第 2 回 SPARC Japan セミナー2015

「科学的研究プロセスと研究環境の新たなパラダイムに向けて
－e-サイエンス，研究データ共有，そして研究データ基盤 -」

研究資源としての
データセットの共同利用について

大山　敬三
（国立情報学研究所）

講演要旨

情報技術の分野では、研究と実用化の距離が急速に縮まってきており、実運用のシステムで生成される大規模な実データが必須の研究資源となってきている。国立情報学研究所に設置されたデータセット共同利用研究開発センターでは、インターネット上で事業を展開している企業等を中心に、各種のデータセットを受け入れ、「情報学研究データリポジトリ」（NII-IDR）を通じて広く研究者に提供する活動を行っている。本講演では、NII-IDR の活動を中心として、現状と将来展望について述べる。

大山　敬三

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。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・教授を経て現職。2015年4月に設置された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・利用の研究に従事。

情報学の分野では、研究と実利用の間の距離が狭まってきています。その結果、実サービスから生成される大規模な実データが研究資源として不可欠になってきています。民間企業が実データを多く持っていますが、それをどうやって提供してもらえばいいのでしょうか。彼らにとっては社会貢献、現在の課題解決や将来技術の開発、共同研究、学生など人材のリクルートなどがインセンティブになります。しかし、このタイプのデータでは、オープンアクセスはその解にはなり得ません。また、このようなデータを研究者がそれぞれ勝手に集めようとすると、面倒な問題が出てきます。

そこで、共通のデータを使おうという話になってきます。これには次のようなメリットがあります。研究者にとっては、研究の再現性や透明性が保証されること、研究結果を互いに比較するのが簡単になること、認知されているデータを使うと研究自体をアピールしやすくなることなどです。

研究者のコミュニティーにとっては、1 点目は、共通のデータセットを使って技術の比較評価のためのプラットフォームをつくれることです。これは個々の研究者にとってもそうですが、コミュニティーとして研究を加速するためには非常に重要で、共通の研究課題を設定したり、評価手法を定義したり、研究成果を蓄積して過去のものと比べられるような環境をつくっていったりすることが含まれてきます。2 点目は、コミュニティーを強化するだけではなく、コミュニティー

の間をつなぐような連携が、このようなデータセットを通じて可能になってくることです。

データを提供する側にとっては、データセットの認知が広まることによって、社会貢献がより広く知られるようになることと、企業として研究コミュニティーに対してオープンである、あるいは公平であるという姿勢をアピールすることができるということです。

データセット共同利用研究開発センターの活動

データセットの共同利用のために、2015 年 4 月に私がセンター長を務めるデータセット共同利用研究開発センターができました。実際には 10 年以上前から幾つかの関連する活動があり、それらをまとめる形でつくられました。情報学研究の推進を目的として、データセットの提供に関係する活動を行っています。データの収集、受け入れ、配布だけではなく、そのために必要なノウハウを蓄積し、いろいろなところと共有していきます。さらに、コミュニティーをつくったり、つないだりすることによって、研究を活性化します。研究者だけでなく、データを持っている人、提供してくれる人、つくる人、ユーザー、そしてわれわれ自身もコミュニティーの一部だと思っています。

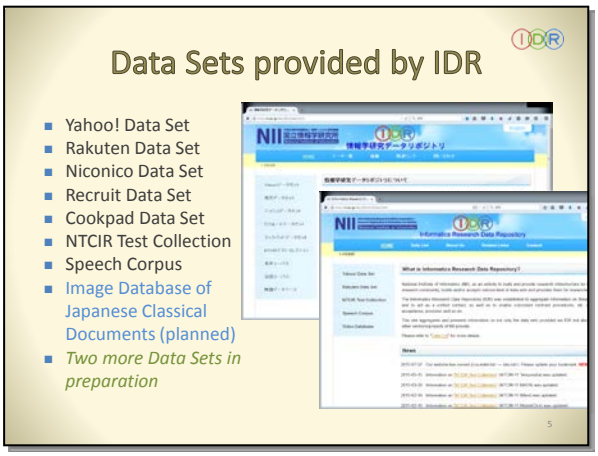
センターで特にデータ提供に関するサービスを「情報学研究データリポジトリ（IDR）」と呼んでいます。図 1 にあるようなデータを受け入れた上で、研究者に提供しています。これらのデータは三つのカテゴリーに分類されます（図 2）。一つ目は、民間企業の特に商用のインターネットサービスから生成された実データです。Yahoo!データセット、楽天データセット、ニコニコデータセット、リクルートデータセット、クックパッドデータセットなどがあり、これらはいずれもインターネット上のかなり大規模なサービスで蓄積されたデータを提供してもらっています。二つほど現在準備中のものがあります。これも皆さん結構おなじみのもので、近々公開できると思うので、楽しみにしていてください。

