

研究者識別子と名寄せ

オープンサイエンス時代の学術情報環境の構築

蔵川 圭

国立情報学研究所

orcid.org/0000-0002-7031-1846

自己紹介

- 経歴

- 2012 年 – 現在:
国立情報学研究所 学術コンテンツ課 特任准教授
- 2006 年 – 2012 年:
国立情報学研究所 学術コンテンツサービス研究開発センター 特任准教授
- 1999 年 – 2006 年:
奈良先端科学技術大学院大学 情報科学研究科 助手
- 1994 年 – 1999 年:
東京大学大学院工学系研究科修士課程・博士課程
- 1990 年 – 1994 年:
東京大学工学部精密機械工学科
- 博士(工学):2001 年

- 研究・開発テーマ

- 2006 年 – 現在:
学術情報データベース
関連サービス研究開発
- 1999 年 – 2006 年:
HCI
グループウェア
ソフトウェア工学
- 1994 年 – 2001 年:
設計工学
設計学

発表の構成

- オープンサイエンス時代
- 学術知識と研究者の情動的性質
- 同姓同名の分析
- 研究者の識別子
- 研究者の名寄せ
- 研究者の識別子と名寄せのアプローチの対比分析と統合デザイン
- おわりに

Open science era

オープンサイエンス時代

学術の情報環境の変化

- 1980年代からインターネット環境の学術利用開始
- 1990年代からインターネット環境の学術利用の一般化とWebによる学術情報流通の拡大
- 2003年10月、インターネット上の学術知識のオープンアクセスを謳うベルリン宣言
- 2013年、分野を超えた研究データ共有を目指すRDA(Research Data Alliance)が結成

学術にオープンが求められる意味

- 公共的性質
 - 学術知識の創成にかかる環境は公共投資や寄付によって支えられ学術文化の醸成に寄与する
 - 人類の共有財産としての学術知識は、必要とする誰もが平和利用可能である
- 知的活動の人的可能性
 - 知識の創成が限られた専門家集団によるだけでなく、すべての人に機会が与えられる
- 知識の構造
 - 学術知識は関連しあって新たな知識を構成するという構造があることから、適切に引用されなければならない
- 知識生産のプロセス
 - 学術知識や学術資源が今まで以上にオープンとなり共有されることによって、知識の創成やイノベーションのプロセスを加速できる

論文数の年次増加量推移

世界で最初の論文が1665年に現れて(Philosophical Transactions of the Royal Society)から、下図のように年次増加し、世界の累積論文数は、2009年時点で5000万と推計[Jinha, 2010]

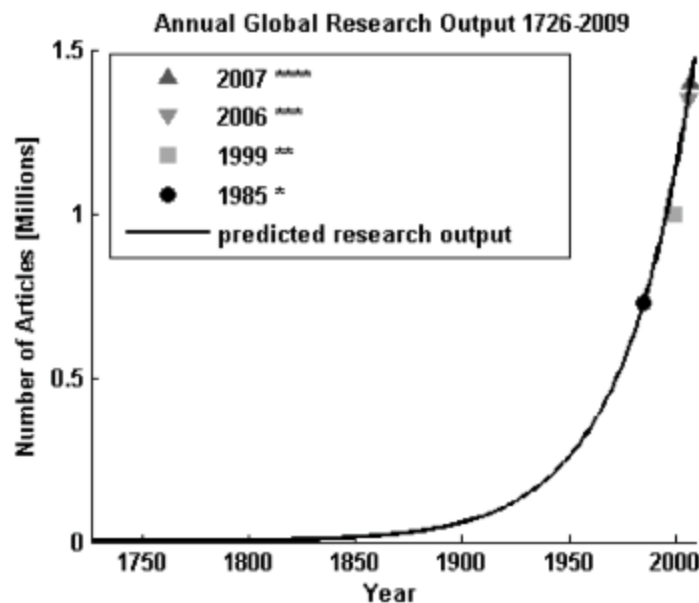


Figure 2. Estimated annual global research article output at 3% annual growth

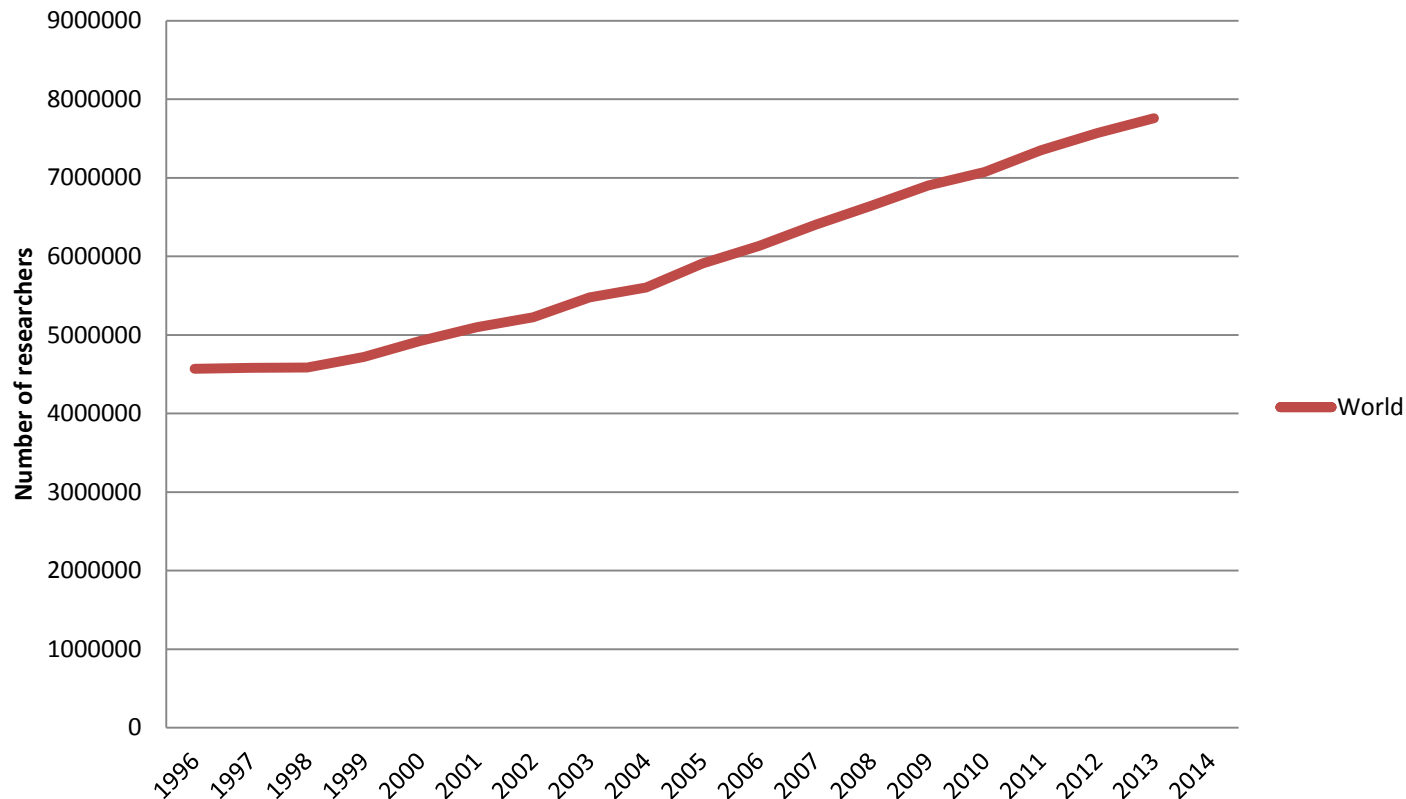
*Year 1985 (2009 – 25 years): the doubling time for annual output for articles of just under 24 years; **1999: corresponds to estimates by Tenopir and King,⁶ for research output in the late 1990s – 1 million articles per year; ***2006: corresponds very closely to Björk et al.'s,² estimate for 2006 – 1.35 million articles; ****2007: corresponds closely to Ware's,⁴ estimate for the same period – 1.4 million articles per year.

Jinha, A. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. Learned Publishing, 23(3), 258–263. <http://doi.org/10.1087/20100308>

世界の研究者FTEの推移 (1996-2013)

世界の研究者数は年々増加しており、2013年で775万人(FTE)

