Ascent Co., Ltd.
*Teijin Pharma Ltd.
*Kimura Information Technology Co., Ltd.
*Astellas Pharma Inc.
*Pfizer Japan Inc
*Ono Pharmaceutical Co., Ltd.
*Daiichi Sankyo Co., Ltd.
*Kaken Pharmaceutical Co., Ltd.

On the Job training programs for business person participants

National Center for Global Health and Medicine
国立国際医療研究センター

J-DREAMS(診療録直結型全国糖尿病データベース事業)を始めとした電気カルテ情報を活用した医療用ビッグデータの構築・管理・運用

Diabetes EMR database

National Center for Neurology and Psychiatry
国立精神・神経医療研究センター

①脊髄小脳変性症、筋ジストロフィー、プリオン病などの疾患データベースを対象とした統計解析。
②疾患データベース、MRI画像データ、脳波データ、髄液データなどを用いた機械学習

Analyzing Heterogeneous Medical databases

Japanese Foundation for Cancer Research
がん研究会

病理部におけるデータのデータベース作成と解析

Data Analysis for Diagnostic Pathological Images

東京医科歯科大学 Tokyo Medical and Dental University

Deep Learning等によるAI創薬プログラムを使用した計算創薬演習

New programs (2018)

Tohoku medical megabank organization (Miyagi)	GWAS using original Biobank datasets
Keio university, Tsuruoka campus (Yamagata)	Metabolomics

提案手法：医療アクセス制限研究の類似オープンデータ順位付け

医療研究分野ではオープンデータの情報が少なく、
初心者が機械学習モデル構築を簡単に試行できない

＜従来のデータサイエンス応用の流れ＞

電子カルテ
データセット

医療アクセス制限研究の
データは利用不可

（IRB承認後）
通常：数カ月

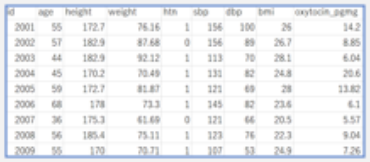
機械学習モデル
の構築

アクセス制限研究の
機械学習モデルへの
応用可能性を検討

＜提案手法＞

オープンデータは、即時利用可能

①オープンデータセットのメタデータから属性キー情報構造化



id	age	height	weight	htn	sbp	dbp	bmi	oxytocin_pgmg
2001	55	172.7	76.16	1	156	100	26	14.2
2002	57	182.9	87.68	0	156	89	26.7	8.85
2003	44	182.9	92.12	1	113	70	28.1	6.04
2004	45	170.2	70.49	1	131	82	24.8	20.6
2005	59	172.7	81.87	1	121	69	28	13.82
2006	68	178	73.3	1	145	82	23.6	6.1
2007	36	175.3	61.69	0	121	66	20.5	5.57
2008	56	185.4	75.11	1	123	76	22.3	9.04
2009	55	170	70.71	1	107	53	24.9	7.26

②属性キーを基にオープンデータ間で属性共起分析、評価尺度からオープンデータを優先順位付け

③高優先順位のオープンデータを用いて属性値予測の機械学習モデル構築

提案手法：
＊属性キー情報から、オープンデータを優先順位付け、高順位データで機械学習モデル構築

解析手順

■作業の流れ

①JDREAM属性キーのMeSH term注釈付け、キュレーション
②オープンデータの属性キーのMeSH term注釈付け、キュレーション

[1] Google Dataset Searchから表データ抽出
（using Google Custom Search API）
[2]表から列ラベル抽出
[3]列ラベルについてMeSH term自動割当
[4]MeSH termキュレーション

③データセット間のterm分析
④クエリtermsに対するデータセットランキング

マニュアルキュレーションの部分を外注

①JDREAMデータベースと属性キーのMeSH termキュレーション

オープンデータを探すクエリとなるアクセス制限研究

QUERY

■事例とするアクセス制限研究

National Center for Global Health and Medicine
（国立国際医療研究センター）

全国糖尿病患者電子カルテ「J-DREAMS」プロジェクト

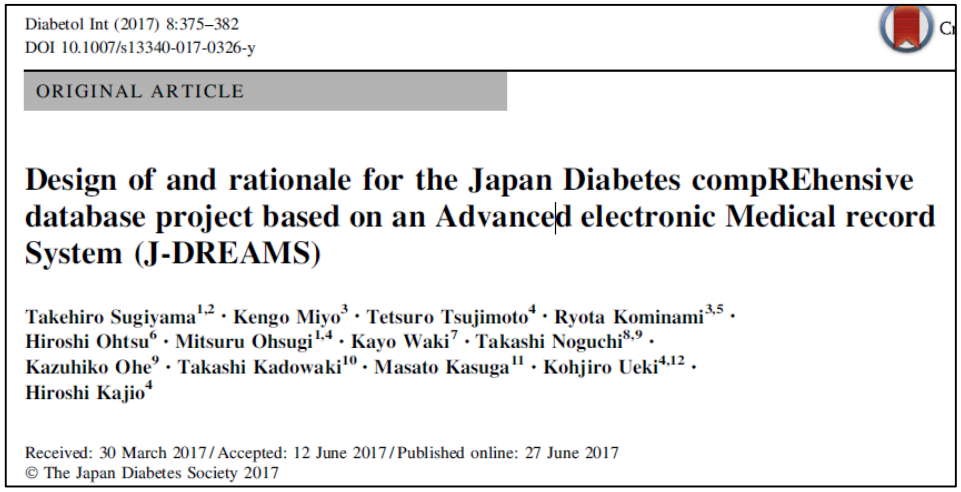
＊2016年2月にスタート
＊ SS-MIX2拡張ストレージに格納
＊匿名化した上で国立国際医療研究センター内のデータベースに登録

