Vol.2018-IFAT-129 No.10
2018/2/17

情報学研究データリポジトリ IDR における データセット共同利用

大須賀智子[†1]　大山敬三[†1†2]

概要：国立情報学研究所では，情報学研究に資するため，情報学研究データリポジトリ（IDR）を運営し，大量のデータを保有している産学界と，それらを必要としている大学等の研究者の橋渡しをしている．本稿では，データセットの共同利用を行うことの背景や意義を述べるとともに，提供中のデータセットとそれらの利用状況や，IDR が民間企業や研究者に向けて行っている活動内容について紹介する．

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，情報学研究データリポジトリ

Informatics Research Data Repository (IDR) as a Hub of Sharing Datasets for Informatics Research

TOMOKO OHSUGA[†1]　KEIZO OYAMA[†1†2]

Abstract: Large scale datasets are indispensable for informatics research in these days. Since there are various difficulties in practice, however, National Institute of Informatics has set up Informatics Research Data Repository (IDR) to promote dataset sharing, having it accept datasets from private companies, etc. and distribute them to researchers. This article overviews the background and the meanings of dataset sharing. It also presents IDR's activities for private companies and researchers concerning dataset sharing, as well as the datasets currently provided and the usage status of them.

Keywords: Large Scale Datasets, Shared Use, Open Sciences, Informatics Research Data Repository

1. はじめに

最近は人工知能（AI）という言葉が毎日のように耳に入るほど社会に浸透してきているが，これはビッグデータと呼ばれる実社会で生成された大規模データを，ディープランニングに代表される統計的手法により処理する技術やその計算機環境等が整ってきたことが背景にある．

このうち技術や環境に関しては，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，誰でも最先端の技術を利用できる状況になりつつあるが，一方でデータに関してはその性質から公開できるものばかりではなく，データの取得元でなければ利用には制限がある．情報アクセス技術やデータ解析技術などの研究においては，実社会で生成された大規模データが不可欠な研究資源となっているものの，特に大学等の学術研究機関では自らデータを取得することは難しい．自然言語処理の分野ではテキストデータを取得するだけでなく，それらに必要なアノテーションを施すのにも多大な労力がかかり，音声情報処理やコミュニケーション研究の分野では音声データや映像データの収録自体にコストがかかるため，多機関で協力してコーパスを構築することが多いが，それらを流通させるには課題が多い．

†1 国立情報学研究所
National Institute of Informatics
†2 総合研究大学院大学
SOKENDAI

このような状況を鑑み，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）の活動を通して，情報学研究に必要な各種データを取得元から受け入れ，適切な権利処理を施すことによって，大学等の学術研究機関を中心に，多くの研究者に提供できるように取り組んでいる．NII は情報学分野における大学共同利用機関として，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（これを「共同利用」という）を使命の一つとしており，その一環として実施しているものである．

以下では，データセット共同利用の意義について改めて述べるとともに，IDR が現在提供を行っている各種データセットとその利用状況等を示す．他に研究コミュニティの支援に向けて行っている活動について紹介し，今後のデータセット共同利用の深化に向けた取組みについても述べることにする．

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おり，実用化への道筋を示すためにも，実社会で生成された大規模データが研究資源として必要になっている．研究室内で疑似的に生成したデータでは量的に充分とは言えず，研究の透明性に欠けるという問題もある．知り合いなどを頼ってデータを所有する民間企業等と共同研究契約を交わしデータを提供いただくケースもあるが，個別に提供

を受けた場合には，研究の透明性に加え再現性にも欠けるという点で大きな問題がある．公的研究資金を用いた研究成果はオープンデータ化を推進することが閣議決定されたところであるが[1]，それに伴って実験用に作成したデータを公開することや，外部から提供を受けたデータを使用した場合にはそれを特定することが義務化されつつある．

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[2]．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．もちろん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．音声データや映像データも著作権や個人情報の問題があり，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．

そこで IDR では，情報学に関連する各種研究用データセットについて，データセットの取得元や作成者から受け入れ，より多くの研究者に提供できるようにするため，その一元的な窓口となることを目指している．

大学等の研究者にとっては，実社会のデータを使用できるだけでなく，使用したデータセットを特定できることにより，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，その後の労力は大幅に削減することができる．また民間企業においては，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．

このようなデータ提供の窓口となる組織としては，国内では言語資源協会（GSK）[a] 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）[b] 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[c] や欧州の ELRA[d] が言語資源を大規模に収集・提供しており，Microsoft[e] 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，原則として無償で提供している組織はあまり類を見ない．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みつつある．

