情報処理学会論文誌　デジタルプラクティス　Vol.2　No.2　47–56 (Apr. 2021)

情報学研究データリポジトリ IDR における研究用データセット共同利用の取り組み

大須賀　智子[1,a)]　大山　敬三[1,2]

受付日 2020年8月25日, 再受付日 2020年11月9日,
採録日 2020年12月21日

概要：データサイエンス研究の進展には，現実性のある十分な規模のデータを多くの研究者が共通に利用できることが不可欠であるが，多くの領域ではそのようなデータの確保に種々の課題がある．国立情報学研究所の情報学研究データリポジトリ（IDR）では，情報学および関連分野に資するため，産学界等と大学等の研究者とを媒介し，データセットの共同利用に取り組んでいる．本稿ではまず，データセット共同利用の意義について述べる．次に，IDR におけるデータセットの取り扱いに関して，企業等からの受け入れ，研究者への提供，提供後の利用者の管理と利用状況の把握，およびデータ DOI の付与について各処理内容を説明する．続いてデータセット提供実績および利用者による研究成果の状況，並びに研究者コミュニティに対する活動支援の取り組みについて示す．筆者らは本活動を通してデータセット共同利用の実現における課題を発掘し理解を深化させ，対応策を案出して実践しており，データセットの受け入れにおける調整事項や利用上の制限事項などはこれらの知見を反映させたものとなっている．

キーワード：大規模データセット，共同利用，オープンサイエンス，データリポジトリ

Sharing Datasets for Informatics Research through Informatics Research Data Repository (IDR)

Tomoko Ohsuga[1,a)]　Keizo Oyama[1,2]

Received: August 25, 2020, Revised: November 9, 2020,
Accepted: December 21, 2020

Abstract: It is indispensable for the development of data science research that realistic and sufficiently large scale data are commonly available for many researchers. In many research area, however, there are variety of issues to be overcome for acquiring such data. The National Institute of Informatics (NII) is operating Informatics Research Data Repository Service (IDR) for promoting dataset sharing and collaboration by mediating industries and academic researchers in order for supporting informatics and the related research field. In this paper, the authors first describe the significance of dataset sharing, next explain the process in each aspect of dataset handling, i.e. acceptance from companies, provision to researchers, management of users and grasp of usage after provision, and assignment of DOI. They then show the statistics of dataset provision and research achievements made by the uses, and introduce the activities for supporting researcher community. Through these activities, they found various issues to solve for realizing dataset sharing, deepened their understanding on those issues, and devised and practiced the countermeasures, and their findings have been reflected on the negotiation issues at the dataset acceptance and on the rules of dataset usage.

Keywords: large scale datasets, shared use, open sciences, data repository

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] 総合研究大学院大学
The Graduate University for Advanced Studies, SOKENDAI, Kanagawa 240–0193, Japan
[a)] osuga@nii.ac.jp

1. はじめに

近年のデータサイエンスの進展には，ディープラーニングに代表される統計的学習手法等の人工知能（AI）技術やそのための計算環境等が整ってきたことに加えて，大量

のデータが利用可能になってきたことが背景にある．最近では最先端の研究成果を取り入れたツールが次々と公開され，また高性能の GPU を搭載したサーバが一層安価に利用できるようになっているため，様々な領域におけるデータ活用の広がりが一層期待されるところである．

しかし，データに関しては，現実性を有しかつ十分な規模があるものを，多くの研究者が共通して利用できるようになっている領域は一部に限られる．そのような領域が拡大しない主要因としては，データ保有者にとって，データを提供することによるメリットが明確でなく，またリスクの管理が困難なこと，データ提供の準備にかかる技術的および事務的コストが負担となること，データ提供後に利用者に対応する組織がないこと，また民間企業においてはこれらの理由から経営者層の理解が得られないこと，などが挙げられる．

このような状況を受けて，情報学分野の大学共同利用機関である国立情報学研究所（NII）では，「情報学研究データリポジトリ」（IDR）の事業において，情報学研究に必要な各種データを民間企業や大学研究者等から受け入れ，適切な権利処理を施し，利用者に対する一元的な窓口となって，大学を中心に多くの研究者に共通のデータセットを提供することにより，データセットの共同利用[*1]に取り組んでいる．

2. データセット共同利用の意義

大学などには実用的な研究成果を要請する声が高まっている．しかし，研究者が実験用として構築したデータには，量的に充分か，あるいは現実を適切に反映しているかなどの問題がつきまとう．また，民間企業等と契約を交わしてデータ提供を受ける場合でも，研究の透明性や再現性などが大きな問題となる．このため，実社会で生成された大規模データや多くの研究者が協力して構築したデータを研究資源として研究コミュニティが共有できるようにすることが重要になる．

一方で，大規模データを取り扱う民間企業，とくにインターネット上で事業展開する企業では，先進技術をいち早く事業に取り入れることが重要であるが，社内に十分な研究開発能力を備えているところは多いとはいえず，保有する大規模データを十分に活用できていないのが現状である．このため，大学との共同研究などを通じて技術開発や若手の人材確保などを図るため，大学などにデータを提供しようとするインセンティブが働いている [1]．ただし企業にとっては，たとえ学術研究用途といえども，業務用システムからデータを抽出して個人情報の秘匿化など必要な加工を施し，機密保持や知的財産処理に関する契約の交渉を個々に行うなど，データの提供にあたっては多大な労力を要する．また，このようなデータは保有する企業の経済的利益にも関わるものであり，広くオープンにすることはできない．著作権や個人情報などに係わるデータも，やはりオープンにすることは難しい．これは実験や観測データのオープンデータ化を進めている自然科学の諸分野とは対照的な点である．

