# J-STAGEを活用した日本の学術論文データの整備

信州大学　学術研究・産学官連携推進機構/IR室
久保琢也*・伊藤広幸
*kubotaku@shinshu-u.ac.jp

2021年3月22日
大学評価コンソーシアム
継続的改善のためのIR/IEセミナー2021
実務担当者セッション

# 概要

## ❖ 背景

- 文部科学省の国立大学運営費交付金「成果を中心とする実績状況に基づく配分」
- 課題は日本の学術雑誌に掲載される論文データの整備

## ❖ 取り組み内容

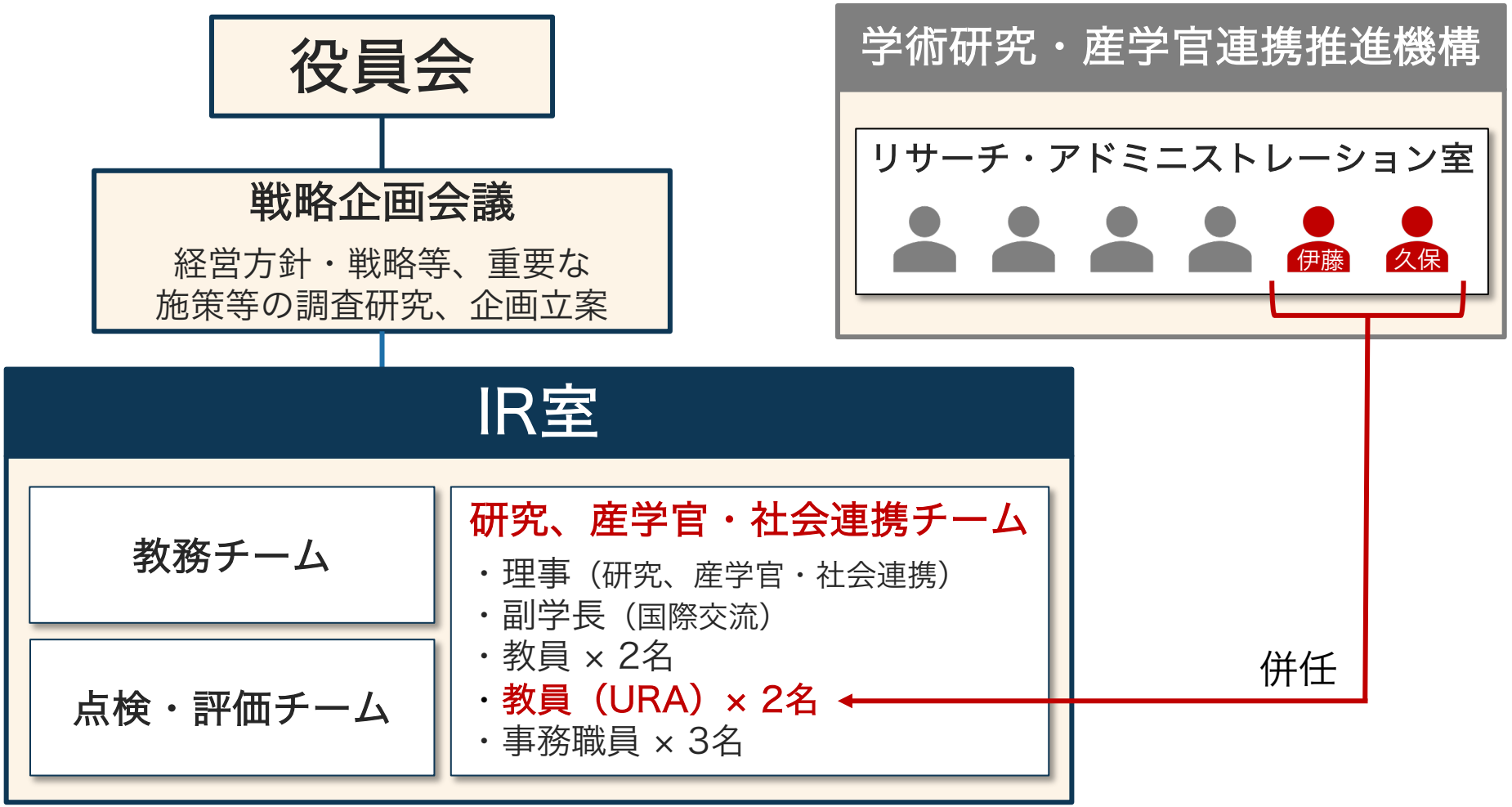
JSTと契約を結び、J-STAGEに採録される学術雑誌の論文データの収集と整備

## *Take-home message*

- J-STAGEを使うことで、日本の学術雑誌の論文データの整備が比較的に容易になる
- 論文タイプや査読の有無等についてはデータの精査が必要である

# 信州大学のIR体制：組織図

## ❖ 信州大学のIR体制



# 信州大学のIR体制：研究IRチームの業務

## ❖ 分析基盤の整備

- データ取集・蓄積・共有
- 独自のツール開発（公開・未公開）

## ❖ 評価対応（研究関係）

- 第３期中期目標・計画期間における法人評価
- 運営費交付金：重点支援の評価
- 運営費交付金：成果を中心とする実績状況に基づく配分

## ❖ 各種調査

- 理事等の執行部の依頼による調査
- URAや部局の依頼によるボトムアップ調査
- 定常的な調査（大学ランキング等）

# 取り組みの背景

# 取組の背景： 文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 「成果を中心とする実績状況に基づく配分」とは

- 運営費交付金の一部を共通する指標に基づく評価によって傾斜配分
- 重点支援の枠組み内での相対評価
  - 重点支援①地域貢献（55大学）
  - 重点支援②特定分野（15大学）
  - 重点支援③世界トップ（16大学）