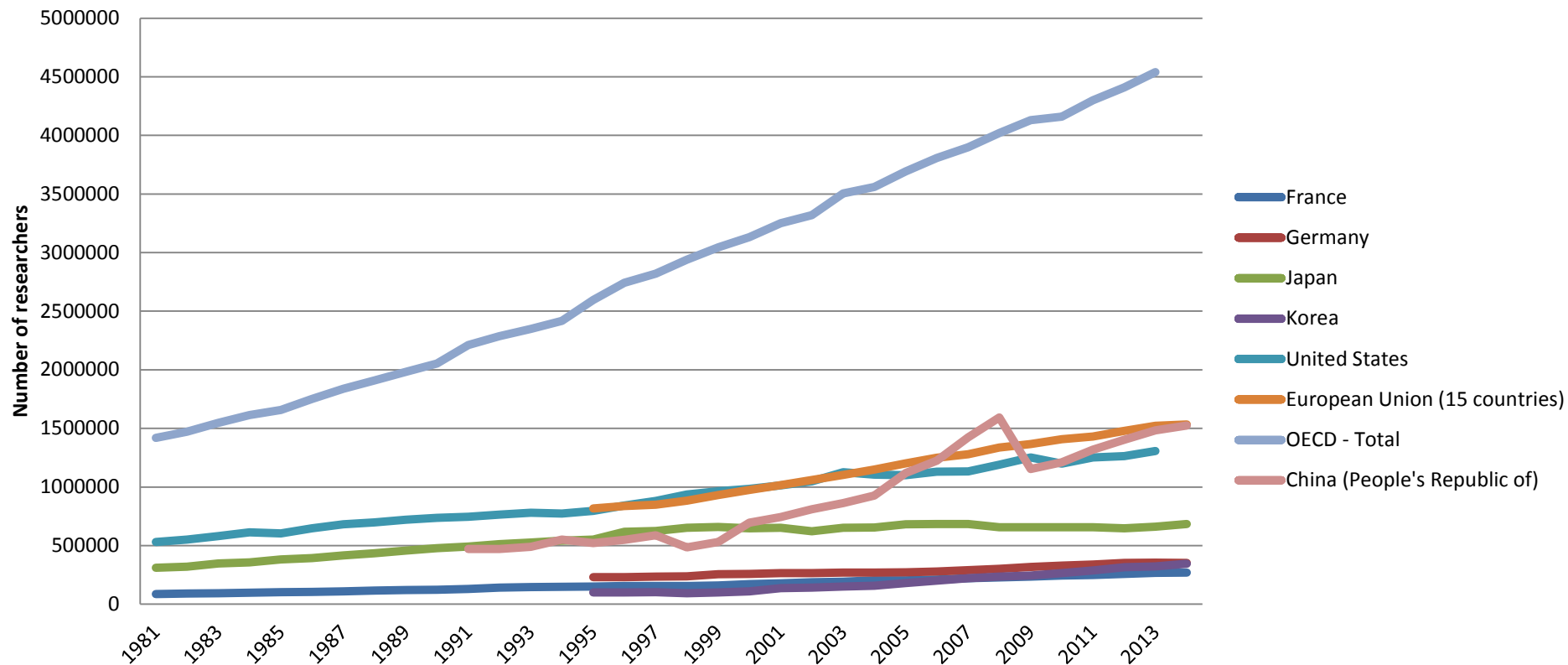
Researchers (FTE) - Total



各国の研究者FTEの推移 (1981-2014)

世界各国の研究者FTEは年々増加しているが、日本は1998年以降横ばいで推移

Total researchers (FTE)



Main Science and Technology Indicators: <http://stats.oecd.org>. Data extracted on 09 Oct 2016 13:16 UTC (GMT) from OECD.Stat

The nature of information resources of academic knowledge and the researcher

学術知識と研究者の情報的性質

学術知識の情報的性質と文化的側面

- 情報的性質
 - 学術知識は無限に拡張する
 - 学術知識は永続して存在しつづけると仮定する
- 文化的側面
 - あらゆる学術知識は、それに貢献した者の存在が明示されるべき
 - すべての貢献した者を含み、貢献の度合いが明示されるべき

文化的側面の未完成

- 重要な学術成果だけ、それに関わった者の貢献が明確に記されている
- 学術成果一般の寄与者は、姓名と所属によって記されている
- 学術成果に関わった寄与者の全ての姓名が記されているわけではない

姓名の記載は貢献の明示ではない

- 名前の曖昧性の問題
(Name Ambiguity Problem)
 - 同姓同名
 - 旧姓
 - ペンネーム
 - 漢字異体字
 - ジャーナルごとに異なる姓名表記フォーマット
 - 姓名の順
 - イニシャル表記
 - 大文字・小文字

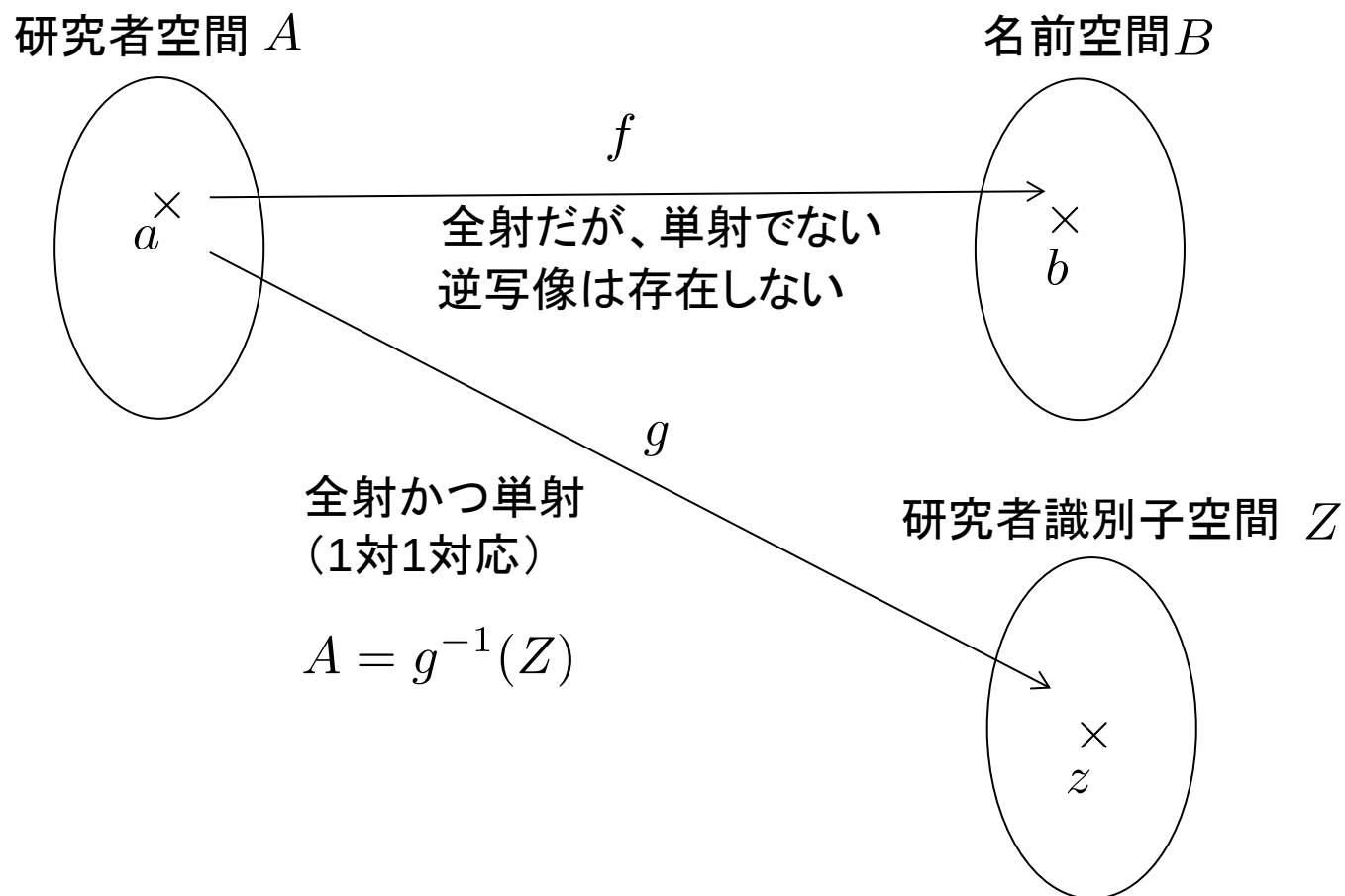
ある研究者の学術知識の集合

- ある研究者の貢献した学術知識の集合を考える
 - 学術知識の集合は、たとえば論文を元とする
- それは、どのような作り方があるか？

2つのアプローチ

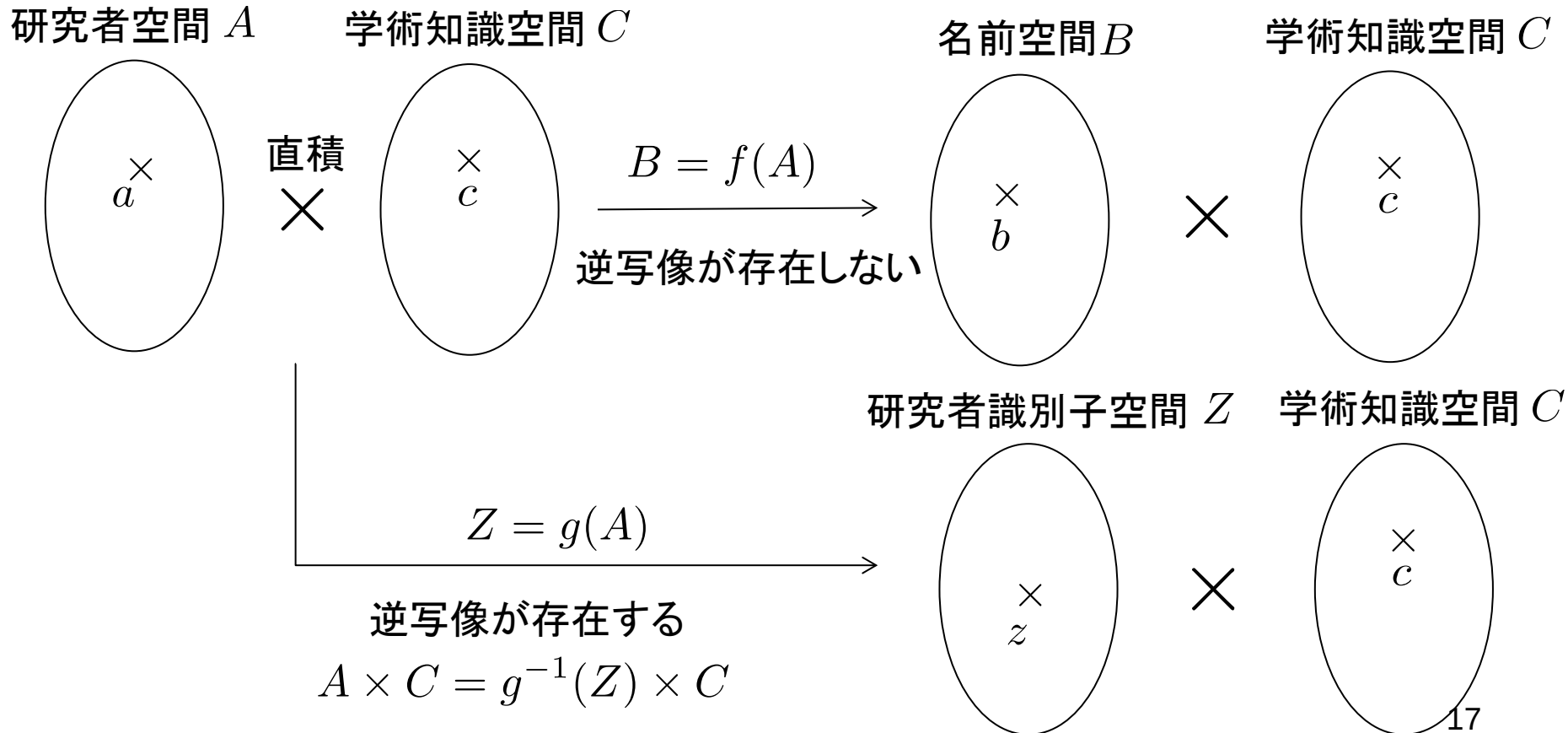
- 研究者の識別子
 - “研究者”からのアプローチ
 - 研究者に識別子を付与し学術知識を関連付け、集合を作る
 - 特徴
 - 唯一無二の研究者の識別子は、同一の目的に世界で一つあれば十分
 - 人手によって行われ、充足させるには手間と時間がかかる
- 研究者の名寄せ
 - “学術知識”からのアプローチ
 - 学術知識を研究者名寄せし、集合を作る
 - 特徴
 - データベースの管理主体ごとに名寄せを実施
 - 主に計算機を用いた手法が用いられるが、十分な性能を獲得することが難しい

