



第24回図書館総合展

フォーラム「学術情報発見サービスの表と裏」

2022.11.16

# CiNii Researchを支える データ整備の裏話

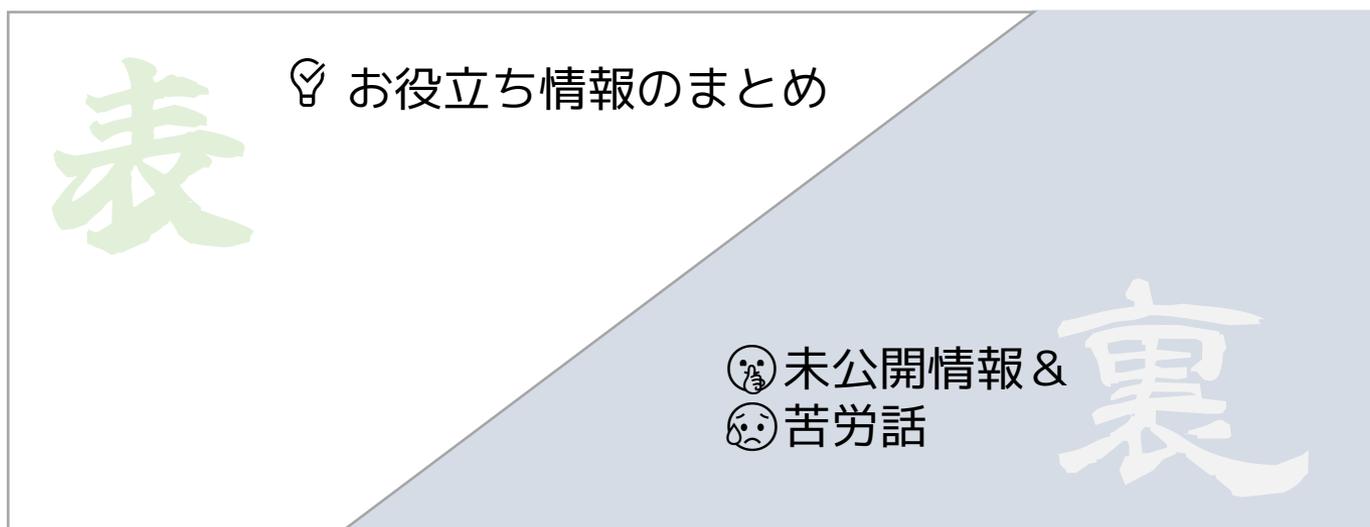
Lv.2 Exp.5

金沢 輝一

(国立情報学研究所;  
CiNii Research データ品質向上チームリーダー)

# 本日のねらい

- CiNii Researchの利用者から寄せられる疑問（謎）にお答えして、情報発見をもっと究めるための情報をご紹介します
  - お役立ち情報のまとめ（ヘビーユーザーには周知の事実かも）
  - 未公開の裏技や、データ整備に日々奮闘する舞台裏の苦労話など



# CiNii Researchの謎, その1

 Degradation Analysis of Pt/Nb-Ti<sub>4</sub> Catalysts with Controlled Arc Plasma-deposition Information)

MA Yongbing, KAJIMA Hirokata, SHIMASAKI Yuta, NAGAKAZUYUKI, KURODA Yoshiyuki, ISHIHARA Akimitsu, MITSU...  
...2-10 wt% Pt was deposited on Magnéli-phase niobium-Ti<sub>4</sub> mounted on glassy carb

DOI DOI

 総合診療のLO/GOインタビュー記録

宮崎 景 2022-04-25

...We have developed a new curriculum based on these t

DOI HANDLE Web Site



検索結果に本文リンクが2つ以上表示されることがあるけれど違いはあるの？  
どれを参照すればいいの？

本文リンクがたくさん!?

# PID(永続的識別子)を活用しよう!

	DOI	HANDLE	Web Site
リンクのURL	<a href="https://doi.org/10.18999/2002446">https://doi.org/10.18999/2002446</a>	<a href="http://hdl.handle.net/2237/0002002446">http://hdl.handle.net/2237/0002002446</a>	<a href="https://nagoya.repo.nii.ac.jp/records/2002446">https://nagoya.repo.nii.ac.jp/records/2002446</a>
	↓	↓	
最終的に表示されるURL	<a href="https://nagoya.repo.nii.ac.jp/records/2002446">https://nagoya.repo.nii.ac.jp/records/2002446</a>		