＊参加施設の各医師は、所定のテンプレートを使用して通常通りに糖尿病患者の診療記録を入力するだけで済む

J-DREAMSデータベースの属性情報＝43項目

QUERY

Basic information
Year/month of birth
Sex
Hospital code
Laboratory data
Blood samples
Blood cell count
Total protein
Aspartate transaminase
Alanine transaminase
Gamma-glutamyl transpeptidase
Creatine kinase
Total cholesterol
High-density lipoprotein cholesterol
Low-density lipoprotein cholesterol
Triglycerides
Blood urea nitrogen
Creatinine
Potassium
Hemoglobin A1c
Glycoalbumin
1,5-Anhydroglucitol
Blood glucose
Cancer antigen 19-9
Brain natriuretic peptide
Cystatin C
Carcinoembryonic antigen
Thyroid-stimulating hormone
Free triiodothyronine
Free thyroxine
Insulin
C-peptide
Anti-glutamic acid decarboxylase antibodies
Anti-islet antigen 2 antibody
Islet cell cytoplasmic antibody
Zinc transporter 8 antibody
Anti-insulin antibody
Hepatitis B surface antibody
Hepatitis C antibody
Urine samples
Qualitative urinary test
Protein
Albumin
Creatinine
C-peptide
Prescription
All of the patient's prescription information obtained from the participating facility



Diabetol Int (2017) 8:375–382
DOI 10.1007/s13340-017-0326-y

ORIGINAL ARTICLE

Design of and rationale for the Japan Diabetes compREhensive database project based on an Advanced electronic Medical record System (J-DREAMS)

Takehiro Sugiyama[1,2] · Kengo Miyo[3] · Tetsuro Tsujimoto[4] · Ryota Kominami[3,5] · Hiroshi Ohtsu[6] · Mitsuru Ohsugi[1,4] · Kayo Waki[7] · Takashi Noguchi[8,9] · Kazuhiko Ohe[9] · Takashi Kadowaki[10] · Masato Kasuga[11] · Kohjiro Ueki[4,12] · Hiroshi Kajio[4]

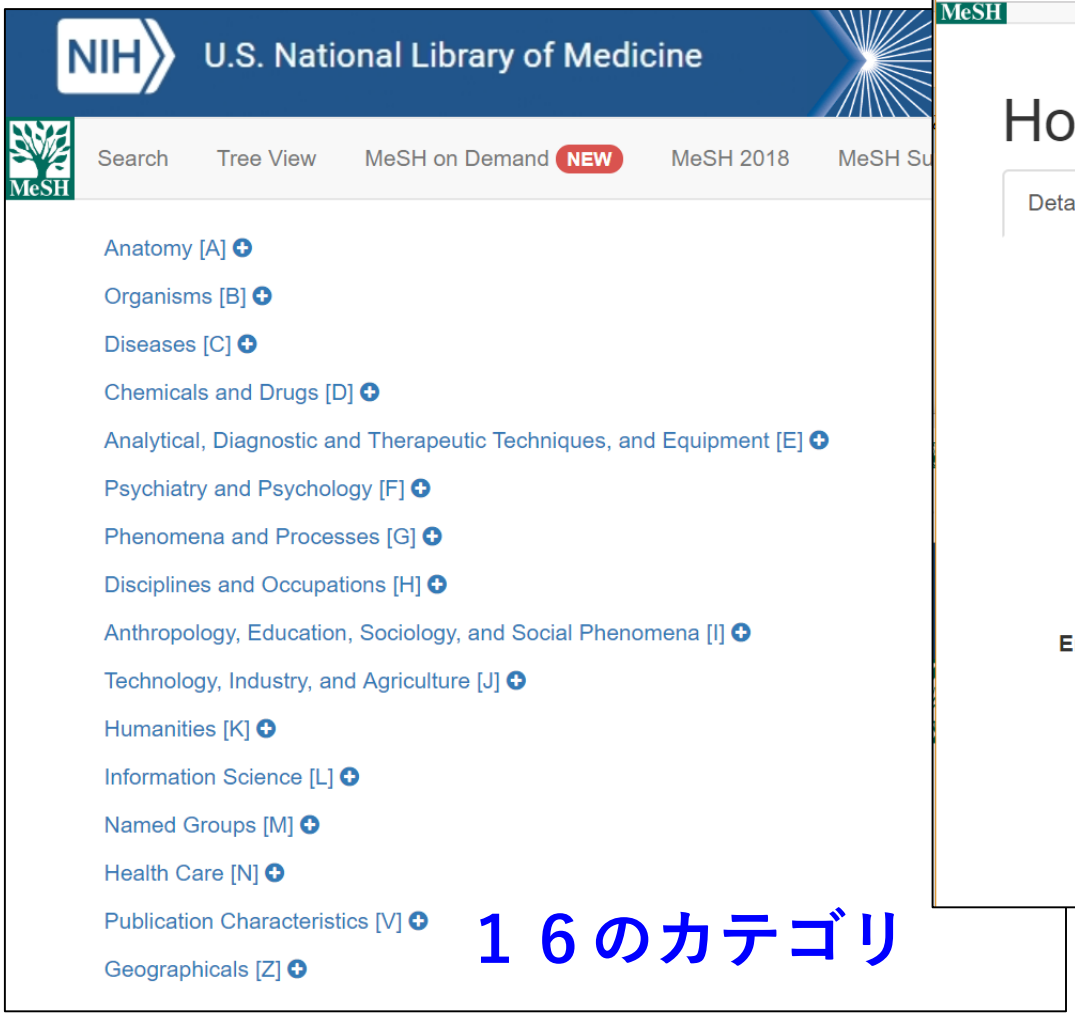
Received: 30 March 2017 / Accepted: 12 June 2017 / Published online: 27 June 2017