研究者空間の写像

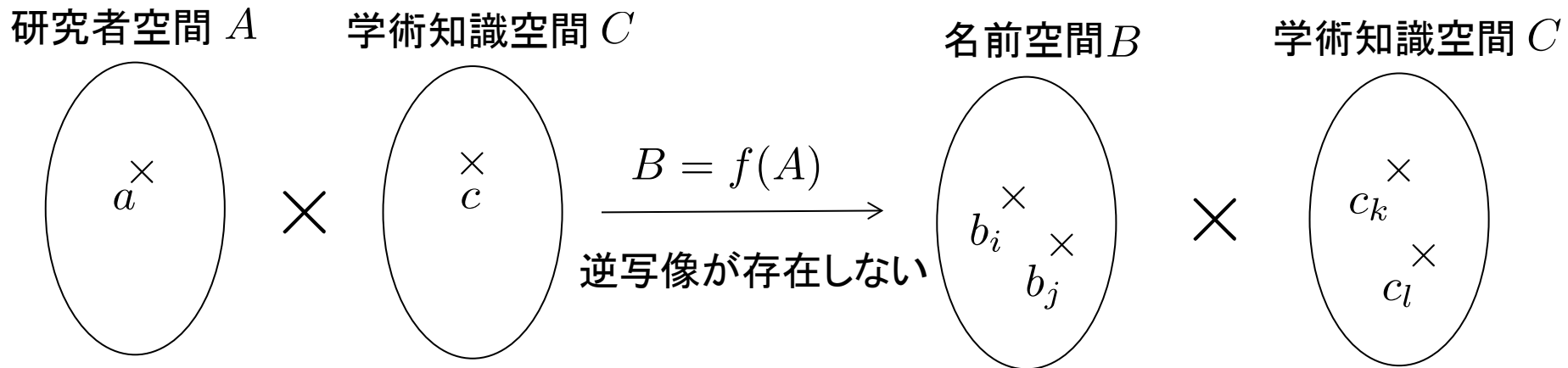


貢献の写像

研究者の貢献は、研究者空間 A と学術知識空間 C の直積 $A \times C$ で表すことができる



同一研究者による貢献の判断



名前 B 、学術知識 C 、および C から抽出できる特徴 X^C を用いて、
貢献 $B \times C$ の元 (b_i, c_k) と (b_j, c_l) が同一研究者によるものか判断する。
ただし、 $k \neq l$ とする。

The analysis of the same family and first name

同姓同名の分析

同姓の分布

- 同姓の数の分布はベキ乗則に従うことが観測されている

$$P(k) \sim k^{-\gamma}$$

k は、同姓の数を示し、 γ は定数

- γ は、文化的な振る舞いや社会的ダイナミクスに依存して決定される

- 同姓の数のランク順分布 (同姓の数/ランク) はベキ乗則に従う (Zipfの法則)

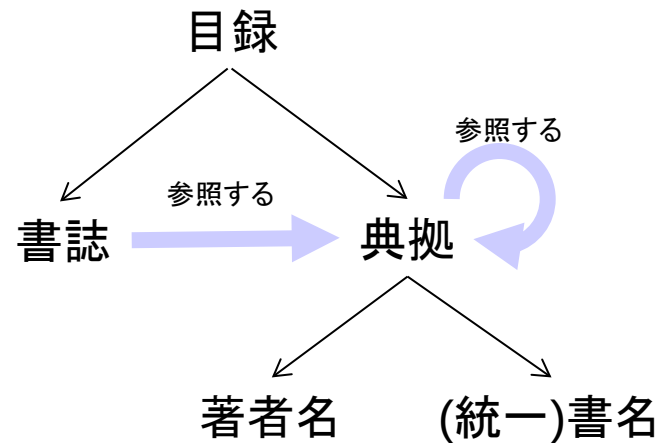
[Baek, et.al. 2007]

TABLE I. Summary of the empirical results for family name distributions. The family name distribution function is written as $P(k) \sim k^{-\gamma}$ with k the size of the family, which is the number of individuals who have the same family name. As the sampled population size N is increased, the number N_f of observed family names increases either logarithmically (China and Korea) or algebraically (other countries), giving us two distinct groups.

Region	γ	N_f
China [7]	1.0	$\ln N$
Korea [8]		$\ln N$
Argentina [9]	2.0	$N^{0.84}$
Austria [10]		$N^{0.83}$
Berlin [11]		$N^{0.90}$
France [10]		$N^{0.77}$
Germany [10]		$N^{0.75}$
Isle of Man [12]	1.5	$N^{0.65}$
Italy [10]	1.75	$N^{0.69}$
Japan [13]		$N^{1.0}$
Netherlands [10]	2.16	$N^{0.81}$
Norway [22]	0.46–1.83	$N^{0.73}$
Sicily [14]	1.9	$N^{0.69}$
Spain [10]		$N^{0.27}$
Switzerland [10]	1.94	
Taiwan [12]	1.43	
United States [11,15]		
Venezuela [16]		
Vietnam [23]		

日本の図書館目録

- NACSIS-CAT
 - 国立情報学研究所の運営
 - 日本の大学図書館の所蔵する図書・雑誌の総合目録
 - 洋書が多く含まれる
 - USMARC準拠
 - 日本目録規則, およびAACR2
- JAPAN/MARC
 - 国立国会図書館の運営
 - 国会図書館の所蔵する, 日本で刊行された出版物および外国で刊行された日本語出版物の目録
 - UNIMARC準拠
 - 日本目録規則
- TRC/MARC
 - 図書館流通センターの運営
 - 日本で刊行される出版物の目録
 - UNIMARC準拠
 - 日本目録規則