また，大学等の研究者等が構築したデータにおいても，たとえば自然言語処理の分野でテキストデータに多大な労力をかけてアノテーションを施したコーパスや，音声情報処理やコミュニケーション研究の分野で多くのコストをかけて収録した音声や映像のコーパスなど，著作権や個人情報などに係わるなどの理由でオープンにすることが難しいデータがある．

そこで IDR では，情報学に関連する研究に資する各種データを保有者から受け入れ，より多くの研究者に提供できるようにするため，一元的な窓口としての役割を担っている．第 5 期科学技術基本計画にオープンサイエンスの推進が明記され，特に公的研究資金を用いた研究成果は，論文だけでなく研究データのオープン化も求められているが，但し書きとして，「商業目的で収集されたデータなどは公開適用対象外とする」こと，「データへのアクセスやデータの利用には，個人のプライバシー保護，財産的価値のある成果物の保護の観点から制限事項を設ける」ことが記されている [2]．IDR におけるデータセット共同利用は，このようなオープン化が難しいデータを，適正な管理の下で，契約に基づき利用可能とするための取り組みであり，データの提供先や利用目的は制限をしながらも，データに関する情報や一定条件下での利用機会についてオープン化を目指すものである．

大学等の研究者にとっては，実社会のデータや実用性の高いデータを使用できるだけでなく，使用したデータセットが特定可能となることにより，研究の透明性・再現性が担保され，他の研究との比較も容易となる．第三者の権利侵害などのおそれもなくなる．データの収集や前処理が不要になることで，研究に取り掛かるまでの労力が大幅に軽減される．

データ提供者の観点からは，最初に提供機関内（民間企業の場合は経営者や事業部門など）との調整やデータの準備にコストがかかるのはやむを得ないが，その後はほとんど労力をかけることなく幅広い研究者にデータを活用してもらえるようになる．また特に民間企業にとっては，当該分野の研究者や学生に対して社会貢献の周知やオープン性・公平性のアピールを図れるとともに，研究成果のフィードバック，将来の共同研究や人材確保の可能性が期待できる．

IDR 以外にこのようなデータ提供の窓口となる組織と

[*1] 個々の研究者や大学などでは整備できない研究資源を構築して大学などの研究者に提供することをいう．

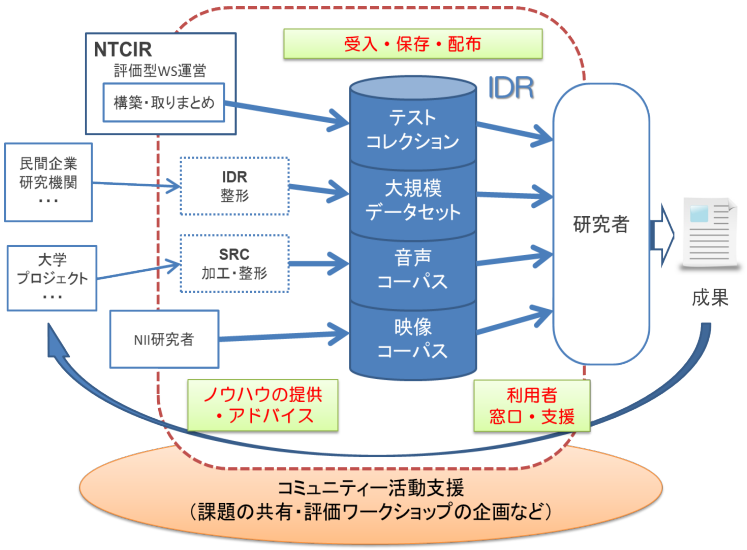

図 1　データセット提供に関わる活動
Fig. 1　Activities related to datasets provision.

しては，国内では言語資源協会（GSK）[*2] が主にテキストコーパスや辞書データを，高度言語情報融合フォーラム（ALAGIN）[*3] が主に情報通信研究機構により構築された各種言語資源を取り扱っている．海外では米国の LDC[*4] や欧州の ELRA[*5] が言語資源を大規模に収集・提供しており，Microsoft[*6] のように民間企業でも研究用に作成したデータセットを公開している例はあるが，IDR のように，特に民間企業から実サービスにより生成されたデータを受け入れ，無償で提供している組織はほとんど類を見ない．このような背景やこれまでの提供実績から，IDR では，新規企業よりデータ提供の申し出を受けることが増加し，それにより利用者層がさらに拡大するという好循環を生んでいる．

3.　IDR のデータセット提供活動におけるプラクティス

3.1　IDR の活動

IDR では，データの保有者からデータセットを受け入れ，保存・管理し，希望する研究者に配布するという基本的な活動に加え，データセット提供者も巻き込んだ研究コミュニティの構築と活性化に努めている．**図 1** に，データセットの提供に関する IDR の主な活動の概念を示す．

現在は，民間企業等からのデータセット受け入れに加え，「音声資源コンソーシアム」（SRC）[*7] にて受け入れた音声コーパスの配布窓口も担っている．また NII では，前身時代の 1997 年より「NTCIR プロジェクト」（NTCIR）[*8] を推進し，評価フォーラムを通じて情報アクセス技術評価用テストコレクションを構築しているが，過去の NTCIR プロジェクトにて構築されたテストコレクションの研究者への配布窓口も順次 IDR へ移管している．民間企業のデータセットについては 2010 年度以降毎年新規データの提供を開始しており，NTCIR テストコレクションや音声コーパスも順調に取り扱い数を増やすとともに，オンライン申請の仕組みも整えるなど，配布の効率化も図っている．