多額の公費により支えられている国立大学等に対して、厳格な評価とそれに基づく資源配分が求められていることから、国立大学等の法人化のメリットを活かした各大学の成果や実績を相対的に評価することで、一層の経営改革を推進するもの

令和元年度国立大学法人運営費交付金における 新しい評価・資源配分の仕組みについて（成果を中心とする実績状況に基づく配分の仕組みの創設）
https://www.mext.go.jp/content/1417264_001.pdf（2021/3/6 アクセス）

# 取組の背景： 文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 各年度の予算額と配分率

| | R1 | R2 | R3（案） |
|---|---|---|---|
| 予算額 | 700億円 | 850億円 | 1000億円 |
| 配分率 | 90%～110% | 85%～115% | 80%～120% |

予算額と配分率の幅は増加傾向にあり、国立大学等の経営に与える影響は大きくなっている

# 取組の背景： 文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 評価項目の一覧

| 分類 | 評価項目 |
|---|---|
| 教育 | 卒業・修了者の就職・進学等の状況 |
| 教育 | 博士号授与の状況 |
| 教育 | カリキュラム編成上の工夫の状況 |
| 研究/経営 | 若手研究者比率 |
| 研究 | 交付金等コスト当たりTop10%論文数<br>（※重点支援３を選択した大学のみ） |
| 研究 | 常勤教員当たり研究業績数 |
| 研究 | 常勤教員当たり科研費獲得額・件数 |
| 経営/研究 | 常勤教員当たり受託・共同研究 受け入れ額 |
| 経営 | 人事給与マネジメント改革状況 |
| 経営 | ダイバーシティ環境醸成の状況 |
| 経営 | 会計マネジメント改革状況 |
| 経営 | 寄付金等の経営資金獲得実績 |
| 経営 | 施設マネジメント改革状況 |