The variables collected through J-DREAMS are listed in Table 2 (the basic information, prescription history, and clinical laboratory data stored in the SS-MIX2 standardized storage) and in Supplementary Fig. 1
(the clinical information collected using the SDMT and stored in the SSMIX2 extended storage).

MeSH termとは

■**MeSH**とは

Medical Subject Headings の頭文字であり、米国国立医学図書館 (NLM) が定める生命科学用語集（シソーラス）である。 wikipediaより



J-DREAMS属性情報43項目のMeSH term割り当て

QUERY

■ 専門家による手作業注釈

A	B	C	D	E
CURATED QUERY	QUERY(ORIGINAL VAR)	VAR Category	MeSH TERM(curated)	MESH UNIQID(curated)
Birth	Year/month of birth	Basic	Term Birth	D047929
Sex	Sex	Basic	Sex	D012723
Hospital code	Hospital code	Basic	Hospitals	D006761
Blood	Blood samples	Laboratory	Blood	D001769
Blood cell count	Blood cell count	Laboratory	Blood Cell Count	D001772

MeSH TermとUnique ID

①QUERY
手作業キュレーション

②MeSH TERMのUnique IDを
手作業キュレーション

結果：JDREAMSの属性Keyを、35項目のユニークMeSH Termsに紐づけた

②糖尿病オープンデータを
Google Dataset Search結果から抽出、
属性KeyをMeSH termキュレーション

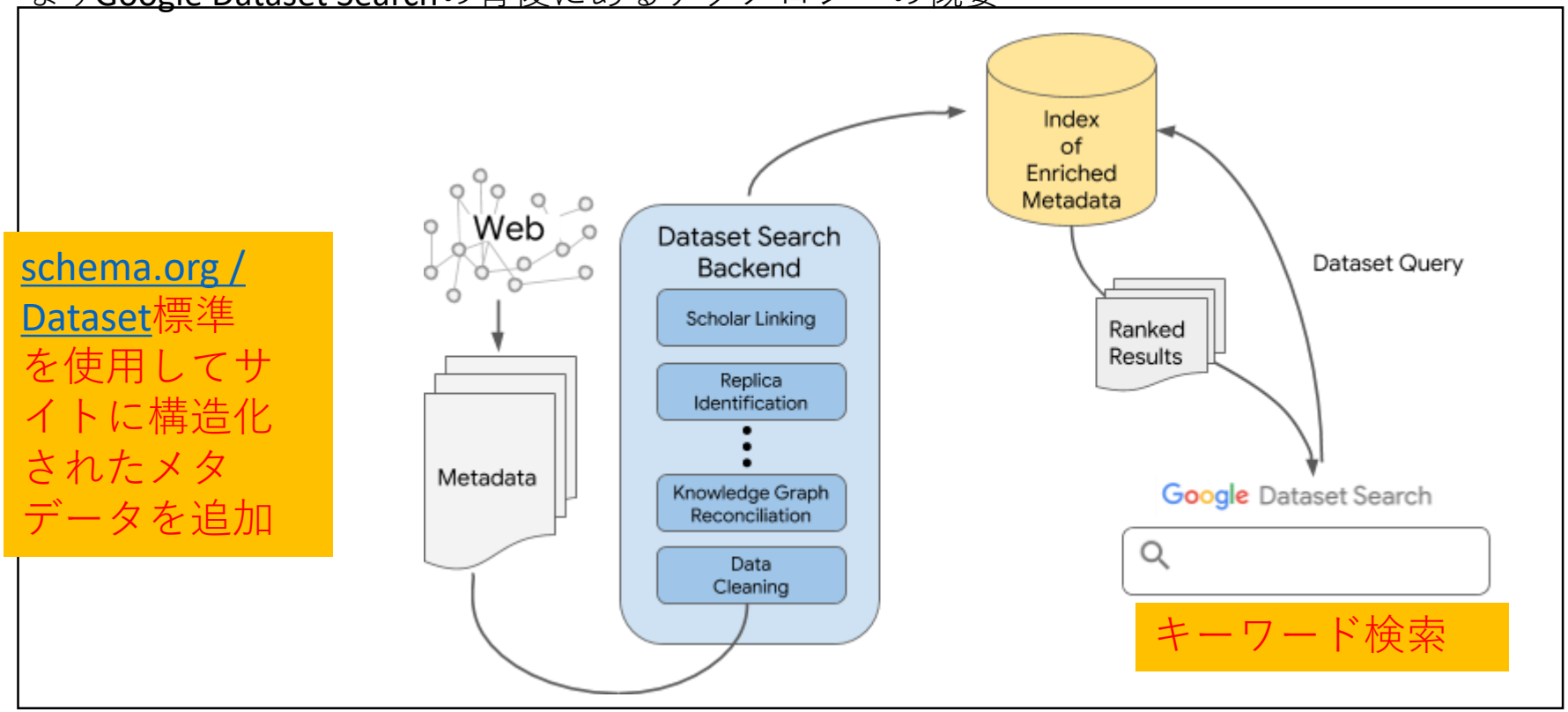
Google Dataset Searchによる訓練データセット検索

■ Google Dataset Search：**オープンデータ**の検索システム

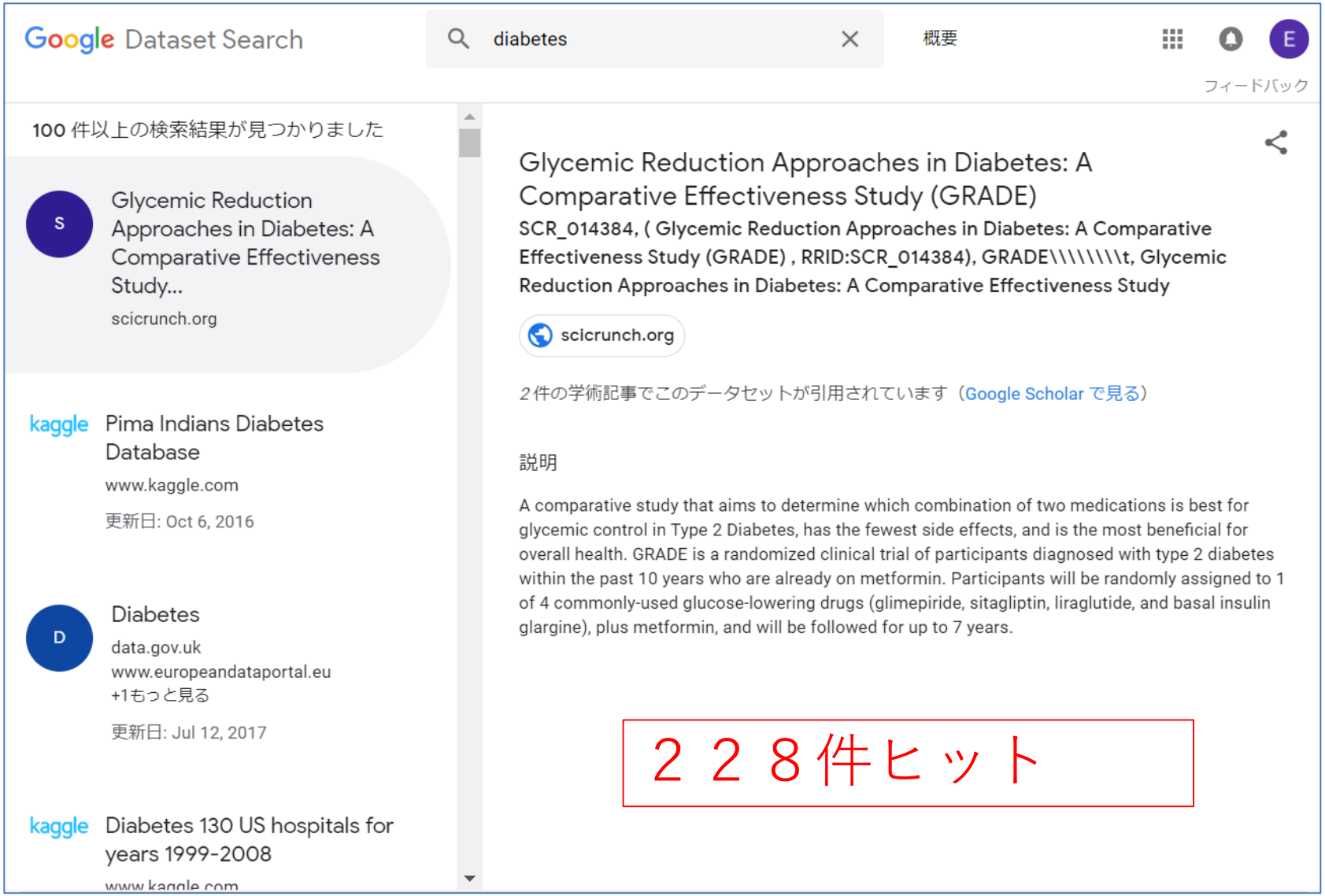
データセットの検索は、二段階で行われます。第一段階では、インターネットクローラーにより、データセットが存在するウェブページがインデックスされます。第二段階では、これらインデックスされたページがランクされます。
データセットが検索されるためには、データ所有者がページに「タグ付け」する必要があります。データセットのタグは、Schma.orgで定めた辞書を用いる必要があります。