- DOI, HANDLE等のPID（永続的識別子）はWeb Siteと同じURLに辿り着く
- 💡 Web SiteのURLは変更されることがあるので（例えばhttpがhttpsになるとか、大学・組織の改組・改名に伴ってドメイン名が変更されるなど）、ブックマークや参考文献のポインタとしてはPIDが適している
- CiNii Researchでは、1つのコンテンツのメタデータを複数の情報源から収集していて、全てのURL, PIDを列挙している

# 3つもDOIが並んでいる理由



	DOI	DOI	DOI
リンクのURL	https://doi.org/10.51069/data.jsmbe.19919807. <b><u>v2</u></b>	https://doi.org/10.51069/data.jsmbe.19919807. <b><u>v1</u></b>	https://doi.org/10.51069/data.jsmbe.19919807
			↓
バージョン	v2	v1	最新 (現時点ではv2)

- 💡 参考文献・データのポインタとしてはバージョンが明示されたPIDの方が適している
- CiNii Researchでは、バージョン違いのPIDを一つのコンテンツにまとめて掲載している
- 💡 CiNii Researchでは、最新バージョンのPIDがどれ、といった履歴の可視化は現時点では行っていないが、同じ種類のPIDの中ではなるべく一番左に最新バージョンが並ぶようにしている

# CiNii Researchの謎, その2

## 画像の内容を説明するテキストを利用した画像解析

渡辺 靖彦, 長尾 真 人工知能 13 (1), 66-74, 1998-01-01

<p>In this paper, we propose a new method for image analysis by using the explanation text of image data. First, we describe the method of

DOI “ 被引用文献6件 参考文献11件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 MIRU'96 1996

“ 被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 画像の認識 理解シンポジウム(MIRU'96) 2 271-276, 1996

“ 被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺靖彦 画像の認識 理解シンポジウム (MIRU'96) 2 271-276, 1996

“ 被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 靖彦, 長尾 真 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション  
パターン情報と自然言語情報を相補的に用いると、受け手にわかりやすく正確に作  
でパターン情報と自然言語情報を統合的にとりあつかう必要性が高まっている。本  
を説明している自然言語テキストの情報を利用して画像を解析する方法について立  
いるテキストから対象の色の情報を取り出し、その情報を利用…

“ 被引用文献3件 参考文献9件

## 画像の内容を説明するテキストを利用した画像解析

渡辺靖彦 信学技報 PRU95-138, 1995

“ 被引用文献1件



検索結果に同じタイトルで  
同じ著者の論文が並ぶこと  
があるけど、本当に全部  
違う論文なの？

# CiNii Researchの謎, その2

## 画像の内容を説明するテキストを利用した画像解析

渡辺 靖彦, 長尾 真 人工知能 13 (1), 66-74, 1998-01-01

<p>In this paper, we propose a new method for image analysis by using the explanation text of image data. First, we describe the method of

DOI “被引用文献6件 参考文献11件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 MIRU'96 1996

“被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 画像の認識 理解シンポジウム(MIRU'96) 2 271-276, 1996

“被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺靖彦 画像の認識 理解シンポジウム (MIRU'96) 2 271-276, 1996

“被引用文献1件

## 画像の内容を説明するテキストを利用した画像解析

渡辺 靖彦, 長尾 真 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション  
パターン情報と自然言語情報を相補的に用いると、受け手にわかりやすく正確に作  
成してパターン情報と自然言語情報を統合的にとりあつかう必要性が高まっている。本  
を説明している自然言語テキストの情報を利用して画像を解析する方法について立  
ているテキストから対象の色の情報を取り出し、その情報を利用...

“被引用文献3件 参考文献9件

## 画像の内容を説明するテキストを利用した画像解析

渡辺靖彦 信学技報 PRU95-138, 1995

“被引用文献1件



検索結果に同じタイトルで  
同じ著者の論文が並ぶこと  
があるけど、本当に全部  
違う論文なの？



いいえ、残念ながら一つの  
論文が2回(以上)表示される  
ことがあるんです

# 同じタイトルの論文を比較してみる

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺 靖彦, 長尾 真 人工知能 13 (1), 66-74, 1998-01-01  
<p>In this paper, we propose a new method for image analysis by using the explanation text of image data. First, we describe the method of

DOI “ 被引用文献6件 参考文献11件

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺 MIRU'96 1996  
 “ 被引用文献1件

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺 画像の認識 理解シンポジウム(MIRU'96) 2 271-276, 1996  
 “ 被引用文献1件