二つ目は、研究目的で研究者や研究機関によってつくられたデータです。音声コーパスは、音声研究、音声認識、音声合成、音声解析などを行っている研究者や企業等でつくられた研究用のデータです。古典籍データセットは近々公開しますが、日本語の歴史的典籍の画像データを中心としたデータセットです。これは国文学研究資料館がいろいろな大学と連携してデータベースをつくり、その公開について IDR がお手伝いするというものです。

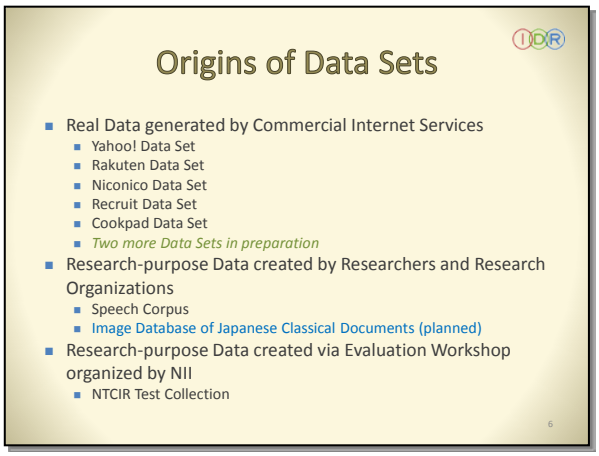
三つ目は、「評価型ワークショップ」を通して得られたデータです。ここでの評価型ワークショップとは、特に情報アクセス技術の評価のために行われているワークショップで、その過程でつくられた NTCIR テストコレクションというデータがあります。

このようにつくられたデータをどうやって配布しているかという手続きも大きく分けて三つあります。

一つ目は、契約書ベースの提供です。データの提供者には、それぞれの考え方やデータの性質などに応じ



（図 1）



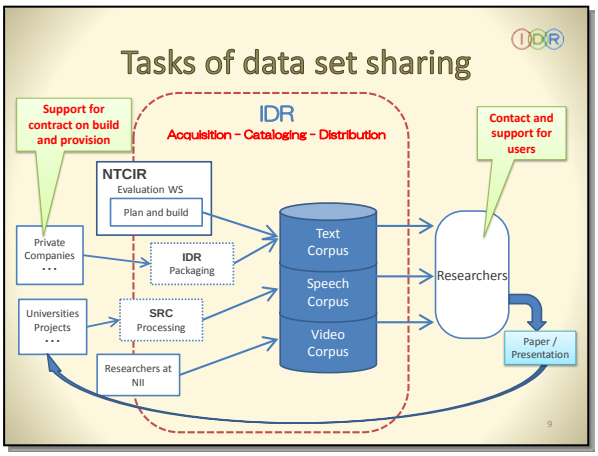
（図 2）

て、どのようにしたいという希望があるため、契約書ベースにもバリエーションがあります。NII がサブライセンスを受けて契約する、企業が直接契約する、契約までは要らず、利用規約に同意書を出してもらえばいいといったレベルが幾つかあります。

二つ目は、オンライン登録すれば使えるというものです。ニコニコデータセットは、皆さんよくご存じのニコニコ動画のデータセットで、名前、電子メール、その他を少し登録してもらえば使えます。名前も実名でなくハンドルネームでも何でもいいという、かなり緩やかなものになっています。NTCIR テストコレクションについては、きちんと名前、所属、電子メールを入れることになっています。

三つ目は、オープンアクセスです。古典籍データベースについては、CC BY-SA でオープンアクセスを予定しています。

次に、データを提供する側は、いろいろ心配事があるので、データの使い方に制約を掛けることがあります。特に民間企業の場合は心配事が大きく、コピーライト、プライバシーや個人情報、ユーザーからの非難、経済的利益の機会を逃すことなどが挙げられます。特に商用のインターネットサービスの事業者では、サービスのユーザーのプライバシーや個人情報に関して非常に慎重です。また、法律的には全く問題なくても、使い方によってユーザーから非難を受けて炎上してしまうというようなことは起こしたくないと考えています。本当はこれが一番怖いです。



（図 3）

このような心配があるので、データを第三者に提供してはならない、商用利用してはならない、学術的な出版物や研究会等でも、個人や団体を特定できる情報は公表してはならない、といった条件が付きます。データセットによっては、インターネット上の情報とのマッチングが禁止されています。提供いただくデータは個人データを除去してありますが、テキストはそのままで入っているので、それをインターネット上でそのまま検索すると、ユーザーのハンドルネームなどが分かるため、個人情報に近いものが判明してしまうのです。それを分析するとますますまずいことが分かる可能性があるので、これを気にして禁止することがあります。一部のデータセット提供者は、研究成果を発表する際に事前チェックを要求します。これも同じようなことを気にしているからです。