3. データセット提供の現状

3.1 IDR 発足の経緯

IDR の活動の起源は，「Yahoo!知恵袋データ（第 1 版）」をヤフー株式会社から受け入れ，2007 年 4 月に研究者に対して提供を開始したところに遡る．当初は単発のデータセット提供であったため特にサービス名称は定めていなかったが，利用者も増え提供データセットの拡大も見込まれるようになったことから，2010 年 1 月に名称を「情報学研究データリポジトリ（IDR）」として活動することとなった．

一方 NII では，前身時代の 1997 年より「NTCIR プロジェクト」（NTCIR）[f] 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，研究者に配布してきた．また 2006 年に「音声資源コンソーシアム」（SRC）[g] を設置し，音声コーパスの調査・カタログ化・提供を行ってきた．そして，これらと IDR の活動を総合し，情報学及び関連分野を対象とした研究用データの収集・構築・提供及びこれらに係わる共同研究を強化する目的で，2015 年 4 月に「データセット共同利用研究開発センター」（DSC）[h] が設置され，IDR は現在，このセンターの下で活動を行っている．主に IDR が関与しているデータセットの提供に関係する活動の概念を図 1 に示す．

IDR での民間企業からのデータセット受け入れは順調に進み，2018 年 1 月の時点で 8 企業から 15 種類のデータセットを提供するまでになっている．他にも，NII が構築・提供に取り組んできた NTCIR テストコレクション（一般研究目的利用の大部分）や音声コーパスの提供について窓口を IDR に集約するとともに，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．また現在は新設された「人文学オープンデータ共同利用センター」に移管したが，国文学研究資料館が作成している「国文研古典籍データセット」に関しても公開当初の支援を行った．

a) http://www.gsk.or.jp/
b) http://www.alagin.jp/
c) https://www.ldc.upenn.edu/
d) http://www.elra.info/en/
e) http://research.microsoft.com/en-US/projects/data-science-initiative/datasets.aspx
f) http://research.nii.ac.jp/ntcir/index-ja.html
g) http://research.nii.ac.jp/src/
h) http://www.nii.ac.jp/research/facilities/dsc/

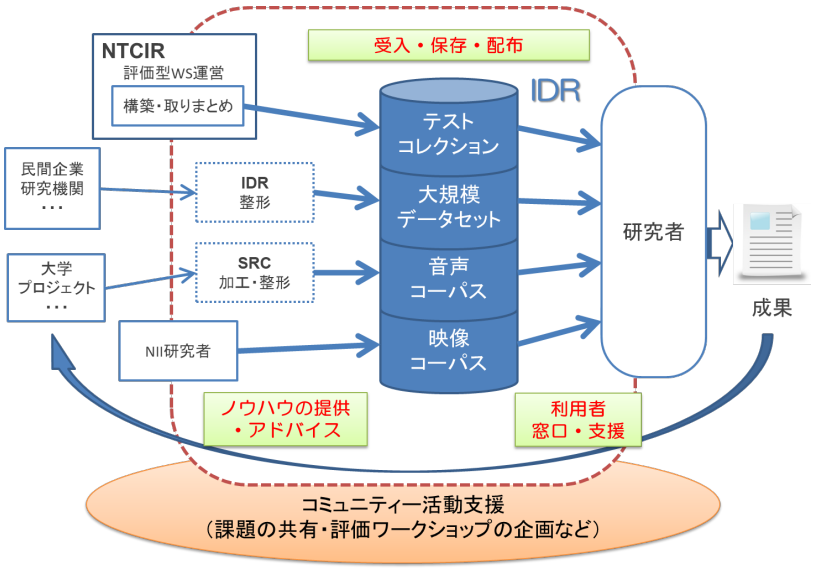

図 1　データセット提供に関わる活動

3.2 取り扱い中のデータセット

IDR では提供者となる民間企業や大学あるいは研究者（以下，提供者という）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を行っている．2018 年 1 月現在提供しているデータセットの一覧を表 1 に示す．