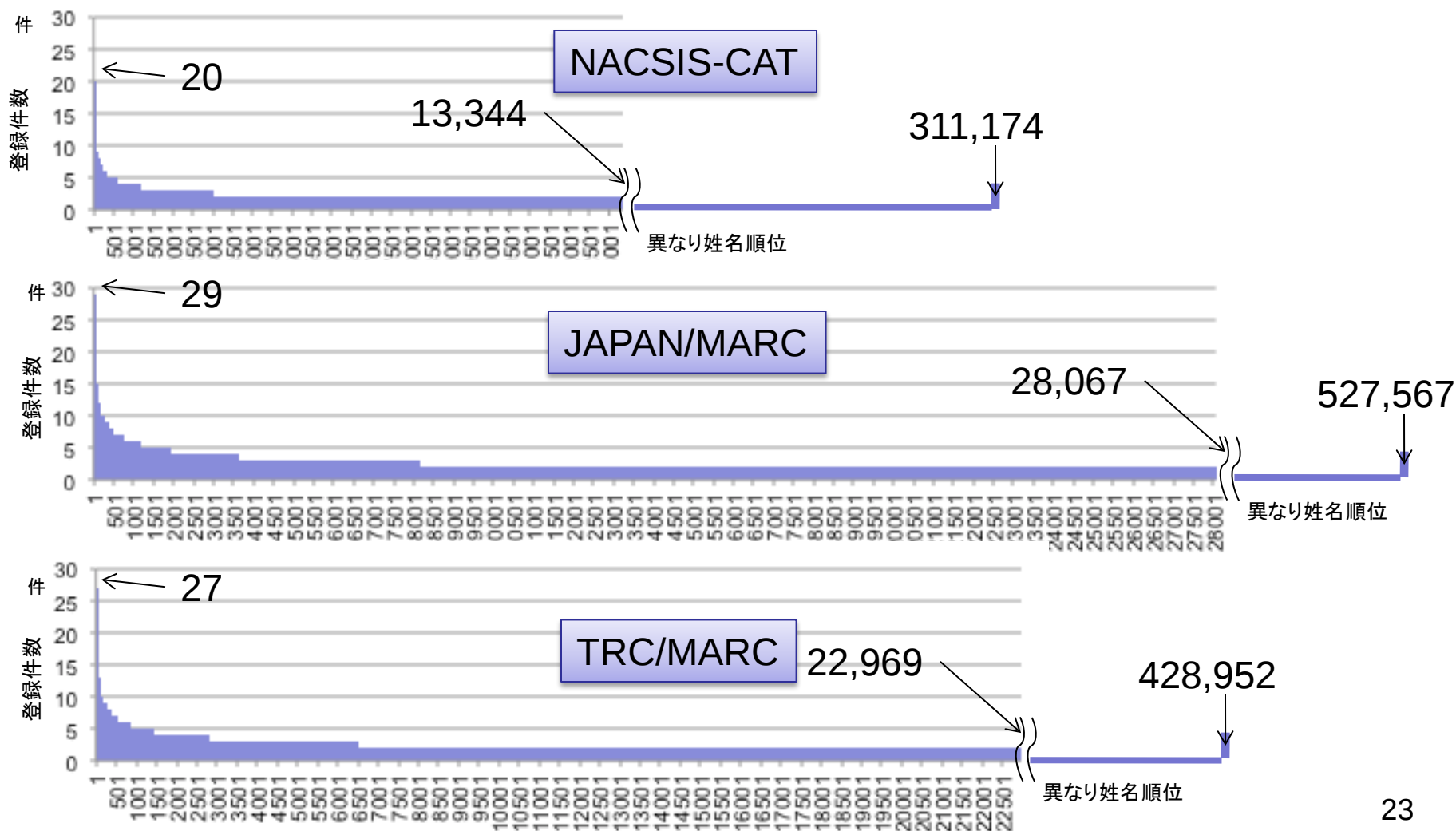


図書館目録の著者名典拠

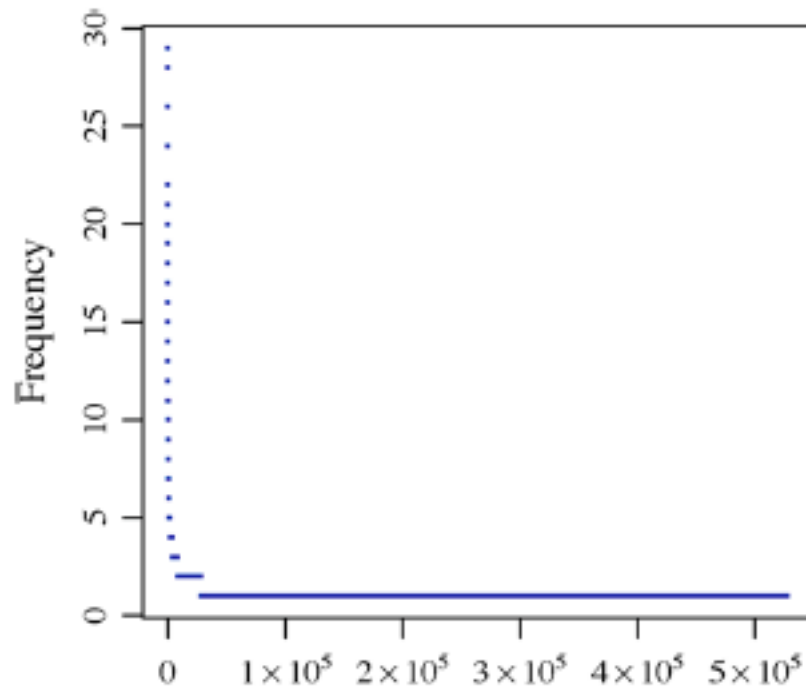
- NACSIS-CAT著者名典拠ファイル
 - 個人名, 団体名の典拠
 - 個人名1,263,685件(西洋人著者名含む, 2008年12月18日)
- JAPAN/MARC典拠ファイル
 - 個人名, 家名, 団体名, および統一書名の典拠
 - 個人名681,924件(西洋人著者名含む, 2008年7月5日)
- TRC/MARC著者名典拠ファイル
 - 個人名, 機関名の典拠
 - 個人名566,249件(西洋人著者名含む, 2009年3月29日)

人名データベース	登録件数 (漢字圏の東洋 人の統一形標目 を抜粋)	同一姓名が 複数存在する 登録件数	同一姓名が 複数存在する 異なる姓名数	同一姓名に 対する最大 登録件数	同一姓名が 複数存在する 登録件数の割合
NACSIS-CAT 著者名典拠ファイル (2008年12月18日)	329,864	32,034	13,344	20	9.71%
JAPAN/MARC 典拠ファイル (2008年7月5日)	572,638	73,138	28,067	29	12.77%
TRC/MARC 著者名典拠ファイル (2009年3月29日)	464,962	58,979	22,969	27	12.68%

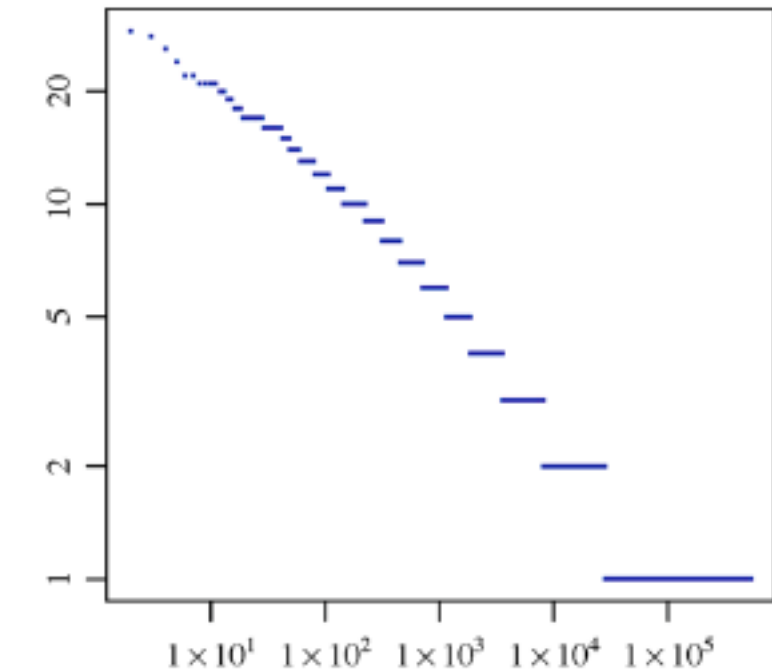
著者名典拠における 異なり姓名ごとの登録件数の分布



Japan/MARC 著者名典拠 異なり姓名の頻度分布



Normal scale



Log scale

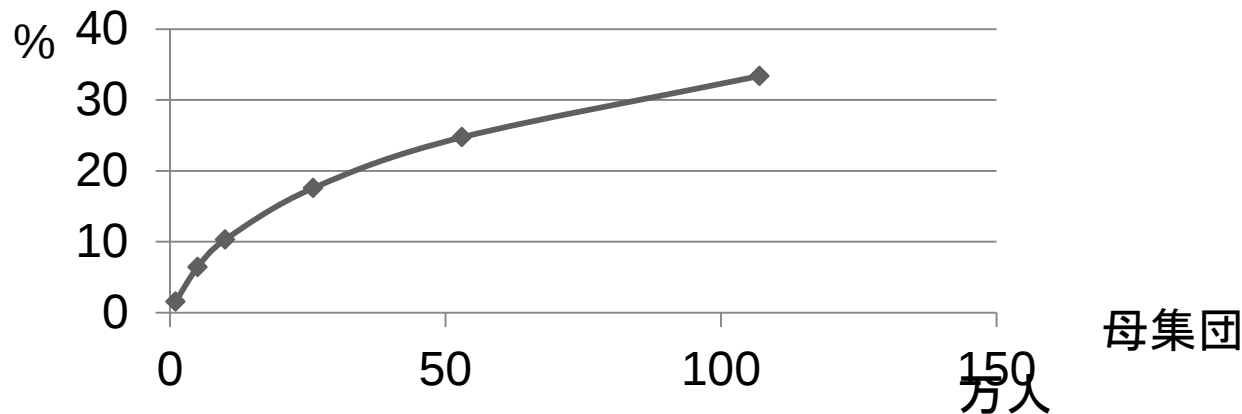
著者名典拠における 異なり姓名順位トップ20

NACSIS--CAT			JAPAN/MARC		TRC/MARC	
順位	姓名	登録件数	姓名	登録件数	姓名	登録件数
1	高橋徹	20	鈴木博	29	鈴木博	27
2	鈴木博	17	田中実	29	田中実	26
3	佐藤進	17	伊藤博	28	小林茂	24
4	田中実	16	小林茂	26	高橋徹	23
5	伊藤博	16	鈴木一郎	24	鈴木実	22
6	高橋進	13	高橋一郎	22	佐藤進	21
7	高橋清	13	佐藤正	22	渡辺誠	19
8	鈴木一郎	13	高橋徹	21	佐藤正	19
9	小林茂	13	鈴木実	21	伊藤博	19
10	吉田豊	13	田中豊	21	田中稔	18
11	高橋誠	12	(李=)	21	小林一郎	18
12	田中宏	12	鈴木茂	20	鈴木隆	17
13	渡辺誠	12	吉田稔	20	鈴木茂	17
14	渡辺茂	12	田中宏	19	田中宏	17
15	小林哲夫	12	佐藤進	19	吉田豊	17
16	田中明	11	高橋和子	18	佐藤博	17
17	佐藤正	11	渡辺誠	18	高橋進	16
18	中村宏	11	渡辺宏	18	田中豊	16
19	高橋豊	10	高橋清	17	田中茂	26
20	高橋正明	10	(陳=)	17	田中一郎	16

母集団の大きさに依存する 同姓同名の割合

- 文献によると

- 田中康仁, 同姓同名の発生頻度, 計算言語学 10-1, 1977
- 昭和51年当時の日本人の漢字姓名107万人の名簿を用いて機械的に数え上げ
同姓同名がいる人の割合(%)



Researcher identifiers

研究者の識別子

研究者識別子とは

- 定義
 - 研究者を識別することを目的として付けられた英数文字列
- 研究者識別子の品質
 - 研究者本人と識別子との間の1対1対応をどれだけ忠実に実現できているか？
 - 重複がないか？
 - 網羅性があるか？
 - 架空登録がないか？
- 研究者識別子の品質保証
 - オーソリティ
 - 本人認証