https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html
よりGoogle Dataset Searchの背後にあるテクノロジーの概要

schema.org / Dataset標準を使用してサイトに構造化されたメタデータを追加



Google Dataset Searchの「Diabetes(糖尿病)」検索結果



オープンデータの抽出（属性KeyはMeSH Term割当へ）

GDS_HIT_ID	Name	Kaggle	Provider	Type	Number of Instances	Number of Attributes
query	JDREAM			Individuals		43
3	Pima Indians Diabetes Database	Y	UCI Machine Learning	Individuals	768	9
5	Diabetes 130 US hospitals for years 1999-2008	Y	Humberto Brandão	Individuals	100000	55
17	6ヶ月の毎日の糖尿病対策	Y			182	30
23	糖尿病フランクフルト病院	Y			2000	9
	アフリカの成人を対象とした全国サンプ					

①糖尿病のキーワード検索のヒット
228件から17データセットを抽出

②MeSH割り当てへ

③１７オープンデータセットの特徴

属性KeyのMeSH term
手動キュレーション結果

結果①：抽出済の糖尿病オープンデータの属性Key数は約30件

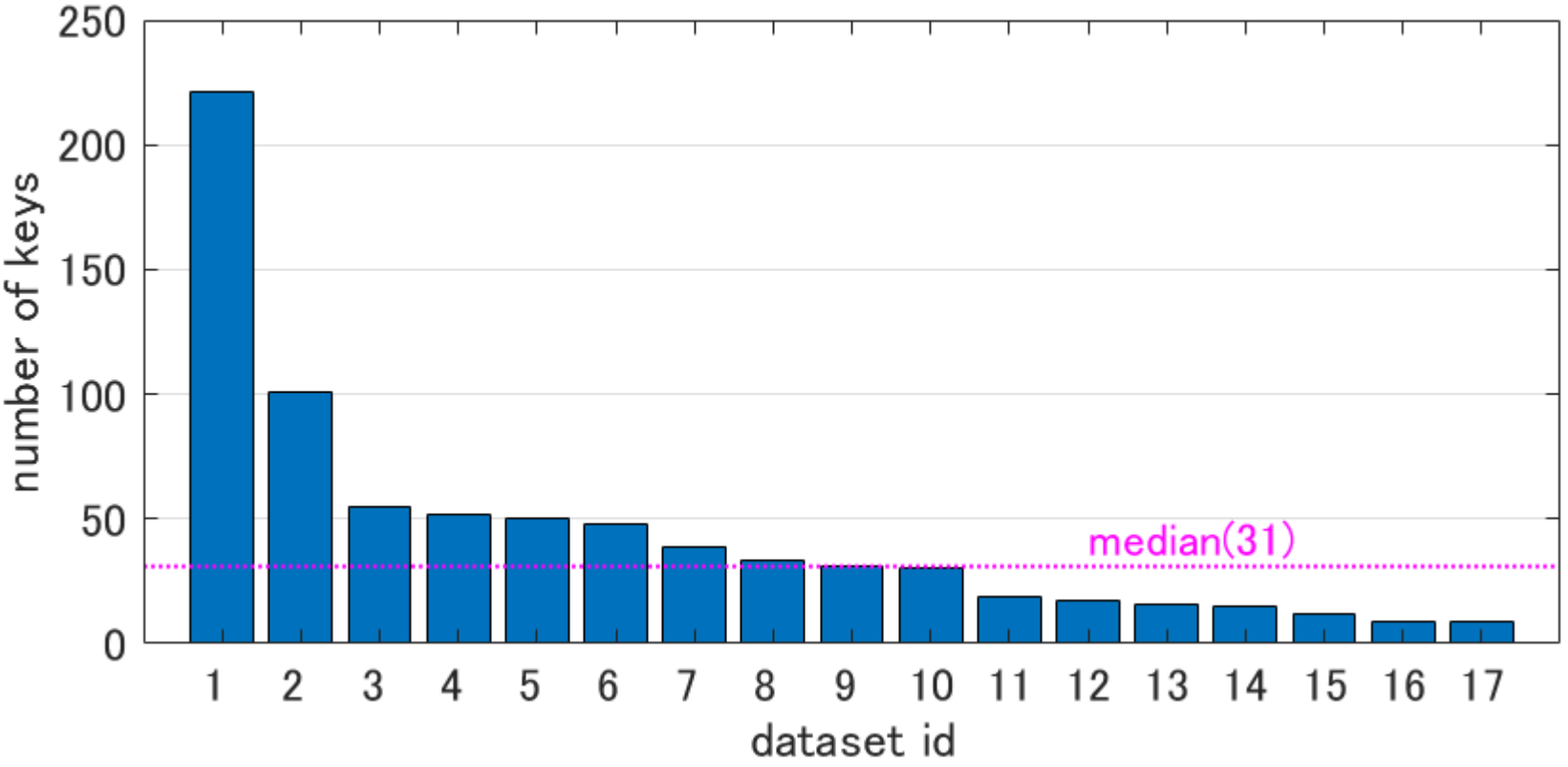
■Google Dataset Searchで“糖尿病“検索。228データセット中、17件でデータが存在