ここの話をします

# 取組の背景：　文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 常勤教員当たり研究業績数（R3）

2017年度～2019年度における学系別の研究業績数（査読付き論文、学術図書、作品等）

**【学系】**

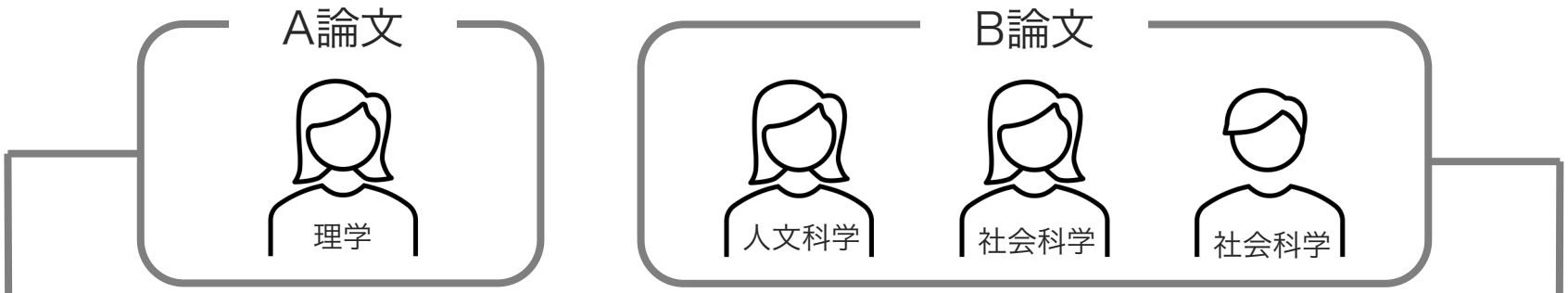
国立大学法人評価の教育研究に関する評価における「現況分析」での評価単位。部局別に予め指定。（人文科学系、社会科学系、理学系、工学系、農学系、保健系、教育系、総合文系、総合理系、総合融合系）

**【査読付き論文のカウント方法】**

- ある学系に所属する常勤教員が単独で書いた論文は当該学系において「１」と集計
- 複数の常勤教員が著者として含まれる場合、１編の論文について各学系ごとに「１」を上限として集計

# 取組の背景：　文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 例えば・・・




| | 人文科学 | 社会科学 | 理学 | … |
|---|---|---|---|---|
| A論文 | 0 | 0 | 1 | … |
| B論文 | 1 | 1 | 0 | … |
| … | … | … | … | … |
| Z論文 | 0 | 1 | 0 | … |
| 合計 | ○○ | ○○ | ○○ | ○○ |

# 取組の背景： 文部科学省 国立大学運営費交付金 成果を中心とする実績状況に基づく配分

## ❖ 必要なデータの形式の例

| 論文ID | 研究者ID | 所属部局 | 学系 |
|---|---|---|---|
| A論文 | a研究者 | 文学部 | 人文科学系 |
| A論文 | b研究者 | 教育学部 | 社会科学系 |
| B論文 | c研究者 | 理学部 | 理学系 |
| C論文 | d研究者 | 工学部 | 工学系 |
| C論文 | e研究者 | 工学部 | 工学系 |
| C論文 | f研究者 | 医学部 | 保健系 |
| … | … | … | … |

論文と該当する研究者の紐付けが大事

# 課題

# 課題：日本の学術雑誌に掲載される論文

## ❖ 国際学術文献データベース

Web of Science　　Scopus

- 上記のデータベースから国際的な学術雑誌に掲載される論文情報は（ある程度）網羅的に収集可能
- Web of Scienceでダウンロード可能なデータから、（ある程度）著者名と著者の所属の紐付けが可能（確かScopusも）

一般的に、研究IRでは国際学術文献DBを用いることが多く、「成果を中心とする実績状況に基づく配分」でも国際学術誌の論文情報は上記のDBが活用できる。

**一方、日本の学術雑誌の論文は？**

# 課題：日本の学術雑誌に掲載される論文

## ❖ 日本の学術文献データベース

| データ取得方法 | CiNii Article | J-STAGE |
|---|---|---|
| ダウンロード機能 | 有 | 無 |
| APIの提供 | 有 | 有<br>（※24h以内の利用制限） |

- 信州大学の研究者が関与する論文の書誌情報は取得できるが、どの著者が信州大学に所属するかわからない
- 著者の所属機関情報も取得するためには１件１件アクセスして確認するしかない

# 課題：日本の学術雑誌に掲載される論文

## ❖ 信州大学 学術情報オンラインシステム（SOAR）

- インプット
  - 基本的に教員による手入力
  - Web of Sciene、CiNii Article からインポートも可能
- 用途
  - 研究者総覧
  - 個人業績評価