これらのデータセットは，2 章で述べたようにオープンデータとすることが困難なものであるので，利用契約の締結後に提供することになる．利用契約の締結形態は，データセットの性質や提供者の方針によりいくつかのパターンに類型化される．表 1 の (1) 民間企業提供データセットのうち，Yahoo!データセット，楽天データセット，リクルートデータセット，クックパッドデータセット，LIFULL HOME'S データセット，不満調査データセット（カテゴリ別不満特徴語辞書データを除く）は研究室単位で配布を行っており，NII が提供者からサブライセンスの許諾を受けている場合は NII と利用者との間の覚書，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．ニコニコデータセット，Sansan データセット，不満調査データセットカテゴリ別不満特徴語辞書データについては個人単位で配布を行っており，これらはいずれも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．表 1 (2) の NTCIR テストコレクションや音声コーパスも上記のいずれかに準じた形態となっている．

特に提供者が民間企業の場合，配布先や利用にあたっての条件を様々課されることになるが，IDR としては利用者の立場からできる限り利用申請時の手続きを共通化・簡略化するとともに，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ようにしている．一方で必要に応じて利用申請時に個別に確認を取るなどして，提供者も安心してデータを提供できるよう努めている．

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（http://www.nii.ac.jp/dsc/idr/）を参照されたい．

3.3 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，2017 年 12 月末現在の利用状況を以下に示す．

Yahoo!データセットや楽天データセットをはじめとした研究室単位で提供しているデータセットについては，利用手続きが完了しデータを配布した利用者数（研究室数）の合計は延べ 700，重複を除いた異なり数は 488 である．機関数でみると 206 機関に上る．大部分は日本国内の大学及び公的研究機関である．図 2 には累計の延べ利用者数と異なり利用者数の推移を示す．

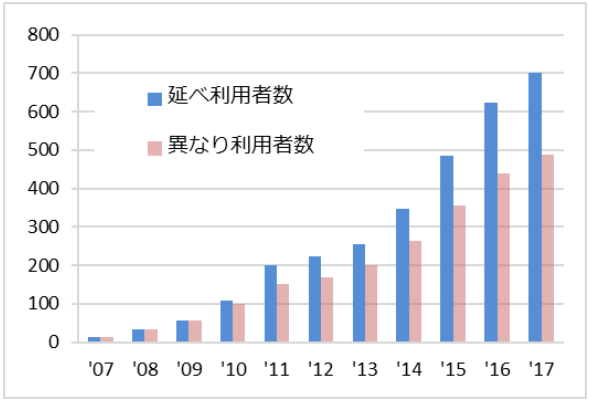

図 2　民間企業提供データセットの累積利用者数の推移（研究室単位で提供中のもの）

表 1　2018 年 1 月現在提供中のデータセット一覧

(1) 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 2 版) 提供 2011/01	2004 年 4 月～2009 年 4 月に解決済みとなった質問(約 1,600 万件),回答(約 5,000 万件)と各種付随データ. 本データを用いたテストコレクション「NTCIR-8 CQA」も併せて提供
【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; 更新 2011/08, 2014/04	全商品データ(約 1 億 5,600 万件), レビューデータ(約 6,400 万件)
・楽天トラベルデータ 提供 2010/08; 更新 2016/01	施設データ(約 13 万件), レビューデータ(約 558 万件)
・楽天 GORA データ 提供 2010/08; 更新 2011/08	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)
・PriceMinister 提供 2017/04	ユーザの商品レビュー (学習用 8 万件, 評価用 3.6 万件), レビューの有効性情報 (学習用 8 万件)
・アノテーション付きデータ 提供 2014/09; 更新 2017/11	・筑波大学文単位評価極性タグ付きコーパス(TSUKUBA コーパス) ・カテゴリラベル付き商品画像データセット ・文字領域アノテーション画像 ・楽天不動産間取り図と壁ラベル
【ニコニコデータセット】	**提供機関:(株)ドワンゴ, (株)大百科ニュース**
・ニコニコ動画コメント等データ 提供 2013/04; 更新 2016/12	2016 年 8 月までに投稿された動画のメタデータ(約 1,400 万件)とコメントデータ(約 35 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全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), 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**
・賃貸物件データ, 画像データ 提供 2015/11 (高精細間取り画像データ 提供 2016/02)	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画像データ(約 8,300 万枚), 高精細度間取り画像データ(約 515 万枚)
【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」に投稿された不満データ(約 525 万件)と投稿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	実物を模したダミーの名刺 3,481 枚をスキャナ等で読み込んだ画像データと, 名刺を構成する 9 種類の領域の画像中の位置座標
(2) その他データセット	
【NTCIR テストコレクション】	**提供機関:国立情報学研究所**
・各種タスクデータ 提供 2012/09 以降(窓口を IDR に移行)	NTCIR プロジェクトで構築したテストコレクションのうち 13 分野, 32 種類のタスクデータ
・文書データ 提供 2012/09	NTCIR WEB タスク用に主に日本の Web サイトから収集したデータ ・NW100G-01(2001 年版, 約 1,100 万文書, 100GB) ・NW1000G-04(2004 年版, 約 1 億文書, 1,400GB))
【音声コーパス】	**提供機関:大学, 民間企業等**
・各種音声コーパス 提供 2010/01(窓口を IDR に統合)	音声資源コンソーシアム(NII-SRC)がさまざまな機関やグループから受け入れた 43 種類の多様な音声研究用コーパス