*Number of datasets : 17

*Total number of keys : 757

*Median (number of keys) : 31

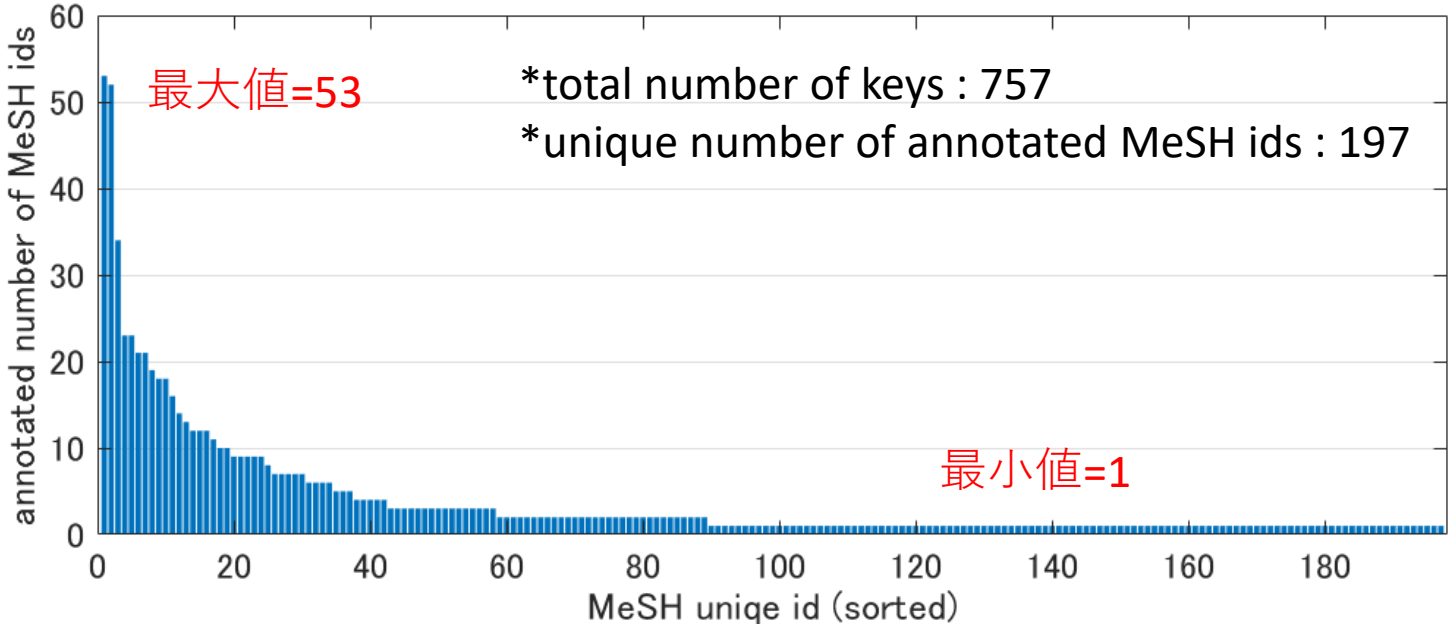
*Min,Max (number of keys) : 9, 221



1データセットで属性キー数の最大値は221,最小値は9

結果②：注釈数上位のMeSH Term

■MeSH idのユニーク数は197 (id単位の注釈数平均値=3.84,中央値=1)



■MeSH Term (+NA)の注釈割当数トップ7

注釈数順位	注釈割当数	MeSH unique id	MeSH Term
1	53	D011788	Quality of Life
2	52	NA	
3	34	D004452	Echocardiography
4	23	D001786	Blood Glucose
4	23	D001794	Blood Pressure
6	21	D003920	Diabetes Mellitus
6	21	D009273	Age Groups