いろいろな研究者識別子

カテゴリ	サービス名	著作者の識別子例	著作者の識別子説明
図書館目録	LC MARC21	n78095332	LCCN, 最大12桁英数字
図書館目録	JAPAN/MARC	00054222	レコード管理番号, 8桁数字
図書館目録	NACSIS-CAT	DA00151899	レコードID, 10桁英数字
図書館目録	TRC/MARC	110000737940000	典拠ID, 15桁数字
著作者識別子	VIAF	96994048	VIAF ID, 桁数浮動の数字
商用研究者ディレクトリ	ResearcherID	A-1009-2008	ResearcherID, 11桁英数字
商用学術論文データベース	Scopus	7404300349	Scopus Author Identifier
学術論文データベース	CiNii	1000030413925	13桁数字(クラスターID)
学術情報データベース	J-Global	200901004071465920	J-Global ID, 18桁数字(著作者を含むID)
サブジェクトリポジトリ	arXiv	http://arxiv.org/a/warner_s_1	arXiv Author ID, オプティン
サブジェクトリポジトリ	INSPIRE	J.C.Wang.2	姓名と数字(クラスターID)
サブジェクトリポジトリ	Author Claim (RePEc)	pma2	サブジェクトリポジトリRePEcの著者サービス
研究者識別子	Names Project	http://names.mimas.ac.uk/individual/345	Names ID
研究者識別子	DAI	info:eu-repo/dai/nl/123456785	Digital Author Identifier
研究助成データベース	KAKEN	30107370	科研費研究者番号, 8桁数字
研究者ディレクトリ	VIVO	http://vivo.ufl.edu/individual/n25562	個別VIVOシステムごとにURI
研究者ディレクトリ	ReaD&Researchmap	1000255420	ReaD研究者コード, 10桁数字
研究者識別子	Researcher Name Resolver	1000010332769	研究者リゾルバーID, 13桁数字
研究者識別子	ORCID	1422-4586-3573-0476	ORCID, 16桁数字
クリエイター識別子	ISNI	1422 4586 3573 0476	ISNI, 16桁数字

研究者識別子と関連学術知識

カテゴリ	サービス名	関連学術知識
図書館目録	LC MARC21	図書
図書館目録	JAPAN/MARC	図書
図書館目録	NACSIS-CAT	図書
図書館目録	TRC/MARC	図書
著作者識別子	VIAF	図書
商用研究者ディレクトリ	ResearcherID	学術論文
商用学術論文データベース	Scopus	学術論文
学術論文データベース	CiNii	学術論文、雑誌
学術情報データベース	J-Global	学術論文
サブジェクトリポジトリ	arXiv	学術論文(数物系)
サブジェクトリポジトリ	INSPIRE	学術論文(高エネ物理)
サブジェクトリポジトリ	Author Claim (RePEc)	学術論文(経済)
研究者識別子	Names Project	機関リポジトリの学術成果一般
研究者識別子	DAI	学術成果一般
研究助成データベース	KAKEN	助成課題および成果報告書
研究者ディレクトリ	VIVO	学術成果一般
研究者ディレクトリ	ReaD&Researchmap	学術成果一般
研究者識別子	Researcher Name Resolver	学術成果一般
研究者識別子	ORCID	学術成果一般
クリエイター識別子	ISNI	楽曲、図書

研究者識別子とURI

カテゴリ	サービス名	URI例
図書館目録	LC MARC21	http://lccn.loc.gov/n78095332
図書館目録	JAPAN/MARC	http://id.ndl.go.jp/auth/ndlna/00054222
図書館目録	NACSIS-CAT	
図書館目録	TRC/MARC	
著作者識別子	VIAF	http://viaf.org/viaf/96994048/
商用研究者ディレクトリ	ResearcherID	http://www.researcherid.com/rid/A-1009-2008
商用学術論文データベース	Scopus	
学術論文データベース	CiNii	http://ci.nii.ac.jp/nrid/1000030413925
学術情報データベース	J-Global	
サブジェクトリポジトリ	arXiv	http://arxiv.org/a/warner_s_1
サブジェクトリポジトリ	INSPIRE	http://inspirehep.net/author/J.C.Wang.2
サブジェクトリポジトリ	Author Claim (RePEc)	http://authorclaim.org/profile/pma2/
研究者識別子	Names Project	http://names.mimas.ac.uk/individual/345
研究者識別子	DAI	info:eu-repo/dai/nl/123456785
研究助成データベース	KAKEN	http://kaken.nii.ac.jp/r/30107370
研究者ディレクトリ	VIVO	http://vivo.ufl.edu/individual/n25562
研究者ディレクトリ	ReaD&Researchmap	
研究者識別子	Researcher Name Resolver	http://rns.nii.ac.jp/nr/1000010332769
研究者識別子	ORCID	http://orcid.org/1422-4586-3573-0476
クリエイター識別子	ISNI	

研究者識別子と品質保証

カテゴリ	サービス名	オーソリティ	オーソリティを高める方法	本人認証
図書館目録	LC MARC21	○	キュレーター	なし
図書館目録	JAPAN/MARC	○	キュレーター	なし
図書館目録	NACSIS-CAT	○	キュレーター	なし
図書館目録	TRC/MARC	○	キュレーター	なし
著作者識別子	VIAF	○	典拠のリユース、クラスタリング	なし
商用研究者ディレクトリ	ResearcherID	△	本人申告	あり
商用学術論文データベース	Scopus	×	クラスタリング	なし
学術論文データベース	CiNii	×	クラスタリング	なし
学術情報データベース	J-Global	×	クラスタリング	なし
サブジェクトリポジトリ	arXiv	△	本人申告	あり
サブジェクトリポジトリ	INSPIRE	△/○	本人申告、キュレーター	あり
サブジェクトリポジトリ	Author Claim (RePEc)	△	本人申告	あり
研究者識別子	Names Project	×	クラスタリング	なし
研究者識別子	DAI	○	キュレーター	あり
研究助成データベース	KAKEN	○	キュレーター	なし(e-rad上はあり)
研究者ディレクトリ	VIVO	△	本人申告	あり
研究者ディレクトリ	ReaD&Researchmap	△	本人申告	あり
研究者識別子	Researcher Name Resolver	○	KAKENの識別子のリユース	なし
研究者識別子	ORCID	△/○	本人申告、外部組織承認	あり
クリエイター識別子	ISNI	○	キュレーター	なし

研究者識別子ORCID

- ORCID (orcid.org)は国際的な研究者識別子としてもっとも認知されている
- 研究者識別子を意識しながら、研究者プロフィールのサービスを展開
- 世界有数の著名な出版社や大学、研究機関、学会、ファンディング機関を巻き込んで識別子付与(500以上のメンバー組織)
- オープンソースを高める仕組みが内蔵されている
- ORCIDを中心とした学術エコシステムを構想

The screenshot shows the ORCID iD profile for Kei Kurakawa. At the top, the ORCID logo is displayed with the tagline "Connecting Research and Researchers". Navigation tabs for "FOR RESEARCHERS", "FOR ORGANIZATIONS", and "ABOUT" are visible. The profile section includes the name "Kei Kurakawa", the ORCID ID "0000-0002-7031-1846", and a list of other identifiers: ResearcherID: A-7273-2008, Scopus Author ID: 56032837500, and Researcher Name Resolver ID: 1000 010332769. The "Country" is listed as Japan. The "Keywords" section lists "digital libraries, library and information sciences, design engineering, researcher identifier". The "Biography" section contains a paragraph about the researcher's interests in information technology and digital libraries. The "Education" section lists "The University of Tokyo: Tokyo, Japan" from 1996-04 to 2001-12, with a Ph.D. in Engineering. The source is cited as "Source: Kei Kurakawa".

ORCID
Connecting Research and Researchers

FOR RESEARCHERS FOR ORGANIZATIONS ABOUT

SIGN IN REGISTER FOR AN ORCID ID LEARN MORE

Kei Kurakawa

ORCID ID
id.orcid.org/0000-0002-7031-1846

Also known as
蔵川圭

Country
Japan

Keywords
digital libraries, library and information sciences, design engineering, researcher identifier

Other IDs
ResearcherID: A-7273-2008
Scopus Author ID: 56032837500
Researcher Name Resolver ID: 1000 010332769

Biography

Digital library for scholarly communications on the Web is the researcher identifier management system, the Researcher Nan identifiers for Japanese researchers. The identifiers are used to institutional repositories for both publications and data. I am a Grants-in Aid for Scientific Research in Japan. It characteristic achievements. Identifiers of scholarly digital objects and creation knowledge and research trends.

My research interests include information technology to develop analyze the nature of research. Web-oriented technology, semi learning, and statistical models are the keywords.

Past research interests are design science, design engineering, engineering. These fundamental researches affect my current development.

▼ Education (3)

The University of Tokyo: Tokyo, Japan
1996-04 to 2001-12
Ph.D. in Engineering
Source: Kei Kurakawa

Researcher name disambiguation

研究者の名寄せ

書誌データベースの一つの問題

- データベースの一つとして書誌データベースを想定する。それに関連する代表的な問題の一つに、著者名曖昧性解消がある。

「小林 誠」 $\neq$ 「小林 誠」

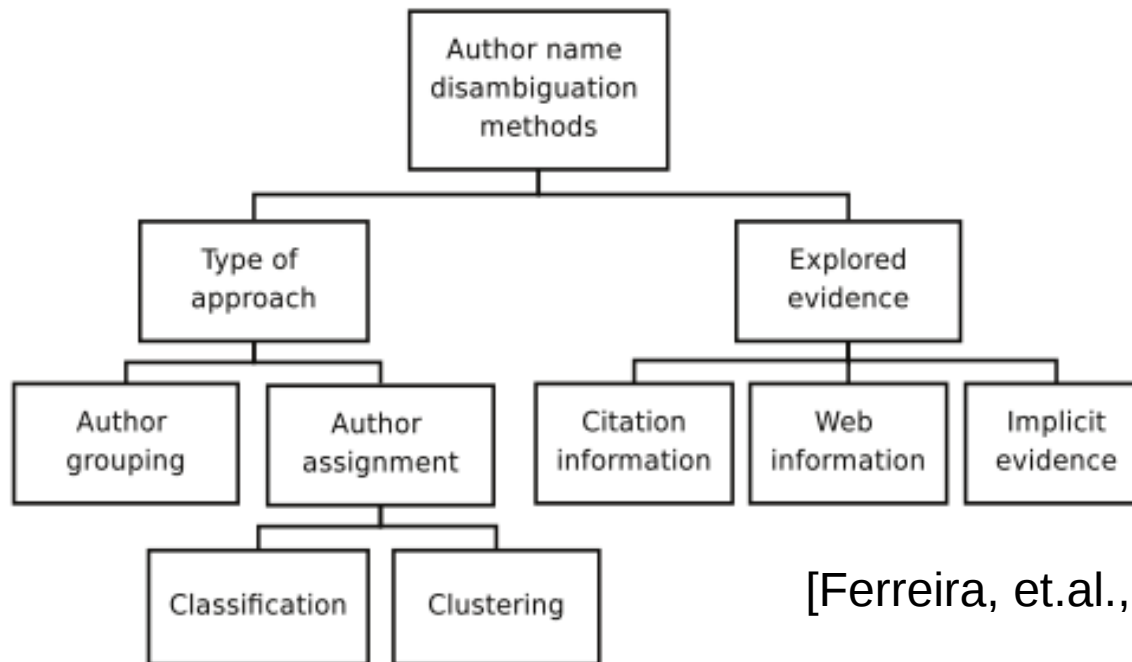
小林 誠, 渡部 欣忍, 松下 隆", 大腿骨骨折をmipoで治療する場合の問題点, 骨折, (2009)