ここ約 10 年間の新規利用申請数を見ると、ニコニコデータセットを公開したところで急増しています。日本中に情報学分野でこの種のデータを使う研究者グループがどのぐらいあるか推定するのは難しいのですが、使う可能性のある研究者はかなりの比率で使っていると考えています。

データセット共同利用の流れ

図 3 は、データセット共同利用研究開発センターにおける IDR のサービスに関する業務を図示したものです。データを受け入れ、保管し、概要を公開し、オンラインで研究者に配布します。研究者は研究成果を発表したものを、年に 1 回報告してもらいます。これらの研究成果は集約して、データ提供元にフィードバックをしています。この全体の過程において、データを提供していただける方には、契約方法やデータ構築方法などに関する支援をしたり、研究者に対する統一的な窓口としての機能を果たしたり、簡単な技術支援をしたりという活動も行っています。

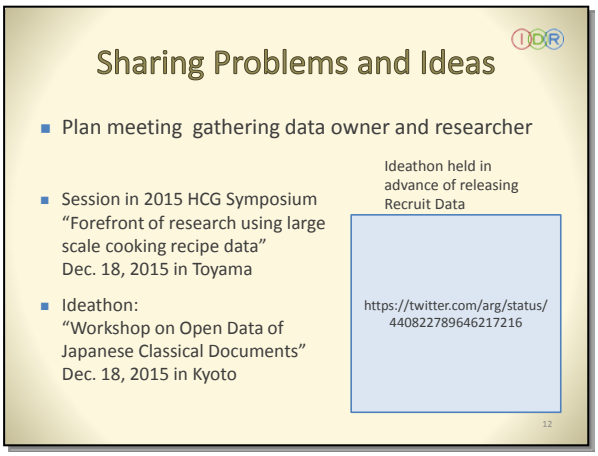
データの提供者と話をするときは、著作権やデータ配布のためのノウハウの提供や、データセットを提供する際のアイデアソンの提案などを行い、研究成果の

フィードバックなど業務にも貢献できることを説明します。

その事例として、特定のデータセットにフォーカスして、情報交換会のような研究集会を開くということが行われています。また、データを配布する前に、研究者の興味を把握したり、現在の課題や意識、アイデアを共有するために、企画型の集会などが開かれています。12 月 18 日にはたまたま同じ日にこのような集会が二つ予定されています（図 4）。一つは、料理データを使ってどういう研究ができるのかについて、研究者やデータ提供者が集まって意見交換をします。もう一つは、古典籍データセットで、どういう研究をしたらいいかということについてアイデアソンを行います。ご興味があれば、ぜひご参加を検討ください。

コミュニティーをつくるということについては、評価型ワークショップを開催し、一つの厳密に定義された課題をみんなで共有して、いろいろな技術でトライする中で、非常に強いコミュニティーが形成されるということがあります。IDR のデータセットを使った事例としては、Yahoo!知恵袋データを使ったコミュニティーQA や、楽天レシピデータを使ったレシピサーチなどがあり、コミュニティーの形成に寄与しています。データセットを使ってコンペティションを行うことも有効な手段で、そのような取り組みも行っています。

また、異なる分野のコミュニティーをつなぐということでいえば、例えばレシピのデータでは、栄養学や経済、環境などの分野の研究者からも使いたいという要望が出てきます。このような共通のデータセットがあると、それに興味を持ったさまざまな分野からのアクセスも期待できるようになります。これらを、アイデアソンなどを使ってつないでいく活動はこれから積極的にやるべきことだと考えています。



（図 4）

今後の課題

まず評価についての課題があります。一つ目は、ユーザーの把握です。契約ベースの提供なら簡単ですが、オンライン登録だとすぐにユーザーが追跡できなくなりますし、オープンアクセスだと一体どうすればいいのか、という問題があります。

二つ目は、ユーザーが生産した研究成果の把握です。これは DataCite やデータジャーナルが一つの解になると思いますが、IDR が扱っているようなデータセットが果たしてこれに載るかどうかは定かではないので、研究が必要なところです。

三つ目は、研究に対してのデータセットの価値の測定です。

四つ目は、このようなデータセットを提供するという IDR の活動の価値の測定です。これらをきちんと評価することが、われわれが活動を継続してゆくためには必須となるので、お知恵があればぜひ教えていただきたいと思います。

また、現在はライセンスや個人情報などが制約となって提供できていないものを提供してもらえるようにする環境づくりを考えています。例えば、個人情報に関わりそうで危ないけれど研究には使いたいデータ、サイズが大き過ぎるデータ、商品価値が高いために流出が懸念されるデータなどを提供してもらうために、クラウドベースのプラットフォームをつくる、あるいは、Evaluation as a Service（EaaS）という形で、ネットワークから隔離されたシステム環境の中で評価をして結果だけ返すという仕組みをつくる方向で考えています。その上で、どのようなライセンスのスキームにすればどこまでデータを提供してもらえるかについては、社会的な面からの研究が必要になってくると考えています。