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺靖彦 画像の認識 理解シンポジウム (MIRU'96) 2 271-276, 1996  
 “ 被引用文献1件

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺 靖彦, 長尾 真 電子情報通信学会技術研究報告, NLC, 言語理解とコミュニケーション  
 パターン情報と自然言語情報を相補的に用いると、受け手にわかりやすく正確に作  
 成されるパターン情報と自然言語情報を統合的にとりあつかう必要性が高まっている。本  
 報告では、説明している自然言語テキストの情報を利用して画像を解析する方法について述  
 べている。また、対象の色の情報を取り出し、その情報を利用する

“ 被引用文献3件 参考文献9件

**画像の内容を説明するテキストを利用した画像解析**  
 渡辺靖彦 信学技報 PRU95-138, 1995  
 “ 被引用文献1件

発行日	収録誌	著者
1998-01-01	人工知能	渡辺 靖彦, 長尾 真
1996	MIRU'96	渡辺
1996	MIRU'96	渡辺
1996	MIRU'96	渡辺靖彦
1995-10-19	信学技報 NLC	渡辺 靖彦, 長尾 真
1995	信学技報 PRU	渡辺靖彦

# 3件の論文・予稿に分類できそう

発行日	収録誌	著者	
1998-01-01	人工知能	渡辺 靖彦, 長尾 真	原著論文(技術論文)
1996	MIRU'96	渡辺	} 同一の会議発表
1996	MIRU'96	渡辺	
1996	MIRU'96	渡辺靖彦	
1995-10-19	信学技報 NLC	渡辺 靖彦, 長尾 真	} 同一の会議発表？
1995	信学技報 PRU	渡辺靖彦	

1995年の2つは同一？別物？この情報だけでは判断がつかない...

# 証拠を求めてネットを巡る裏

学会のサイトに、2つの研究会による共催の発表会だった記録を発見した  
→ 書誌情報としては2件、予稿としては同一の内容だと分かる

電子情報通信学会 技報コンテンツアーカイブ

[https://www.ieice.org/publications/ken/summary.php?contribution\\_id=KJ00002154688](https://www.ieice.org/publications/ken/summary.php?contribution_id=KJ00002154688) より

資料番号	NLC95-33, <u>PRU95-138</u>
発行日	<u>PRU95-138</u> , 1995
研究会情報	
研究会	NLC
開催期間	<u>1995/10/19</u> (から1日開催)

NLC, 言語理解とコミュニケーション 95 (320), 19-26, 1995-10-19

- ☹ CiNiiは会議の共催情報を持ち合わせていないため、外部の情報で確認していただくことになり不便
- ☺ CiNiiのデータ整備の一つとして、収録号の内容を比較して共催関係を推測することはできるかも

# 情報量が異なる原因を探る

発行日	収録誌	著者	NDL	JaLC	CiNii A.
1998-01-01	人工知能	渡辺 靖彦, 長尾 真	✓	✓	✓
1996	MIRU'96	渡辺			✓
1996	MIRU'96	渡辺			✓
1996	MIRU'96	渡辺靖彦			✓
1995-10-19	信学技報 NLC	渡辺 靖彦, 長尾 真			✓
1995	信学技報 PRU	渡辺靖彦			✓

# 情報量が異なる原因を探る

発行日	収録誌	著者	被引用文献	参考文献	NII論文ID
1998-01-01	人工知能	渡辺 靖彦, 長尾 真	✓	✓	110002808028
1996	MIRU'96	渡辺	✓		10022414215
1996	MIRU'96	渡辺	✓		10010708754
1996	MIRU'96	渡辺靖彦	✓		10018300092
1995-10-19	信学技報 NLC	渡辺 靖彦, 長尾 真	✓	✓	110003278328
1995	信学技報 PRU	渡辺靖彦	✓		10000081001

# 共通点が見えてきた？

発行日	収録誌	著者	被引用文献	参考文献	NII論文ID
1998-01-01	人工知能	渡辺 靖彦, 長尾 真	✓	✓	110002808028
1996	MIRU'96	渡辺	✓		10022414215
1996	MIRU'96	渡辺	✓		10010708754
1996	MIRU'96	渡辺靖彦	✓		10018300092
1995-10-19	信学技報 NLC	渡辺 靖彦, 長尾 真	✓	✓	110003278328
1995	信学技報 PRU	渡辺靖彦	✓		10000081001