延べ利用者数，異なり利用者数ともに，最近 3 年ほどは新規データセットの提供開始により増加が加速傾向にある．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．一方で，第 1 版の提供開始から 10 年以上が経過している Yahoo!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いる．

個人単位で提供しているニコニコデータセットや Sansan データセット等については，2017 年 12 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 2,387 であり，所属は大学が 45.2%，民間が 25.0%，研究機関が 2.6%，その他が 27.0%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めると延べ約 4,000 件，音声コーパスについては約 3,600 件の提供実績を有している．

3.4 提供データセットを利用した研究成果

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，発表した論文を報告してもらうこととしている．民間企業提供のデータセットを用いた，2016 年度末分までの発表論文数の合計は約 530 となっており，図 3 にその推移を示す．

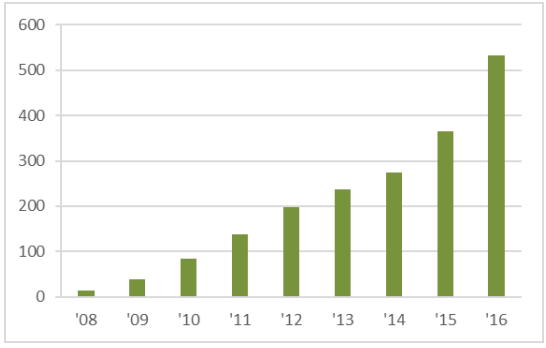

図 3　提供データセットを用いた研究成果の外部発表数（利用報告ベース）の推移

なおこれらの論文リストの一覧を公開すべく，NII で開発しているリポジトリモジュール WEKO を活用したリポジトリの構築を進めている．図 4 に示すように，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他，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．現在のところ民間企業データセットならびに音声コーパスを用いた研究成果を掲載しているが，別途公開している NTCIR テストコレクションを用いた研究成果リスト[i]も統合予定である．提供中のデータセットを用いた研究の参考としてご活用いただければ幸いである．将来的には研究動向分析なども行いたいと考えている．

4. コミュニティの活動支援

4.1 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同に会し，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 名の参加者を得て盛会であった．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[3]をご参照いただきたい．

2017 年度も第 2 回として 12 月に開催し，101 名の参加者を得た．トークセッション 1 件とパネルセッション 2 件では，いずれも学界と産業界の講演者に同時にご登壇いただき，双方の立場を踏まえた上での今後の連携の在り方などについて活発な議論が行われた．データセット利用者によるポスター発表は 27 件あり，データセット提供企業によるポスター展示も含めて，発表者と参加者間で熱心な議論が交わされた．

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いただけたり，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，参加者からは好評をいただいている．2018 年度も開催予定であるので，提供中のデータセットの利用に興味はあるがまだ使ったことがない方や，データの提供に興味をお持ちの企業等の方にもぜひご参加いただき，コミュニティを共に活性化できればと考えている．

4.2 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ることがあり，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．

i) http://ntcir.nii.ac.jp/jp/Papers-on-NTCIR-using-NTCIR/

図 4　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

5. 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，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であり，今後は時系列データやトランザクションログなど性質の異な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上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着実に積み上げていくことも，さらなるデータの提供やより強固な関係の構築につながるものと考える．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努めていきたいと考えている．関連分野の皆様にも，データセットの提供者や利用者という立場だけでなく，研究コミュニティの一員として，IDR の活動を支援していただけたら幸いである．

参考文献

[1] “世界最先端ＩＴ国家創造宣言・官民データ活用推進基本計画（平成 29 年 5 月 30 日閣議決定）”．http://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20170530/siryou1.pdf,（参照 2018-01-15）．

[2] 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. 2015, vol. 30, no. 3, p. 310-317.

[3] 清田陽司．集会報告 NII-IDR ユーザフォーラム 2016．情報管理. 2016, vol. 59, no. 12, p. 867-871.