次節以降では，データセットの提供活動における，受け入れ，提供，提供後の各段階での具体的なプラクティスとそれらを通して得られた知見を述べる．IDR では 10 年以上にわたり各種データセットを様々な研究者に提供しており，その中で改善を重ねたこれらのプラクティスを実践することにより，これまでデータ提供者を巻き込むようなトラブルは生じていない．

3.2　データセットの受け入れ

データセットの受け入れにあたり，まずデータセットを構成するコンテンツ等について，サービス規約等の調査や提供者からの聞き取りにより，各種の項目の詳細な確認を行い，項目ごとに生じうるリスクを提供者と共有する．**表 1** には，これまでの知見に基づき，多くのデータセットに共通する代表的な確認項目の例を示す．これに基づき，研究資源としての利用価値にも配慮して，具体的なデータ項目や加工方法について協議を行う．条件が整わない場合，受け入れを断念することもあるが，将来の受け入れに向けて条件整備の提案を行い，中には受け入れに至ることもある．

[*2] https://www.gsk.or.jp/
[*3] https://www.alagin.jp/
[*4] https://www.ldc.upenn.edu/
[*5] http://www.elra.info/
[*6] https://msropendata.com/
[*7] http://research.nii.ac.jp/src/
[*8] http://research.nii.ac.jp/ntcir/

表 1　データセット受け入れ時のコンテンツ等に関する確認項目の例
Table 1　Check item samples on contents at dataset acquisition.

(1)　著作権等知的財産権の権利処理 ・　学術研究目的での第三者提供のために必要な著作権処理が行われているか. ・　学術研究利用に必要な複製や翻案・改変等に制約がないか.
(2)　個人情報保護への対応 ・　明らかな個人情報や，個人情報になりうる情報があるか. ・　個人情報の取得にあたっては学術研究目的での第三者提供に同意を得ているか.
(3)　第三者の権利の侵害への対策 ・　第三者の権利を侵害するデータの投稿を禁止しているか. ・　権利侵害の通報の受付および対応のための体制が整備されているか. ・　自主的に権利侵害を検知・削除する体制があるか．網羅性はどの程度か.

表 2　事業への影響等について考慮すべきデータセットの利用に関する事項
Table 2　Check item samples on dataset usage potentially affecting business operation.

(1)　利用者 ・　所属組織の種別（大学，公的研究機関，民間研究機関，民間企業，等） ・　構成員の種別（教員，学生，社会人学生，共同研究員，等） ・　競合企業との関係性
(2)　利用方法・利用内容 ・　論文や口頭発表（データの公表条件，クレジット表示，事前確認の要否，等） ・　デモシステムでの利用（利用者の制限，表示内容，等） ・　実習・演習等での教育的利用（対象者，人数，誓約書の有無，等）

次に，データセットの利用について，利用者や利用方法・利用内容の場合ごとに，提供者の事業への影響等に関する懸念を考慮し，一方で期待されるメリットにも理解を得つつ，提供対象者および利用目的の範囲や利用制限事項などの提供条件について協議を行う．**表 2** に，これまでの知見に基づき，ほぼすべての民間企業の提供者に共通する，考慮すべき代表的な事項の例を示す．これらの協議結果に基づき，3.3 節に述べるデータセットの提供手続きの方針を定め，親契約を締結する．

提供者ごとに状況や考え方が異なり，これまでもほぼ毎回新たな課題が出現したが，知的財産権の権利処理に問題がある場合を除いては，提供条件と 3.4 節に述べる利用者管理方法の調整により，ほぼ対応することができた．

このような経験を通して蓄積したノウハウに基づき，現在では新たなデータセットの受け入れに際して多面的なアドバイスが可能となり，社内の法務など関連する部門との調整や契約手続きなどを円滑に行えるようになっている．

また実際に配布するデータセットの仕様や配布形態などについても，利用者（研究者）の立場に立ってアドバイスを行っている．リスクを低減するために過度にデータを加工してしまうと可能な研究テーマが限られることもあるので，適度なバランスを見極めることは，難しいが重要なポイントとなる．

3.3　データセットの提供手続き

IDR で取り扱うデータセットは，これまで述べてきたように主としてオープンデータとすることが困難なものであるので，研究者へは利用契約の締結後に提供することになるが，特に提供者が民間企業の場合，配布先や利用にあたって様々な条件を課されることになる．以下に，提供時の条件等について主なものを述べる．

§ 提供対象者

大半のデータセットは，提供対象者が大学や公的研究機関の研究者に限定され，研究室単位で配布を行っている．一方，民間の研究者でも利用可能な一部のデータセット（企業提供のデータでは現時点で 4 種類）については個人単位で配布を行っており，書類のやりとりを簡略化して，利用者がオンラインで利用規約に同意するという形をとっている．

利用契約の締結形態は，NII が提供者からサブライセンスを受けている場合は NII と利用者との間の契約（覚書の締結もしくは同意書の提出），そうでない場合は提供者と利用者との間の直接契約の形となっている．利用契約形態の選択は，データセットの性質や提供者（特に法務部門）の方針によるが，データセットの研究利用（特に不正または予期しない利用）が提供者の事業に与える影響や，データセットに潜在する問題が IDR の活動に与える影響を総合的に評価し，提供者と NII の協議により決定している．