小林 誠, マーケティング概念のシステム論的再構成: マーケティング・コミュニケーション・ミックスを例として, 社会環境研究, (2003)

小林 誠, Chiral Symmetry, 素粒子論研究, (1970)

著者名曖昧性解消手法の分類

データベースにおける名前曖昧性解消は1950年代から検討*される問題であり、すでにいくつかの解法が提案されているがいまだ十分な解決をみていない。特に「機械学習」と近年呼ばれる技術を組み合わせた解決法の提案がなされてきた。



[Ferreira, et.al., 2012]

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. ACM SIGMOD Record, 41(2), 15. doi:10.1145/2350036.2350040

* NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J., & JAMES, A. P. (1959). Automatic linkage of vital records. Science (New York, N.Y.), 130, 954–959. doi:http://dx.doi.org/10.1126%252Fscience.130.3381.954

機械学習とは

- 訓練集合(training set)の要素ベクトル x が目標ベクトル(target vector) t に対応しているとする。
 - たとえば、手書き数字認識問題を考えると、訓練集合の要素は手書き数字画像であり、目標ベクトルは数字分類である。
- 機械学習によって得られるのは、この対応を表した関数 $y(x)$ である。
- 訓練、あるいは学習によって得られた関数 $y(x)$ を用いて、新たな入力要素に対して分類予測できる。

機械学習の類別

- 教師あり学習(supervised learning)
 - 入力ベクトルと目標ベクトルの対応が与えられる
 - クラス分類問題(classification)
 - 回帰(regression)
- 教師なし学習(unsupervised learning)
 - 入力ベクトルのみ与えられる
 - クラスタリング(clustering)
 - 密度推定(density estimation)
 - 視覚化(visualization)のための次元圧縮
- 強化学習(reinforcement learning)
 - 与えられた状況下で、報酬を最大にするような適当な行動を見つける問題

書誌の著者クラスター化

CID,AFID,JNAME,ENAME,YNAME,JAFF,EAFF,YEAR,CO-AUTH,TITLE,JRNL

16,B-10002917029-1-CJP,松本健一,:NA,マツモト ケンイチ,国立遺伝研・進化,:NA,1996,

"松本 健一,山本 博士,白吉 安昭,竹田 直樹,相沢 慎一,池村 淑道,中辻 憲夫",

遺伝子ターゲティングによる細胞外マトリックス・テネイシンxの機能解析,日本分子生物学会年会プログラム・講演要旨集

20,B-10003797238-1-CJP,松本健一,:NA,マツモト ケンイチ,日新製鋼(株)呉製鉄所 圧延部,Wakamatsu Works Hitachi Metals Ltd.,1997,

"松本 健一,村重 進,石川 淳,森野 泰司,早川 一,空野 博明",

熱間薄板圧延仕上げ前段ミルロールにおける表面損傷の超音波法による評価技術の開発,材料とプロセス:日本鉄鋼協会講演論文集 = Current Advances In Materials And Processes: Report Of The Isij Meeting

20,B-10003797238-1-JIC,松本健一,Matsumoto Ken'ichi,マツモト ケンイチ,日新製鋼 呉製鉄所,Nisshin Steel Co. Ltd. Kure Work.,1997,

"松本 健一,村重 進,石川 淳,森野 泰司,早川 一,空野 博明",

熱間薄板圧延仕上げ前段ミルロールにおける表面損傷の超音波法による評価技術の開発,材料とプロセス:日本鉄鋼協会講演論文集 = Current Advances In Materials And Processes: Report Of The Isij Meeting

CID:=同一著者のクラスター

ID

正解

3. 比較

AFID:=著者フラグメントID

JNAME:=著者名

YEAR:=出版年

CO-AUTH:=共著者名

TITLE:=論文題目

JRNL:=雑誌名

判別解

AFIDのクラスタ化

1. 著者特徴量抽出
2. クラスタリグ

著者クラスタ化の性能指標

- PurityおよびInverse-Purityで測る
 - アイテムが一つのクラスターに属する場合の、クラスタリング評価指標
 - Web people search task I (2007)で使用された

$$\text{Purity} = \sum_i \frac{|C_i|}{n} \max_j \text{Precision}(C_i, L_j) \quad \text{Precision measure}$$

$$\text{Precision}(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|}$$

$$\text{Inverse Purity} = \sum_i \frac{|L_i|}{n} \max_j \text{Precision}(L_i, C_j) \quad \text{Recall measure}$$

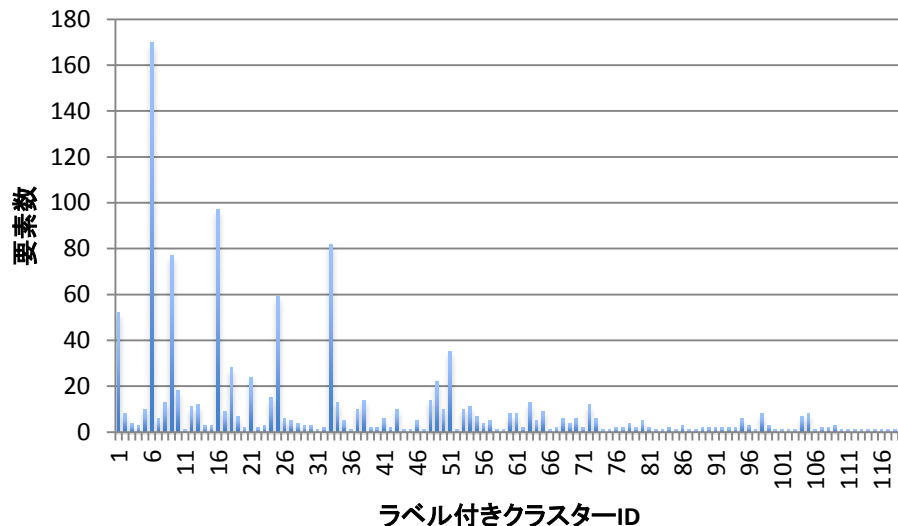
C : the set of clusters to be evaluated

L : the set of categories (manually annotated)

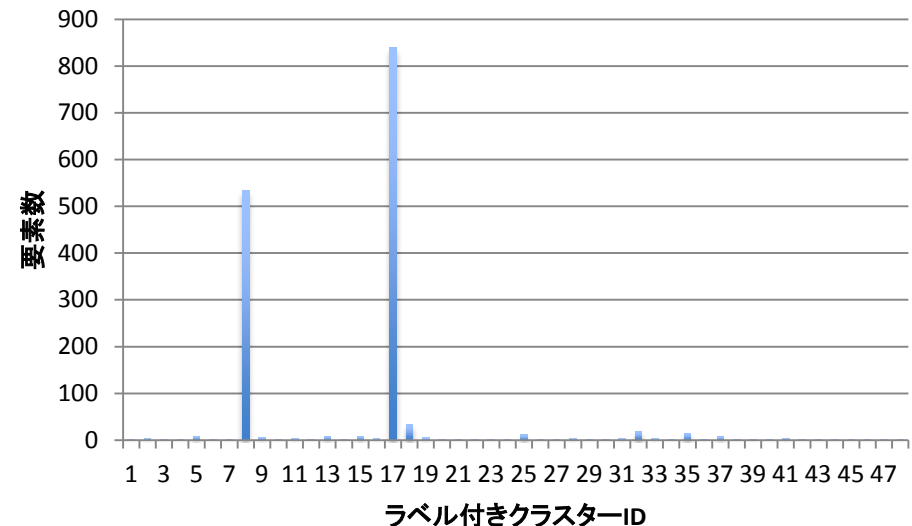
n : the number of clustered elements

2つのデータセットを用いた実験

小林誠データセット (Koba-ci dataset)
:レコード数 = 1121, クラスター数=119
119人の同姓同名が含まれている



松本健一データセット (Matu-ci dataset)
:レコード数=1568, クラスター数=48
48人の同姓同名が含まれている



実験結果の例 (性能)