- 教員による入力作業を必要とするため、データの網羅性に不安が残る
- 手入力による表記揺れのため、教員の共著関係を正しく把握できない

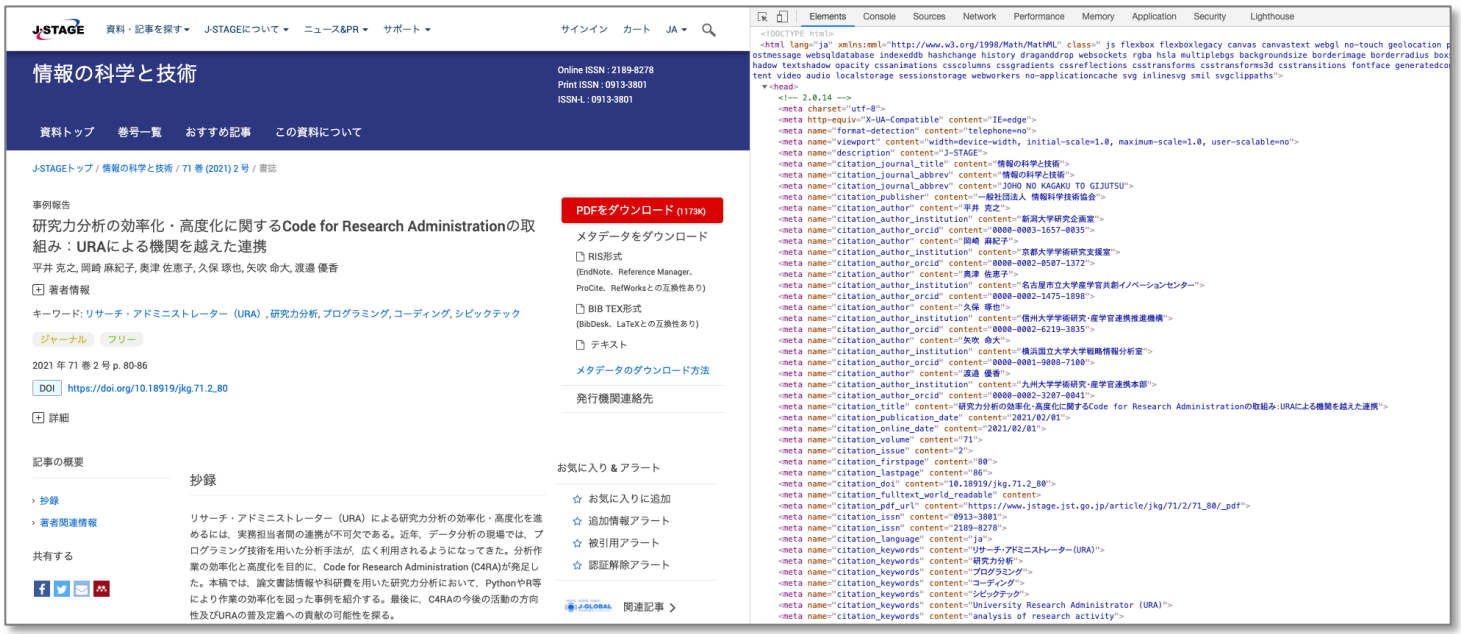
# 取り組みの内容

# 取組の内容：J-STAGEからデータを収集

## ❖ 方法

機械的に該当する論文のページにアクセスして著者名や所属機関等の情報を取得する（クローリング）

Webブラウザで表示　　HTML文書




```
<meta name="authors" content="久保 琢也">
<meta name="authors_institutions" content="信州大学学術研究・産学官連携推進機構">
<meta name="authors_orcids" content="0000-0002-6219-3835">
```

# 取組の内容：J-STAGEからデータを収集

**※ 注意 ※**

J-STAGEから機械的にデータを取得することは
利用規約上、禁止されている

- **JSTにJ-STAGEの利用の許可をもらう**
  - 2020年6月ごろ、JSTの担当者の方に相談

    ～（契約書の修正等）～
  - 2020年10月ごろ、契約締結

※2020年11月初旬「成果を中心とする実績状況に基づく配分」のためのデータ提出〆

# 取組の内容：J-STAGEからデータを収集

## ❖ データ取得の流れ

① APIにより、信州大学の研究者の2017年から2019年の論文情報（DOI、リンク等）を取得（約1200件）

※大学名のバリエーションに注意（信州大学、Shinshu University、信大、信州大）

② 上記のリンクを使ってクローリング

- R言語によりクローラー作成
- アクセス間隔を十分にとっても１日程度でデータの取得は可能

※サーバーに過度な負荷を与えないように注意

③ 査読あり雑誌名を事前に取得しておくことで、査読あり雑誌の論文だけを抽出することが可能

# 取組の内容：J-STAGEからデータを収集

## ❖ 著者情報の所在

– <head>タグ内の<meta>タグに論文の書誌情報や著者情報が整理されている

著者①

```
<meta  name = “authors”  content = “信大 一郎”>
<meta  name = “authors_institutions”  content = “信州大学 社会学部”>
```