§ 利用者の範囲

研究室単位で配布しているデータセットでは，利用申請の段階で，研究代表者の身分や研究実績の有無，研究室内の利用予定メンバーの身分等を確認し，利用者としての適格性を審査するとともに，データセットによっては，申請者の研究室における民間企業所属者の有無や，民間企業との密接な連携（共同研究等）の有無等の確認も行う．

一方，近年は大学等の組織や所属形態も多様化しているため，研究代表者や研究グループ構成員の身分の呼称も様々あるが，IDR が窓口となることにより，提供者が大学等の事情に詳しくない場合でも，実情に合った適切な形での利用契約となるよう支援している．

§ データの使用目的

原則として学術研究目的での利用に制限されており，利用申請書において具体的な使用目的を確認している．大学での研究においても，検索手法の研究等システム開発を伴うことがあるが，たとえば検索結果として提供データを直接引用するようなシステムの場合，データの第三者提供にあたる可能性がありウェブ等での公開は不可である旨，申請時点で了解を得るようにしている．これは実際に学生が作成したシステムが公開された事例が 1 件あったことを受け追加した対応である．

なお，民間の研究者にも提供可能なデータセットには，直接データを販売するのでなければ研究開発目的に利用可能なものもある．

§ データの利用制限

全データセットに共通した利用条件として，第三者への提供や商業利用の禁止，研究発表等での個人や組織の特定につながる情報の開示の禁止が課されている．さらに，サービス利用者による投稿データが含まれるデータセットでは，プライバシー侵害や公序良俗違反への懸念の程度に応じて，インターネット上の情報などとの照合を禁止するものや，研究発表内容の事前確認を求めるものがある．

教育利用の可否に対する考え方も提供者によるため，大学の授業やゼミでの利用希望があった場合，その規模や対象者，データへのアクセス制限等の管理体制を確認したうえで，提供者との調整を行っている．

また，利用申請時に提出された利用申請書の内容に，不明な点や利用条件に抵触する可能性のある点があれば，必要に応じて利用者，提供者，あるいはその両者に個別に確認を取るなどして，提供者が安心してデータを提供できるようにすることも重要である．

以上のように，データセットの提供に際しては様々な制約があるが，データセットごとに注意が必要な点を分かりやすく提示するとともに，これらの制約が企業の立場やデータの性格等の明確な根拠に基づくものであることを説明し理解を求めることにより，これまで，研究者からは利用条件について問題点の指摘を受けたことはなく，また大きく違反する事例も発生していない．ただし，知的財産権の帰属については，契約文書上の表記が大学等のポリシーに適合しないとの理由で契約に至らなかった事例があり，この知見に基づき，以降のデータセットの受け入れでは，大学側の事情を考慮した表記にするようアドバイスを行っている．

なお，データセットの提供は基本的にはオンラインでのダウンロード形式としているが，ダウンロードページの URL を利用者ごとに生成し，アクセス用 ID・パスワードについては郵便にて所属機関宛に別送するなど，研究代表者の実在性の確認やデータセットの漏洩に対しては利用契約での縛りに加えてデータ提供者が許容可能なレベルの対策も実施している．

3.4　データセット利用状況の管理

IDR では，毎年度末に利用者に対し利用報告書の提出を求め，翌年度の継続利用の有無や利用申請内容の変更の有無を確認し，必要に応じて再契約や利用停止の手続きを行っている．多くのデータセットでは利用契約の有効期間は 1 年間（自動更新）に設定されており，当初はデータ提供日を起点としたり年末を区切りとしたりしていたが，これまでの経験から，利用報告の収集は研究室の学生の入れ替わりや研究者の異動が多い年度の区切りに合わせて実施することが効率的であると分かり，近年提供を開始したデータセットでは原則として利用期間の区切りを年度末としている．

なお，利用報告において提供データセットを用いた研究成果として外部発表したものがあれば書誌情報を記載してもらい，それらの論文リストはデータセット提供者にフィードバックするとともに，「DSC リファレンスポータル」（**図 2**，http://dsc.repo.nii.ac.jp/）で一般公開している．これは NII で開発しているリポジトリモジュール WEKO を活用したもので，使用したデータセットをインデックスとして分類し，データセットごとの研究成果を容易に一覧できるほか，論文誌の種類や発表年，著者名などによる検索も可能である．このリストは，我々の活動の成果としてのエビデンスとなるだけでなく，新たに既存データセットの利用を検討している研究者，新規にデータセットの提供を検討している提供希望者にとっても参考事例となり得るものである．

利用を停止する場合には，関連データの消去について書面で確認するなどの対応を取っている．なお文部科学省が定めた研究不正対策のガイドラインに対し日本学術会議が出した指針 [3] において，研究成果のもととなった実験データ等の研究資料は，論文等の発表から原則 10 年間の保存が示されている．IDR が扱うライセンスつきのデータはその対象から除外されてはいるものの，その趣旨を尊重し，データセット提供終了後も，利用者に代わり NII に

図 2　提供データセットを利用した研究成果のリストを公開するリポジトリの画面サンプル

Fig. 2　Screenshot of the DSC Reference Portal that shows research articles using the provided datasets.