結果③：１７データセットの被覆率が高いMeSH Term

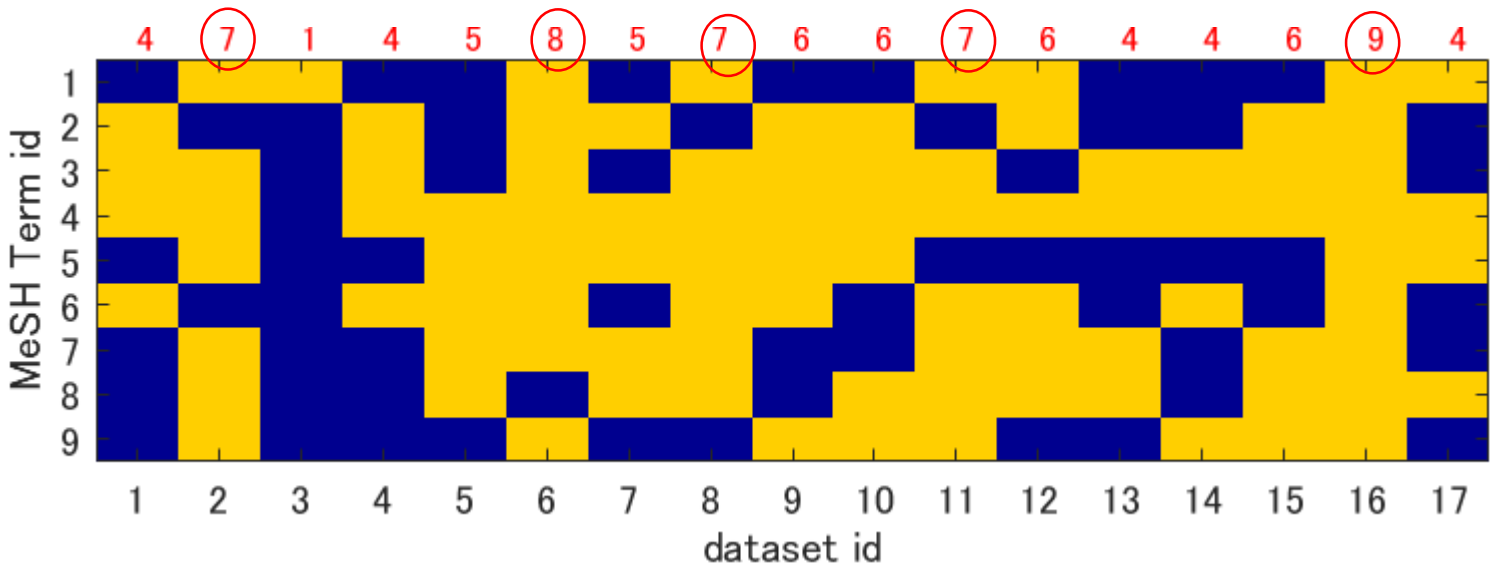
■17 datasets間で、注釈数上位のMeSH Term
(8 datasets以上被覆で赤色強調)

注釈数順位	注釈割当数	MeSH Unique ID	ヒットDataset数(～最大17)	MeSH Term	MeSH Term日本語訳	主要Term id
1	53	D011788	1	Quality of Life	QOL	
2	52	NA	3	NA		
3	34	D004452	1	Echocardiography	心エコー検査	
4	23	D001786	8	Blood Glucose	血糖値	1
5	23	D001794	9	Blood Pressure	血圧	2
6	21	D003920	**12**	**Diabetes Mellitus**	**糖尿病**	**3**
7	21	D009273	16	Age Groups	年齢群	4
8	19	D005951	4	Glucose Tolerance Test	ブドウ糖負荷試験	
9	18	D011570	2	Psychiatry	精神病治療法	
10	18	D013995	4	Time	Time	
11	16	D006442	**9**	**Glycated Hemoglobin A**	**HbA1c**	**5**
12	14	D015992	10	Body Mass Index	BMI	6
13	13	D006262	1	Health	Health	
14	12	D007328	7	Insulin	Insulin	
15	12	D009272	10	Persons	被験者番号	7
16	12	D015444	7	Exercise	Exercise	
17	11	D012723	11	Sex	Sex	8
18	10	D001835	8	Body Weight	Body Weight	9
19	10	D007004	2	Hypoglycemic Agents	血糖降下薬	