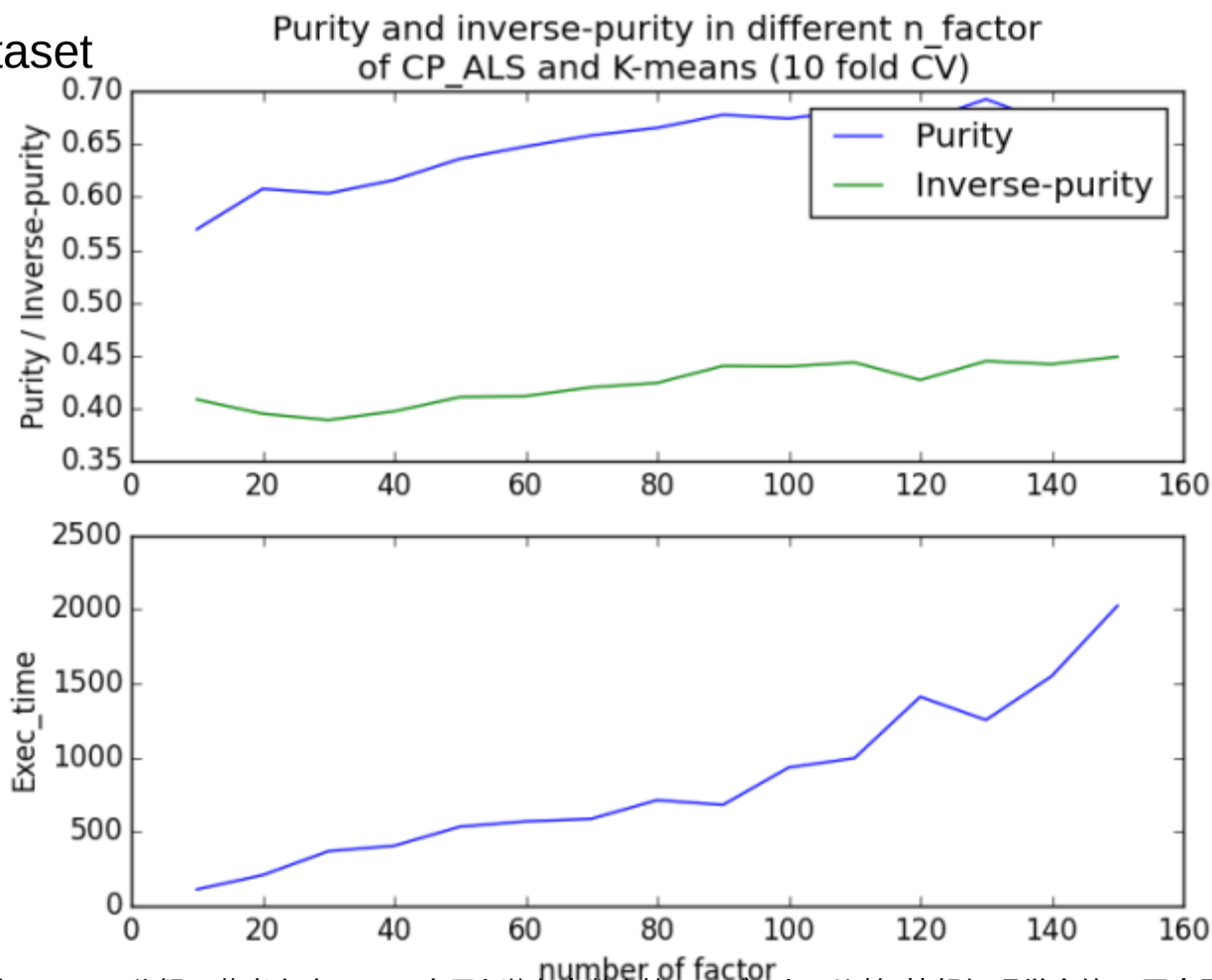
Dataset	CP ALS and K-means				RESCAL ALS and K-means			
	(R=14)		(R=200)		(R=14)		(R=100)	
	Purity	Inv.-	Purity	Inv.-	Purity	Inv.-	Purity	Inv.-
Koba-ci	0.5834	0.3943	0.6949	0.4540	0.6227	0.3782	0.7859	0.4558
Matu-ci	0.8616	0.1888	0.8712	0.3036	0.8929	0.2245	0.8093	0.1849

Dataset	LDA		Random labeling (baseline)	
	Purity	Inv.-	Purity	Inv.-
Koba-ci	0.7565	0.4835	0.2194	0.1320
Matu-ci	0.8948	0.2251	0.5434	0.0721

蔵川圭, 馬場康維, テンソル分解の著者名寄せへの応用と潜在変数を持つモデルとの比較, 情報処理学会第78回全国大会講演論文集, pp. 1-495 – 1-496, 東京, 慶応大学矢上キャンパス, 2016.03.10-12

実験結果の例(性能および実行時間)

Koba-ci dataset



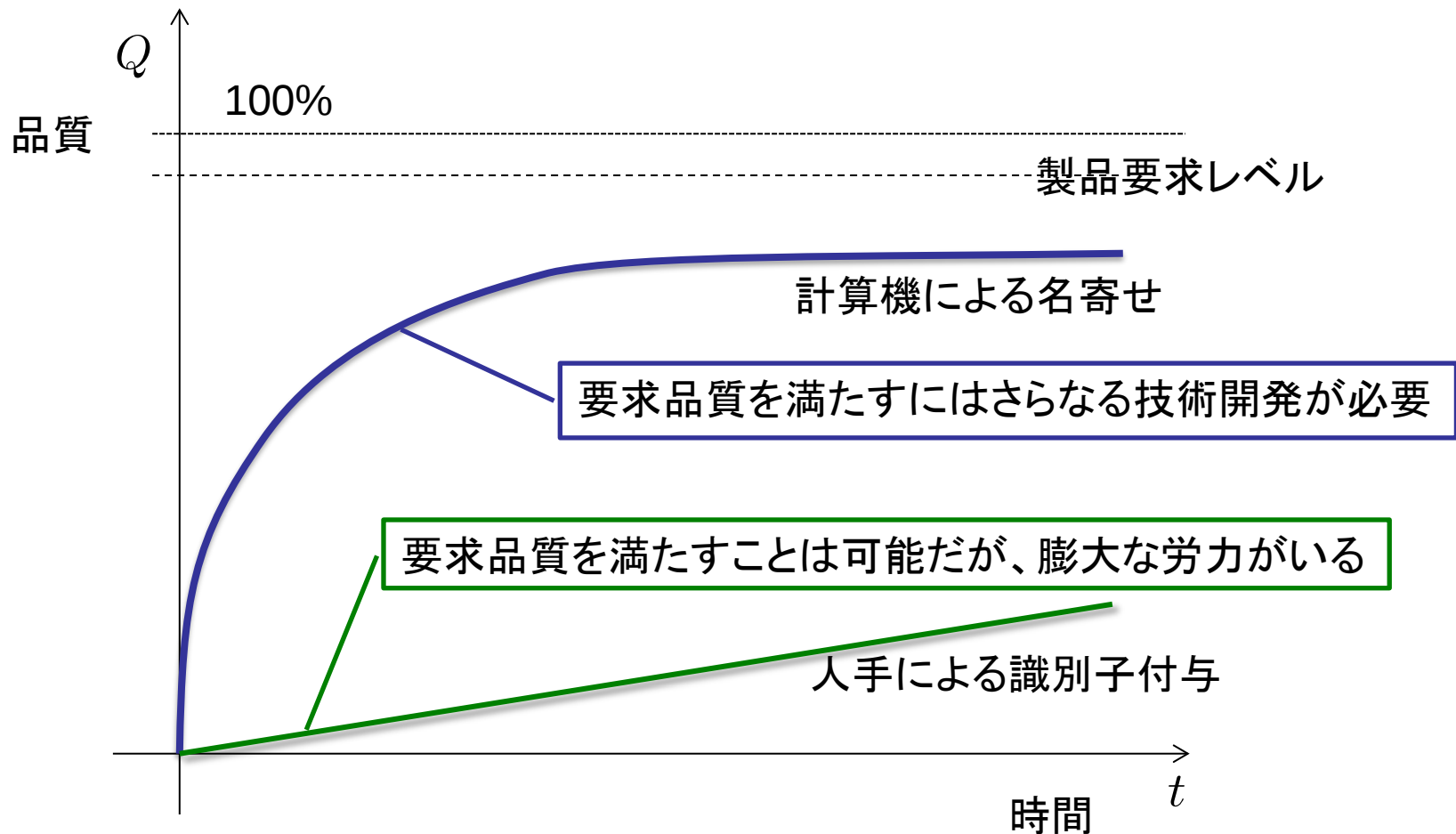
蔵川圭, 馬場康維, テンソル分解の著者名寄せへの応用と潜在変数を持つモデルとの比較, 情報処理学会第78回全国大会講演論文集, pp. 1-495 – 1-496, 東京, 慶応大学矢上キャンパス, 2016.03.10-12

Comparative analysis and integration design between researcher identifier approach and name disambiguation approach