てデータを 10 年間保管し，研究成果について不正の有無を検証する目的に利用できることについて，データセット受け入れ時の提供者との親契約において承諾を得るように努めている．

3.5　データ DOI の付与

2018 年 11 月より，図 2 のリポジトリに提供データセットのメタデータを登録し，ジャパンリンクセンター（JaLC）[*9] を通してデータ DOI の付与を行っている．データ利用者には成果論文に使用したデータセットの DOI を引用してもらうことで，論文の読者にデータセットへの恒久的なアクセスを保証している．また，以前は論文の謝辞欄で提供者名の記載を促していたが，参考文献欄にデータ DOI を記載してもらうことで，NII で開発中の検索基盤（CiNii Research）[*10] により，データセットの作成者や提供者の貢献が可視化されるとともに，他の論文等と有機的に結び付けられ，将来的により効果的な知の循環を生むことを期待している．

データに DOI を付与する際はその粒度が重要な検討課題となる．IDR では，論文等での引用時の効率を優先し，原則としてデータセットの提供単位（利用契約の単位）で

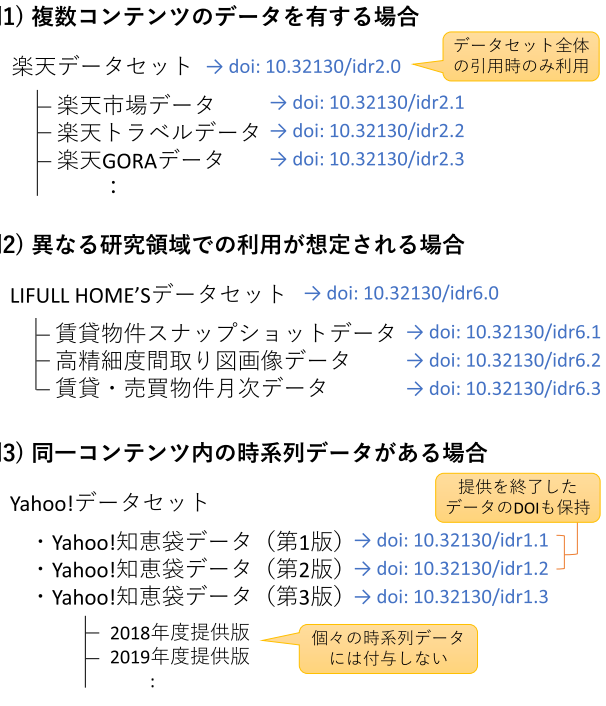

図 3　データ DOI の付与単位の例

Fig. 3　Example of data DOI assignment unit.

DOI を付与しているが，いくつか検討を要したものについて具体例を**図 3** に示す．

例 1 の楽天データセットでは複数の異なるサービスから

[*9] https://japanlinkcenter.org/
[*10] https://rcos.nii.ac.jp/service/research/

取得されたコンテンツが含まれているが，通常，異なるサービスのコンテンツを同時に利用することは想定されないので，サービスごとのデータを単位としてDOIを付与している．このような場合は，データの提供者がどの単位で利用者の引用実績を捕捉したいかという意向も検討材料となる．なお，データ提供者等が引用する場合を想定して，データセット全体としてもDOIを付与している．

例2のLIFULL HOME'Sデータセットでは，通常の不動産賃貸物件データのサブセットとして高精細度間取り図画像データがあるが，空間認識など一部の領域では後者のデータのみを使用することがある．このように異なる領域での利用が想定される場合は，サブセットに対しても独立したDOIを付与している．

例3のYahoo! 知恵袋データ（第3版）では年度ごとに提供データが更新されるが，取得元のサービスは同一であり，複数年度の提供データを同時に利用する場合も十分に想定されることから，第3版データとして単一のDOIを付与し，提供データの更新年度による区別は提供開始年（引用時に記載するデータの出版年）により行うこととしている．

なおYahoo! 知恵袋データについては過去に第1版，第2版データを提供していたが，それぞれデータの仕様が大きく変更され，利用契約も改めて必要であったことから，版ごとにDOIを付与したうえ，現在は提供を終了した第1版，第2版についても，メタデータに関しては恒久的に参照できるようにしている．

3.6　大学等研究者提供データセットへの対応

このようなデータセットの提供活動を続ける中で，音声コーパスに関連するコミュニティから，会話分析や対話処理，コミュニケーション学などの分野の研究を目的として構築された映像コーパスについても受け入れの依頼を受けるようになった．このようなデータでは顔の表情なども重要な要素となるため，顔画像を秘匿化することはできず個人情報を含むことになることから，やはりオープンデータ化は難しいという事情がある．そこで，当初は個別対応としていたが，2019年に「研究者等提供データセット受入要項」を制定し，研究者への提供の枠組みは既存のものを活用することで，受け入れ体制を整えた．すでにこの要項に基づき数種類のデータセットの受け入れを実施したほか，NIIが構築に関与している会話データや手話データといった映像データの提供についても準備を進めている．

4.　データセット提供の現状

4.1　取り扱い中のデータセット

民間企業からのデータセット受け入れは順調に進み，2020年8月の時点で12企業26種類のデータセットを提供するまでになっている．また3.6節で述べた研究者等提供データセットについては4種類のデータの取り扱いを開始している．それらデータセットの一覧を**表3**に示す．