９個のMeSH Termに注目

結果④：主要なMeSH Terms の17データセット被覆数

■データセット毎の9個 MeSH Terms（有＝ ，無＝ ）



データセット毎の MeSH Term ヒット数 合計値

データセット毎に、 MeSH Term(9件）の欠損箇所が異なる。

今後の課題：データセット間での実データの統合方法？

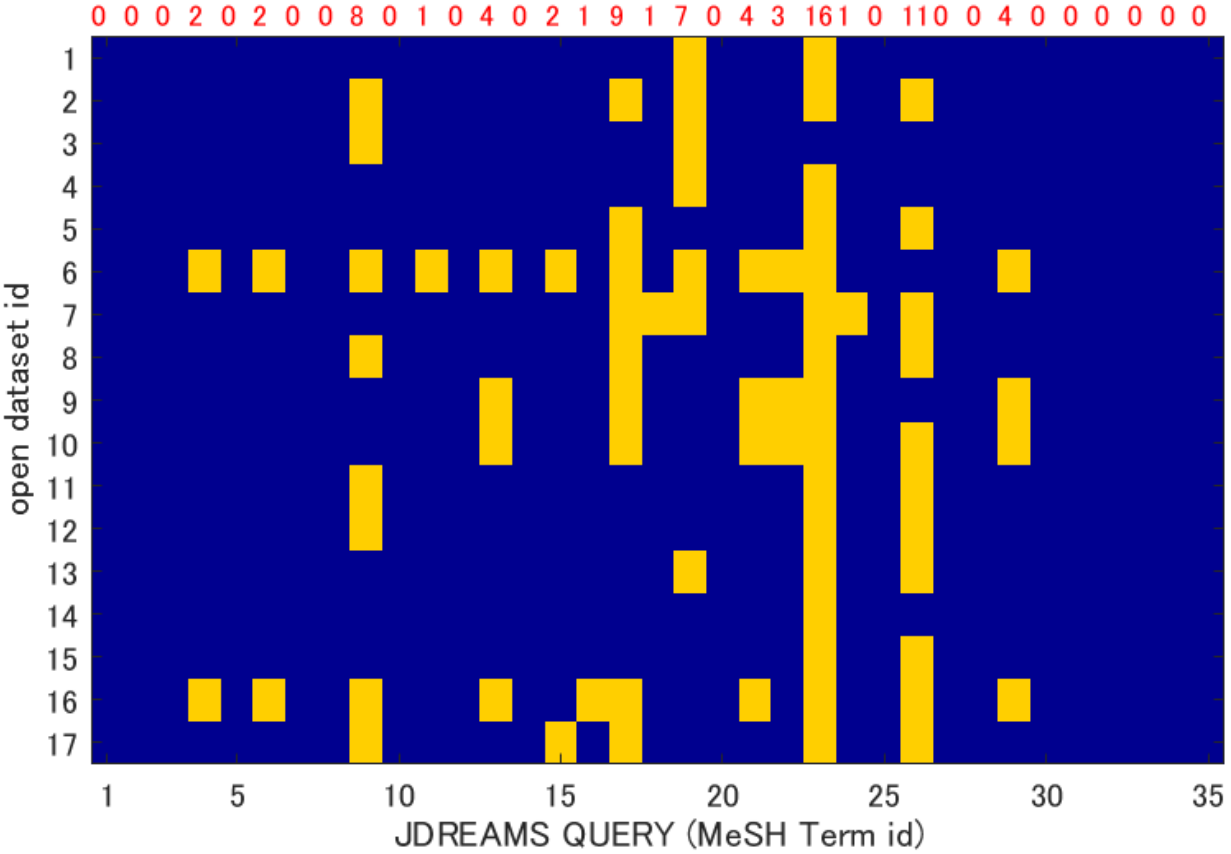
①MeSH Termが同一で、異なるデータ形式時の統合処理
②データセット間での値補正
③欠損値処理
：

④JDREAMS属性Keyとの類似度計算で１７オープンデータセットをランキング

JDREAMSの代替オープンデータを探す

J-DREAMSクエリのMeSH Termの５４％は、17 datasetsに存在しない

■MeSH Term Unique ID (J-DREAMS vs オープンデータセット)



MeSH Terms
（有＝ ，無＝ ）

JDREAMSのMeSH Term 35件中、19件が17オープンデータセットに存在しない

Term id 9 : Blood Glucose (8 datasets)
Term id 17 : Glycated Hemoglobin A (9 datasets)
Term id 23: Age Groups (16 datasets)
Term id 26: Sex (11 datasets)

J-DREAMSクエリのMeSH Termとの類似度計算で、オープンデータセットをランキング

■MeSH Term間の類似度は、Hamming距離で定義

ハミング距離

$$d_{st} = (\#(x_{sj} \neq x_{tj})/n).$$

■JDREQMSクエリに対する、17データセットのTop7ランキング結果

Rank	ID	Hamming 距離	Open Dataset Name	Number of Attributes	Number of Instances
1	6	0.657	Eisenberg et al., PLoS One, 2018. PMID: 29300770	221	92
2	16	0.714	Okamura et al., Int J Obes, 2019. PMID:29717276	31	15,464
3	10	0.800	Heier et al., PLoS One, 2018. PMID: 30359432	39	78
4	7	0.829	McCracken et al., BMJ Open, 2017. PMID:28801438	55	214
5	9	0.829	Andersson et al., PLoS One, 2015. PMID:26186716	33	63
6	2	0.857	Strack et al., Biomed Res Int 2014. PMID:24804245	50	101,766
7	17	0.857	Karakonstantis et al., Mendeley Data, 2018.	12	55
				検査値等項目数	被験者数

今後の発展

■糖尿病オープンデータの統合に向けて

Rank	ID	Hamming 距離	Open Dataset Name	Number of Attributes	Number of Instances
1	6	0.657	Eisenberg et al., PLoS One, 2018. PMID: 29300770	221	92
2	16	0.714	Okamura et al., Int J Obes, 2019. PMID:29717276	31	15,464
3	10	0.800	Heier et al., PLoS One, 2018. PMID: 30359432	39	78
4	7	0.829	McCracken et al., BMJ Open, 2017. PMID:28801438	55	214
5	9	0.829	Andersson et al., PLoS One, 2015. PMID:26186716	33	63
6	2	0.857	Strack et al., Biomed Res Int 2014. PMID:24804245	50	101,766
7	17	0.857	Karakonstantis et al., Mendeley Data, 2018.	12	55
				検査値等項目数	被験者数

課題① NLPを使った属性KeyへのMeSH Term自動割当
課題② 被験者データの統合（クレンジングと項目予測 (ex. 糖尿病)）