研究者の識別子と名寄せのアプローチの対比分析と統合デザイン

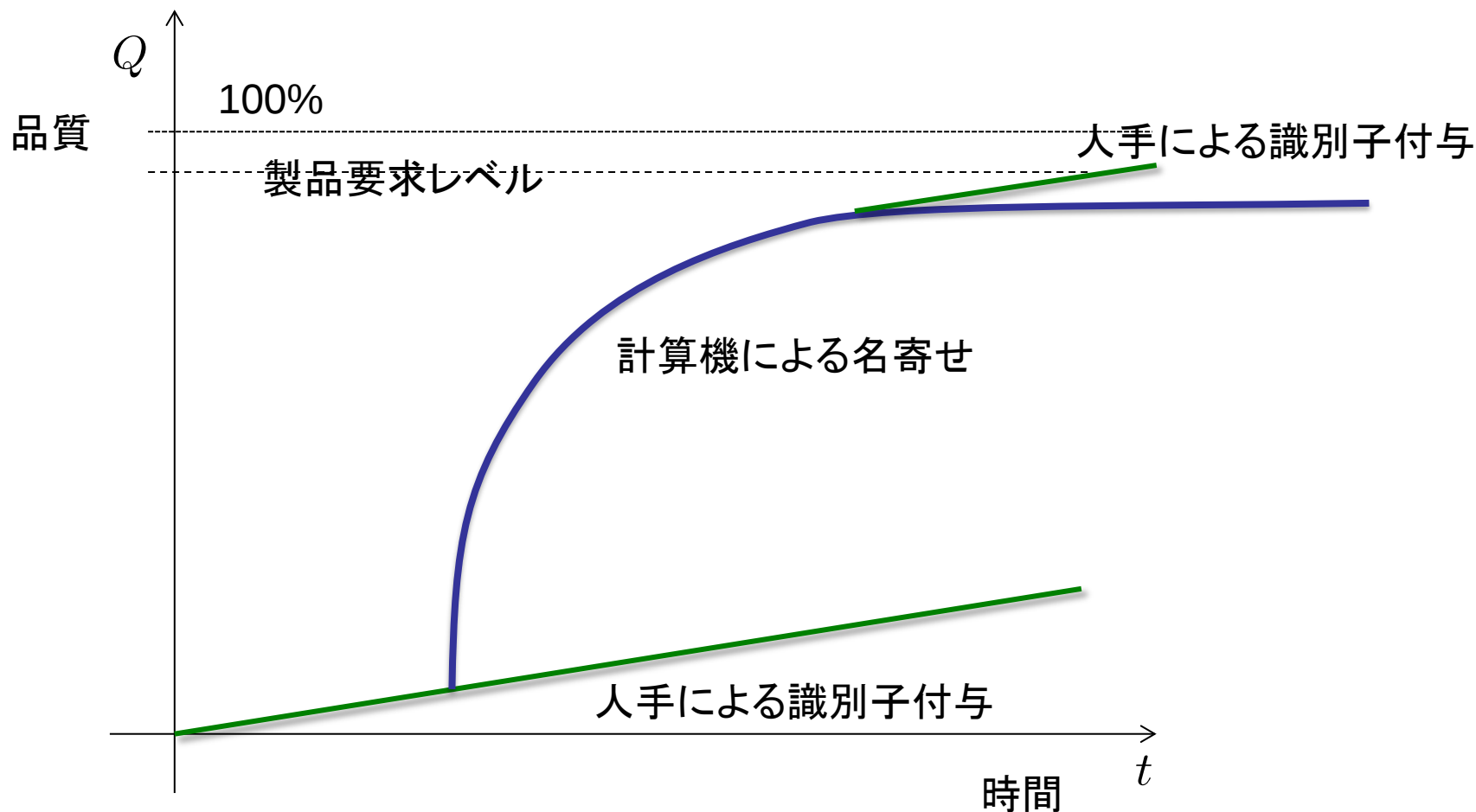
研究者を識別する

2つのアプローチのコストと品質



時間に比例してコストがかかるとする

研究者を識別する ハイブリッドな方法のコストと品質



時間に比例してコストがかかるとする

CERN INSPIREの研究者ページ



Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNames :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

Abe, Fumio

Profile Name

Search

View Profile

Manage Profile

Manage Publications

Help

2016-10-06 12:50:30

Personal Details (HepNames)

Name	Fumio Abe (阿部 文雄)
Current Institution	Nagoya U.
Links	http://researchmap.jp/read0179...
Fields	ASTRO-PH
Identifiers	BAI: F.Abe.1 INSPIRE: INSPIRE-00411685 KAKEN: 80184224

Period	Rank	Institution
	SENIOR	Nagoya U.
	JUNIOR	KEK, Tsukuba

Update Details

Publications Datasets External

1. J-GEM Follow-Up Observations to Search for an Optical Counterpart of The First Gravitational Wave Source GW150914
2. MOA-2010-BLG-328Lb: A Sub-Neptune Orbiting very Late M Dwarf?
3. Interpretation of a Short-Term Anomaly in the Gravitational Microlensing Event MOA-2012-BLG-486
4. Microlensing Discovery of a Tight, Low Mass-ratio Planetary-mass Object around an Old, Field Brown Dwarf
5. The Microlensing Event Rate and Optical Depth Toward the Galactic Bulge from MOA-II
6. Extending the Planetary Mass Function to Earth Mass by Microlensing at Moderately High Magnification
7. A Giant Planet beyond the Snow Line in Microlensing Event OGLE-2011-BLG-0251
8. Microlensing Discovery of a Population of Very Tight, Very Low-mass Binary Brown Dwarfs
9. MOA-2010-BLG-073L: An M-Dwarf with a Substellar Companion at the Planet/Brown Dwarf Boundary
10. MOA-2010-BLG-523: 'Failed Planet' = RS CVn Star

[Click here to see all](#)

Publication Graph



国際的なPIDと統合デザイン

PID(永続的識別子)のタイプ

デジタルオブジェクト、研究データ

研究者

組織

グラント

OpenAIRE

PANGAEA

EUDAT

ePIC (PID services for the European Research Community)

IGSN (international geoscience number)

FORCE11 data citation principles

CrossRef

DataCite

THOR

DOI

Handle System

eduPersonOrgOrcid

ORCID

eduPersonOrgDN

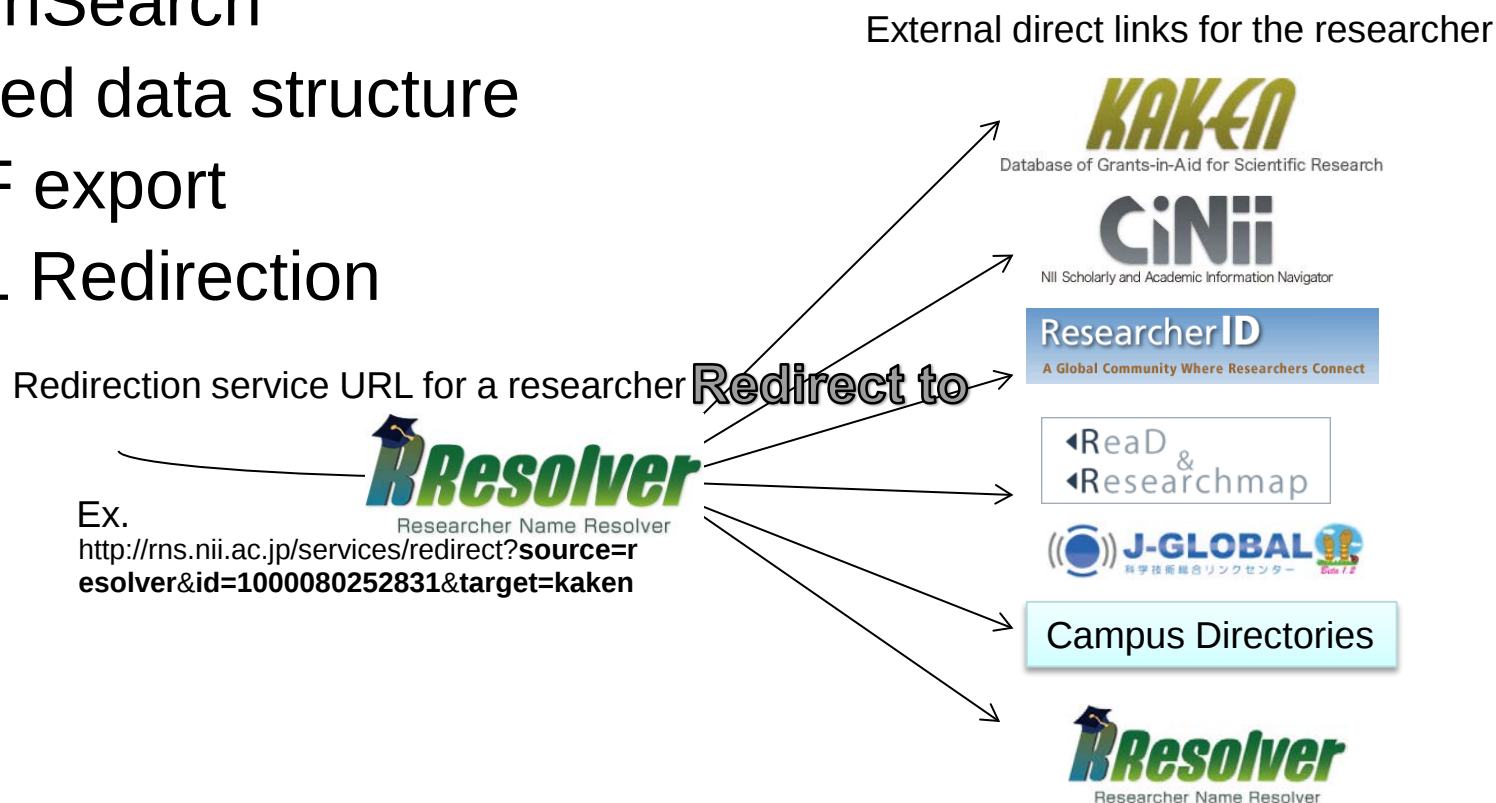
Federation
Identity
Management

PIDの機能

- PIDのブリッジ機能
 - 複数のデータベースの統合利用には、共通の外部識別子がブリッジ機能を果たす
 - DOIだけでなく、ORCIDもそれと同様の機能を果たす
- 研究者識別子のブリッジ機能
 - ORCIDがブリッジ機能を果たすこととして期待される
 - KAKENの研究者番号は、ブリッジ機能を果たすように推奨したため、機関リポジトリには著者番号として付与されていることがある
 - 国内では、2つの研究者番号が並走しているため、これを吸収する対応関係データベースが必要である
 - 研究者リゾルバーは、科研費研究者番号とORCIDの対応関係をもつ

研究者リゾルバーのWebサービス

- Researcher URI
- OpenSearch
- Linked data structure
- RDF export
- URL Redirection



Kurakawa, K., Takeda, H., Takaku, M., Aizawa, A., Shiozaki, R., Morimoto, S., & Uchijima, H. (2014). Researcher Name Resolver: identifier management system for Japanese researchers. International Journal on Digital Libraries, 14(1–2), 39–58.
<http://doi.org/10.1007/s00799-014-0109-z>

学術のオープンデザイン

- オープンサイエンス
 - 論文のオープンアクセス
 - 研究データ共有
- 学術環境のオープンデザイン
 - エコシステム
 - シェアードシステム
 - インターネット上のオールインワンプレイス
- 学術文化の継承
 - 人類の知の創生プロセスの再確認
 - On the shoulder of the giant (巨人の肩の上に立つ)
 - 知の創生プロセスに関わるすべての貢献者への賛辞

おわりに

- 増え続ける学術知識と研究者
- 学術知識への貢献の明示
 - 研究者識別子
 - 研究者名寄せ
- 学術のオープンデザイン
 - オープンサイエンス
 - 学術環境のオープンデザイン
 - 学術文化の継承