各データセットの詳細や利用条件など，興味のある方はIDRのWebサイト（https://www.nii.ac.jp/dsc/idr/）を参照されたい．

4.2　データセットの提供実績

ここでは民間企業提供のデータセットについて，2019年度末時点の利用状況を以下に述べる．

研究室単位で提供しているデータセットについて，累計利用者数の推移を**図4**に示す．毎年度の新規データの提供開始もさることながら，提供開始から5年以上経過しているデータセットに関してもコンスタントに利用申請が続いており，2020年3月末時点で述べ957研究室と，利用者数は順調に伸びている．また重複を除いた異なり数で645研究室であり，機関数でみると238に上る．当初民間企業の研究所にも提供していた「Yahoo! 知恵袋データ」のうち，現在は提供を終了している第1版および第2版を除くと，異なり研究室数523の内訳は大学：466，研究機関：31，高専等：9，海外の大学：17，異なり機関数194の内訳は大学：152，研究機関：16，高専等：9，海外：17となっており，一部海外の大学も含むが，大部分は日本国内の大学および公的研究機関である．異なり数でみた152大学中，国立大学は49大学であり，これは全国86の国立大学のうち医学系や教育系等の単科大学を除くと約8割を占め，本取り組みが広く認知されていることがうかがえる．

また，利用者の分野の広がりについては，参考程度ではあるが，利用者の所属学部・学科等の名称に含まれる単語のうち5回以上出現したものについて，楽天データセットを提供開始した2010年以降3年度末ごとに頻度順に列挙したものを**表4**に示す．IDRの発足後しばらくは情報検索や自然言語処理分野の研究室からの申請がほとんどであったが，クックパッドデータの提供開始により保健や栄養学といった分野，LIFULL HOME'Sデータセットの提供開始により建築分野や画像処理分野，インテージデータセットの提供開始によりマーケティング学や経済物理学の分野という具合に利用者のすそ野が広がっており，異なり利用者の増加につながっている．

個人単位で提供しているニコニコデータセット，Sansanデータセット，不満データセットのうちカテゴリ別不満特徴語辞書については，2020年3月末現在の利用申請者数（登録メールアドレスの異なり数）は2,919であり，所属は大学が44%，民間が25%，研究機関が2%，その他が29%となっている．民間や個人などにもこのようなデータセットへの需要があることが見て取れる．

なお本稿では詳細は省略するが，NTCIRテストコレクションについてはNTCIRプロジェクトからの提供分も含

表 3　2020 年 8 月現在提供中の民間企業データセットおよび研究者等提供データセット一覧

Table 3　List of the datasets provided by private companies and researchers (available as of August 2020)

民間企業提供データセット	
【Yahoo!データセット】	提供機関：ヤフー（株）
・Yahoo!知恵袋データ（第3版） 提供 2019/01	データは年度ごとに更新．3～5年前の3年間に解決済みとなった質問（2020年度提供版は約270万件）と回答（同約838万件）
【楽天データセット】	提供機関：楽天（株）
・楽天市場データ 提供 2010/08; 最終更新 2020/02	全商品データ（約2億8,300万件），商品レビューデータ（約7,000万件），ショップレビューデータ（約2,250万件）
・楽天トラベルデータ 提供 2010/08; 最終更新 2020/02	施設データ（約2.9万件），レビューデータ（約656万件）
・楽天GORAデータ 提供 2010/08	ゴルフ施設データ（1,669件），レビューデータ（約32万件）
・楽天レシピ 提供 2012/08; 更新 2016/01	レシピデータ（約80万件），レシピ画像（約80万枚），Pickupレシピ（1,854件），デイリシャスニュース（362件）
・アノテーション付きデータ 提供 2014/09; 最終更新 2020/02	研究用にアノテーションが付された各種データ（文単位評価極性タグ付きコーパス，カテゴリラベル付き商品画像データセットなど）
【ニコニコデータセット】	提供機関：（株）ドワンゴ
・ニコニコ動画コメント等データ 提供 2013/04; 最終更新 2018/12	2018年11月までに投稿された動画のメタデータ（約1,670万件）とコメントデータ（約38億件）（動画データ本体は含まれない）
・ニコニコ大百科データ 提供 2014/03	2014年2月上旬までに投稿されたすべての記事データと付随する掲示板全データ
【リクルートデータセット】	提供機関：（株）リクルートテクノロジーズ
・ホットペッパービューティーデータ 提供 2014/09	2012年1月～2014年1月に掲載された店舗（約1万件），店舗ブログ（約180万件），メニュー（約52万件），口コミ（約36万件）など
【クックパッドデータセット】	提供機関：クックパッド（株）
・クックパッドレシピデータ 提供 2015/02	2014年9月までに公開されたレシピ（約172万件）とそれを含む献立
【LIFULL HOME'Sデータセット】	提供機関：（株）LIFULL
・賃貸物件スナップショットデータ 提供 2015/11	2015年9月時点で掲載されていた賃貸物件データ（約533万件），間取り図や室内写真などのサムネイル画像（約8,300万枚）
・高精細度間取り図画像データ 提供 2016/02	賃貸物件スナップショットの画像データのうち，間取り図に関しての高精細度版画像（約515万枚）
・賃貸・売買物件月次データ 提供 2017/08	2015年7月～2017年6月の各月に掲載されていた賃貸・売買の物件データ
【不満調査データセット】	提供機関：（株）Insight Tech
・不満調査データ 提供 2016/05; 更新 2017/08	2015年3月～2017年3月に「不満買取センター」に投稿された不満（約525万件）と投稿したユーザのプロフィール情報（約10万人分）
・カテゴリ別不満特徴語辞書 提供 2017/02; 更新 2017/11	「不満買取センター」への投稿データから投稿カテゴリごとに特徴的な単語を抽出した約190万語
【Sansanデータセット】	提供機関：Sansan（株）
・サンプル名刺データ 提供 2017/05	データ分析コンテストで使用された，ダミーの名刺をスキャナ等で取り込んだ画像（3,481枚）と画像内の位置座標等
【インテージデータセット】	提供機関：（株）インテージ
・インテージパネルデータ 提供 2019/09	2017年に京浜エリアのモニター700名から収集した清涼飲料水の購買履歴およびメディア接触ログ，同エリアの小売店の販売データ
【オリコンデータセット】	提供機関：（株）oricon ME
・顧客満足度調査データ 提供 2019/06	2016年以降に実施された，11ジャンル・88業種（200件）の「オリコン顧客満足度」調査データ
【ダイエット口コミデータセット】	提供機関：（株）T.M.Community
・ダイエット商品口コミデータ 提供 2019/10	2008年8月～2019年10月に「ダイエットカフェ」に投稿された，約8,000商品に対する口コミデータ（約16万件）
【弁護士ドットコムデータセット】	提供機関：弁護士ドットコム（株）
・法律相談データ 提供 2020/03	2017年1月～2019年9月に「みんなの法律相談」に投稿された質問とそれに対する弁護士の回答データ（約25万件）
研究者等提供データセット	
・グループコミュニケーションコーパス 提供 2019/08　（TDU-NEDO）	東京電機大学が作成した，グループディスカッションおよびポスターセッションの様子を収録した映像データと各種アノテーションデータ
・立命館ARC所蔵浮世絵データベース 提供 2020/03	立命館大学アートリサーチセンターが所蔵する浮世絵の書誌情報および画像ファイルへのリンク（約11,000件）
・理研記述問題採点データセット 提供 2020/07	大学受験予備校で行われた模擬試験の記述問題に対し，理化学研究所が採点アノテーションを付与したデータ（9問各2,000件）
・大阪大学 マルチモーダル対話コーパス 提供 2020/08　（Hazumi）	対話エージェントと人との対話の様子をビデオカメラおよびMicrosoft Kinectで収録したマルチモーダルコーパス（59名分，各15分程度）