著者②

```
<meta  name = “authors”  content = “信大 花子”>
<meta  name = “authors_institutions”  content = “信州大学 社会学部”>
<meta  name = “authors_institutions”  content = “社会科学研究所”>
```

※１名の著者に対して複数の所属情報がある場合があるので注意

– ソースの確認方法（Google Chrome）
「表示」⇒「開発/管理」⇒「デベロッパーツール」

# 取組の内容：J-STAGEからデータを収集

## ❖ 著者情報の整理方法①

① name属性の値が「authors」または「authors_institutions」のmetaタグを抽出し、name属性の値とcontent属性の値からなるデータフレームを作成

② name属性が「authors」の時は「1」、「authors_institutions」の時は「0」となる著者名フラグ列を作成

③ 著者名フラグ列の累積和であるID列を作成

| | ① | | ② | ③ |
|---|---|---|---|---|
| | name属性 | content属性 | 著者名フラグ | ID（著者名フラグの累積和） |
| 著者① | authors | 信大 一郎 | 1 | 1 |
| 著者① | authors_institutions | 信州大学 社会学部 | 0 | 1 |
| 著者② | authors | 信大 花子 | 1 | 2 |
| 著者② | authors_institutions | 信州大学 社会学部 | 0 | 2 |
| 著者② | authors_institutions | 社会科学研究所 | 0 | 2 |

# 取組の内容：J-STAGEからデータを収集

## ❖ 著者情報の整理方法②

④　name属性の値（authors/authors_institutions）によって２つのデータフレームに分割する。

authorsのデータフレーム

| content属性 | ID |
|---|---|
| 信大 一郎 | 1 |
| 信大 花子 | 2 |

authors_institutionsのデータフレーム

| content属性 | ID |
|---|---|
| 信州大学 社会学部 | 1 |
| 信州大学 社会学部 | 2 |
| 社会科学研究所 | 2 |

⑤　ID列をキーに２つのデータフレームを結合する

| ID | 氏名 | 所属機関 |
|---|---|---|
| 1 | 信大 一郎 | 信州大学 社会学部 |
| 2 | 信大 花子 | 信州大学 社会学部 |
| 2 | 信大 花子 | 社会科学研究所 |