# 情報の薄いレコードの正体



- CiNii Articlesだけが情報源の学術記事のレコードには「参考文献として書かれた情報」だけに基づいているもの（引用レコード）が存在する
  - 引用レコードのメタデータは不十分であることが多い
    - **本文へのリンクがない**
    - 筆頭著者名しか記載されていないことがある
    - 収録誌名が略称のことがある
    - 発行日の情報が省略されていたり、年だけのことがある
  - 会議発表の予稿などに引用レコードが多く、削除してしまうと困る分野もあるので削除はしない方針
- 💡 予稿のメタデータの収集への取り組み
- 学術情報流通インフラでの対応
- ▲ 引用レコードは必ずNII論文番号が「1」+10桁だが、「1」+10桁のレコードが全て引用レコードではないので目安程度

# 引用レコード品質向上への取り組み状況

💡引用レコードをもっと適切に表示できればよい

- 引用レコード間の名寄せがもっとできればよいはず

🔧開発途上

- ハードルを上げている要因⇒引用レコードのメタデータは不十分であることが多い
  - 筆頭著者名しか記載されていないことがある
  - 収録誌名が略称のことがある
  - 発行日の情報が省略されていたり、年だけのことがある

# 引用以外の重複レコード



楽曲分析の手法を用いた合奏指導法試論II



河添 達也

島根大学

収録刊行物



島根大学教育臨床総合研究

島根大学教育臨床総合研究 20 2021

関連プロジェクト



現代日本人作曲家の伝統音楽受容による音楽語法研究

河添 達也 島根大学 2019-04-01 - 2022-03-31 (科研費)

詳細情報

CRID :

1010853567245878272

資料種別 :

journal article

データソース種別 :

KAKEN



楽曲分析の手法を用いた合奏指導法試論 II



DOI



河添 達也

収録刊行物



島根大学教育臨床総合研究

島根大学教育臨床総合研究 20 135-147, 2021-08

島根大学教育学部附属教育支援センター

詳細情報

CRID :

1390853650671423744

NII論文ID :

120007150547

DOI :

10.24568/53830

本文言語コード :

ja

データソース種別 :

JaLC

CiNii Articles

# 正体は科研費の成果物

- 科研費研究の成果物情報が名寄せできていないことがある
  - 科研費の報告書にDOIが記載してあれば名寄せされるが、報告書の作成時点で未発行のためDOIを記載できないこともある
  - DOIを記載する様式になる以前の成果物情報も多数存在する
  - 書誌情報に基づいて名寄せをする仕組みがどうしても必要

楽曲分析の手法を用いた合奏指導法試論II

河添 達也  
島根大学

収録刊行物

島根大学教育臨床総合研究  
島根大学教育臨床総合研究 20 2021

関連プロジェクト

現代日本人作曲家の伝統音楽受容による音楽語法研究  
河添 達也 島根大学 2019-04-01 - 2022-03-31 (科研費)

DOIがない

科研費  
報告書由来

楽曲分析の手法を用いた合奏指導法試論 II

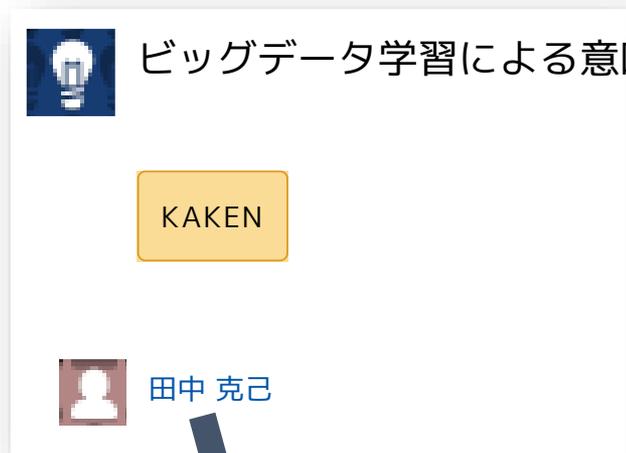
DOI

河添 達也

収録刊行物

島根大学教育臨床総合研究  
島根大学教育臨床総合研究 20 135-147, 2021-08  
島根大学教育学部附属教育支援センター

# CiNii Researchの謎, その3



ビッグデータ学習による意図

KAKEN

田中 克己



プロジェクトの詳細画面では漢字表記の研究者なのに、研究者の詳細画面ではローマ字表記のことがあるようだね



Tanaka Katsumi

KAKEN

関連論文: 3件 関連プロジェクト: 1件

# 新しければよいとは限らない裏

- CiNii Researchの人物詳細情報では**最新の氏名情報**を代表表記として掲載している
  - ☹️最新の情報が英語の論文だった場合、氏名が英語表記だけになってしまう
  - 💡代表表記の選び方をもう少し工夫すれば解消できそう
- 科研費報告書では誤字・異表記が結構あり、KAKENでは多数決で代表を選ぶなどの工夫をしている