表 4　利用者の所属学部・学科等名称の頻出語（頻度 5 以上で頻度降順．太字は頻度 10 以上．）

Table 4　Frequent words in users' faculty/department names (descending order of frequency; 5 times or more (10 times or more in bold face))

2010〜2013 年度： **情報，システム**，社会，総合，メディア，知能
2010〜2016 年度： **情報，システム，経営，知能，メディア，社会，電子，総合，コンテンツ，創成，環境，デザイン，文化，通信，経済，ネットワーク**，電気，生命，国際，コンピュータ，未来，技術，物理，政策，図書館，人間，商学，文学，自然科学，現代，建築，教育，サイエンス，知識，人文，心理，栄養
2010〜2019 年度： **情報，システム，経営，メディア，知能，社会，電子，総合，デザイン，通信，環境，文化，国際，創成，教育，経済，人間，コンテンツ，技術，電気，商学，建築，政策，ネットワーク，サイエンス，コンピュータ，生命，知識，都市**，管理，基盤，図書館，物理，未来，自然科学，創造，ビジネス，学術，現代，産業，人文，先進，先端，文学，医学，栄養，健康，アーキテクチャ，医療，応用，基礎，心理，数理，生産，戦略，農学，文理

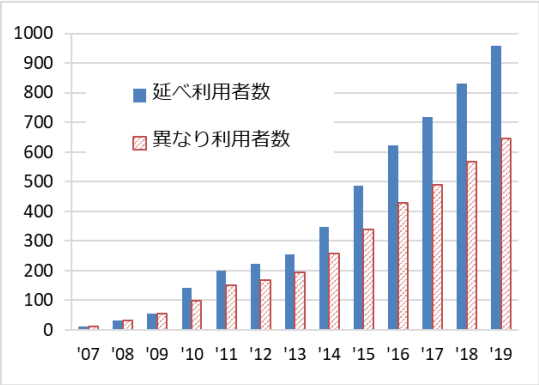

図 4　民間企業提供データセットの累積利用者数の推移（研究室単位で提供中のもの）

Fig. 4　The number of cumulative users (laboratories) of datasets from private companies.

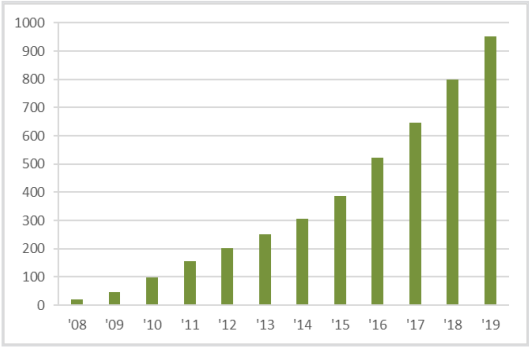

図 5　提供データセットを用いた研究成果の外部発表数（利用報告ベース）の推移

Fig. 5　The number of research articles using the provided datasets from private companies (based on usage reports).