# 取組の内容：J-STAGEからデータを収集

## ❖ 文献タイプの取得


コレ

- 文献タイプ（論文、事例紹介、書評、等）のあるページがあるので、合わせて取得しておくと便利
  - <body> … 

# 取組の内容：J-STAGEからデータを収集

## ❖ J-STAGEの論文データの注意点

### ① 文献タイプ

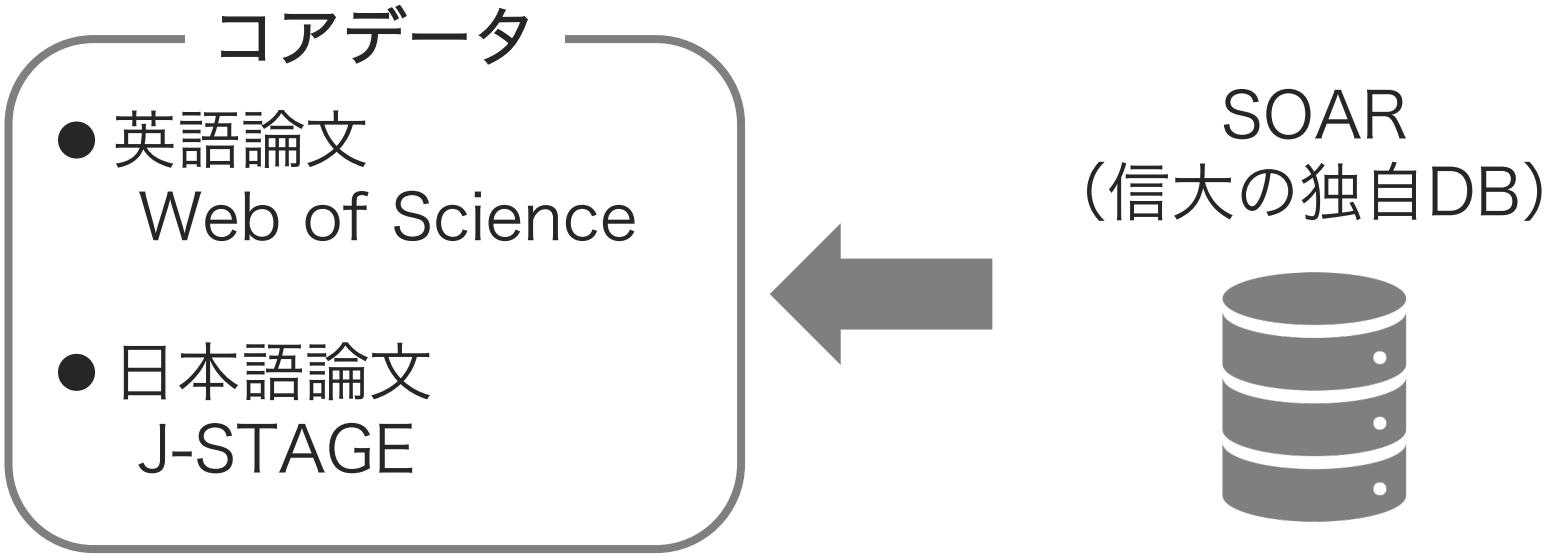
- J-STAGEには様々な文献タイプが混在するため、該当しない文献は削除する必要がある（巻頭言、編集後記、etc.）
- 機械的に文献タイプを取得できない場合は１つ１つ確認するしかない

### ② 査読の有無

- 雑誌の査読の有無は機械的に判断することができるが、１つ１つの論文が査読を経たものであるかは、やはり、それぞれ確認しないといけない

# 取組の内容：データの整理

## ❖ データの統合




- Web of Science、J-STAGEの論文データは重複することがあるため、DOIをIDとして重複を削除
- SOARから重複しない論文データを統合

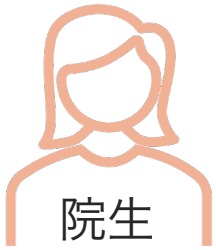
# 取組の内容：データの整理

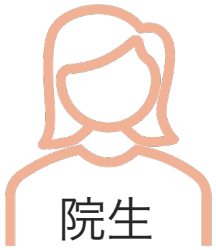
- **著者名の名寄せと常勤教員の抽出**

| 論文ID | 研究者 | | 所属部局 | 学系 |
|---|---|---|---|---|
| A論文 | a研究者 | ↔ | 文学部 | 人文科学系 |
| A論文 | b研究者 | ↔ | 教育学部 | 社会科学系 |
| B論文 | c研究者 | ↔ | 理学部 | 理学系 |
| C論文 | d研究者 | ↔ | 工学部 | 工学系 |
| C論文 | e研究者 | ↔ | 工学部 | 工学系 |
| C論文 | f研究者 | ↔ | 医学部 | 保健系 |
| … | … | ↔ | … | … |

# 取組の内容：データ整備体制

## ❖ アルバイトさんの力は偉大




- 著者名の名寄せ
- 論文種別
- 査読チェック
- 重複チェック

※ 学術図書に関する作業も



- 著者名の名寄せ
- 論文種別
- 査読チェック
- 重複チェック

＋ URA業務

- データの前処理
- データの取りまとめ

頼れるアルバイトさんと出会えるかどうかの差は大きい！

# おわりに

- **所感**
  - 全てを自動化できたわけではないが、日本の学術論文データの整備にかかる負荷はかなり軽減される
  - しかし、それなりの手作業が発生するので、信大クラスの規模（教員数1000+）だからできるのかもしれない
- **今後の発展性と課題**
  - 他大学との比較によるパフォーマンス分析（特に、人文社会科学系）
  - 評価・IR以外にも活用方法の模索（広報、その他）
  - 定常的な業務としてデータを整備できるような仕組みづくり

# 謝辞

本取り組みを実施するにあたり、ご協力いただいた科学技術振興機構の方々にこの場を借りて御礼申し上げます。