藤元 治郎  
藤元 治郎  
藤本 治郎

萩谷 昌己  
萩谷 昌巳  
萩谷 雅己

夏越 祥次  
夏越 祥次  
夏越 祥司  
夏越 祥二

萩原 兼一  
萩原 兼一  
萩原 謙一

- CiNii Researchでは論文の著者名はオリジナルの記載を尊重しているが、イニシャル表記や名字だけの記載には参考情報としてフルネームも掲載してほしい、英語UI画面で漢字表記の氏名にローマ字表記も併記してほしいという要望はある

# CiNii Researchの謎, その4



対馬 かなえ

関連論文: 8件

関連プロジェクト: 1件



同一人物の詳細が複数に分かれているようですが?



対馬 かなえ

KAKEN

関連プロジェクト: 4件

この人物について

| 研究者番号 80754663

| 所属 国立情報学研究所, アーキテク...

## 著者同定が成功している例

 研究 太郎 (000000000)



① 研究成果報告書に記載された  
研究者番号が同一なので  
名寄せする



③ 報告書に記載された  
DOIで名寄せする  
(論文クラスタの中の1件でも  
報告書と名寄せできれば  
クラスタ全体を研究者と  
紐づけることができる)

② CiNiiの論文著者の  
クラスタリング処理で  
同一人物と推定する  
(氏名、共著者、所属等の  
類似性に基づいて機械学習  
している)

# あと一歩で完成するはずなのに...



 対馬 かなえ (80754663)



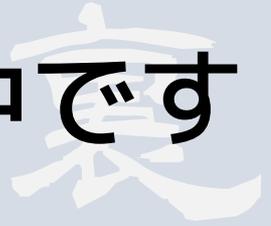
① 研究成果報告書に記載された  
研究者番号が同一なので  
名寄せする

 対馬 かなえ



② CiNiiの論文著者の  
クラスタリング処理で  
同一人物と推定する  
(氏名、共著者、所属等の  
類似性に基づいて機械学習  
している)

③ 報告書に記載された  
DOIで名寄せする  
(研究者番号が記載されていない  
報告書だったため、研究者と  
紐づかない)



# 新しい名寄せを密かに開発中です

 CiNii Researchの論文著者の同定に、研究分野を判断に用いる新しい手法のシステムを開発中

 対馬 かなえ (80754663 )



① 研究成果報告書に記載された研究者番号が同一なので名寄せする



② 著者と同名の研究者の中で、論文と近い分野の人が一人だけいたら同一人物と推定する

(論文と報告書の概要文の分野類似性を深層学習している)



③ 報告書に記載されたDOIで論文と研究課題を名寄せする (報告書に研究者番号の記載がなくても間接的に研究者と紐づけることができる)

# 誤情報を発見されたら フォームからご連絡ください

- 自分の論文ではない、自分の論文が同姓同名の別人のものとして掲載されているといった誤りは、フォームからご連絡ください。訂正いたします。

 金澤, 輝一

関連論文: 18件

この人物について

### 修正のご指摘

名前(必須)

メールアドレス(必須)

指摘対象 金澤, 輝一 (1070292572148442369)

指摘内容 著者情報の修正

詳細(任意)

### 詳細情報

CRID : 1070292572148442369

NII著者ID : 9000004352910  
9000004806742

データソース種別 : IRDB  
CiNii Articles  
NDL

外部サイトでさがす

VIAF  
KAKEN - 研究者を探す  
researchmap  
ORCID  
Google Scholar  
Google  
Yahoo!

### 問題の指摘

著者情報の修正  
その他

# 表に見えないところでも日々成長を続けるCiNii Researchです



- 誤情報を掲載してしまった原因が個別のデータの誤りではなく処理方法にあった場合、CiNii Researchの研究開発チームが専門知識と経験を持ち寄って議論することで、よりよい解決を目指しています



画像はイメージです

開発者インタビューや座談会『研究者、図書館職員らが語る、強みと期待』などCiNii Researchの情報満載の広報誌NII today No.96もお楽しみください!

<https://www.nii.ac.jp/today/96/>

**CiNii Researchの、今後のさらなるデータ品質向上にご期待ください**