めると延べ約 4,800 件超，音声コーパスについても約 4,300 件超の提供実績を有している．

4.3　提供データセットを利用した研究成果

3.4 節で述べたように，提供したデータセットを利用した研究成果については，利用者から毎年度，発表した論文等の報告書提出を受けている．民間企業提供のデータセットを用いた，2019 年度末分までの発表論文数の合計は約 950 となっており，**図 5** にその推移を示す．

5.　その他の活動

5.1　ユーザフォーラムの開催

研究コミュニティの活動支援の一環として，「IDR ユーザフォーラム」と称したイベントを 2016 年度より毎年開催している．これは主に民間企業提供のデータセットを対象として，データセットの提供者と利用者が一堂に会し，直接意見交換できる場を提供すべく企画したものである．

初開催となった 2016 年度は，データセット利用者の招待講演，データセット提供企業登壇のパネルセッションや企業ごとの個別セッションに加え，データセット利用者による 21 件のポスター発表があり，110 名の参加者を得た．当日のパネルセッションでの議論の内容など，イベントの詳細は，データセット提供企業である株式会社 LIFULL の清田氏による報告記事 [4] をご参照いただきたい．

ユーザフォーラムは 2017 年度以降も毎回 100 名以上の参加者を得て盛会となっている．2018 年度からはポスターセッションに加え，研究着手段階での研究アイディアの発表を受け付けるスタートアップセッションを設け，学部生にも多く発表いただくとともに，ポスター賞を受賞したうち数件の発表については，翌年度のユーザフォーラムにて口頭発表いただく機会を設けている．

このように，同じデータセットの利用者と議論を交わすだけでなく，データセット提供者から直接アドバイスを受けたり，逆にデータセット提供者に対し要望を伝えたりできる場はこれまでになく，参加者には好評である．提供中のデータセットの利用に興味がある方や，データの提供に興味がある企業等関係者にも参考となるものと考えている．

5.2　コミュニティ開催イベント等への支援

大学等の研究者から，提供中のデータセットを評価ワークショップやコンペティション，学生向けのイベント等に利用したいという相談を受けた場合には，可能な範囲で提供者への仲介を行っている．このような用途ではデータセットの通常の提供条件の範囲では利用が認められないことが多く，提供者，利用者の双方にアドバイスを行って調整を手助けしている．

また，提供者が企画する研究集会やアイデアソン・ハッカソンなどに講演やデータセット提供といった形での協力も積極的に行っている．

6.　おわりに

本稿では，情報学や関連諸分野の研究を推進するため，IDR が取り組んでいる活動について，その背景や意義とともに紹介した．IDR の活動は，民間企業等の実サービスの中で作成されたデータや，個人情報を含む映像データ等，通常では共有が難しいデータセットを中心に共同利用に供し，研究の透明性と再現性を高め，多くの多様な分野の研究者に平等に研究の機会を提供するという意味で，オープンサイエンスの推進にも寄与するものと考えている．

IDR の活動のうち，データセットの提供は最も基礎となるものであり，取り扱うデータセットの種類を着実に増やしているところではあるが，現在提供している民間企業のデータセットの多くはウェブ上の実サービスに蓄積されているデータのスナップショットや一部の時系列データであり，今後はトランザクションログなど，より厳密な管理を要するデータへ幅を広げることが望まれる．また現状ではリスクを低減させるためにデータを加工して提供せざるを得ないが，研究者からはより原データに近い詳細なデータへの要望も多い．これに応えるためには，データそのものを利用者に開示することなく，利用者が作成したプログラムを実行し結果のみが得られるようにする仕組みを整えるなど，技術的にも安全にデータを共同利用できる環境を構築していく必要がある．

一方で，データセット提供者と利用者との交流を活性化させ，相互理解を進めることも重要である．研究者側も，要望を一方的に伝えるばかりではなく，まずは研究室内のデータの管理や利用者の管理に責任を持ち，提供者の立場も理解したうえでデータセットの利用方法や論文等での言及には細心の注意を払い，利用報告等の利用者の義務をきちんと果たすことが望まれる．このようにして信頼関係を積み上げていくことは，今後のオープンサイエンスの普及にも不可欠であると考える．そのような土壌の醸成にもIDR として一役買うとともに，ユーザフォーラムの開催などを通して，提供者，利用者の双方を巻き込んだ研究コミュニティの活性化に一層努めていきたいと考えている．

参考文献

[1] 森　正弥：ビックデータ時代における E-Commerce での AI 技術活用，人工知能，Vol.30, No.3, pp.310–317 (2015).
[2] 内閣府：第 5 期科学技術基本計画，（平成 28 年 1 月 22 日閣議決定），〈https://www8.cao.go.jp/cstp/kihonkeikaku/5honbun.pdf〉（参照 2020-08-12）．
[3] 日本学術会議：科学技術における健全性の向上について（平成 27 年 3 月 6 日），〈http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-23-k150306.pdf〉（参照 2020-08-12）．
[4] 清田陽司：集会報告 NII-IDR ユーザフォーラム 2016，情報管理，Vol.59, No.12, pp.867–871 (2016).

大須賀 智子（非会員）

2006 年千葉大学大学院自然科学研究科博士後期課程修了．博士（工学）．2003 年日本学術振興会特別研究員（DC1）．2006 年より国立情報学研究所勤務．現在，データセット共同利用研究開発センター特任研究員．音声言語資源の構築・整備やデータセットの共同利用に関する事業に従事．

大山 敬三（正会員）

1985 年東京大学大学院工学系研究科電気工学専攻博士課程修了．工学博士．その後，東京大学文献情報センター助手，学術情報センター助手・助教授・教授を経て国立情報学研究所教授，総合研究大学院大学複合科学研究科教授．データセット共同利用研究開発センター長を兼務．情報検索や Web 情報アクセス・利用技術などの研究に従事．