国立情報学研究所ニュース　ISSN 1883-1966（Print）　ISSN 1884-0817（Online）

NII Today

71
Mar. 2016

National Institute of Informatics News

NII Interview

オープンサイエンスの時代へ

オープンデータの最前線

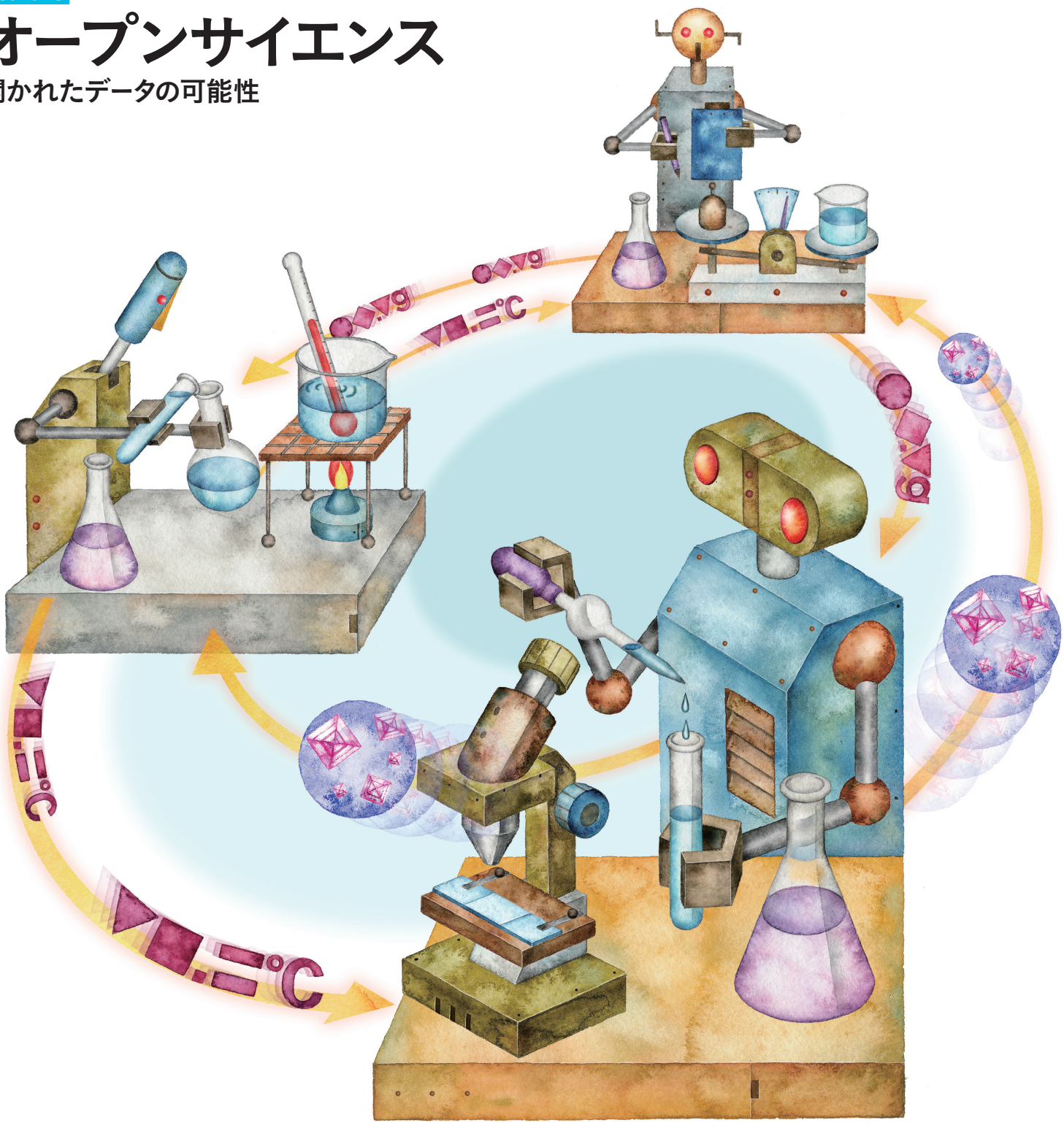
機関リポジトリから
オープンサイエンスへ

実環境データを情報学研究に活かす

Feature

オープンサイエンス

開かれたデータの可能性



大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

オープンサイエンスの時代へ

データ共有化でNIIの果たす役割は?

喜連川 優［国立情報学研究所 所長］
聞き手：**滝 順一氏**［日本経済新聞社 論説委員 兼 経済解説部編集委員］

科学論文の根拠となる実験データなどを公開、共有化する動きが始まっている。過去数百年の間、論文誌や学会発表を通じて行われてきた情報の共有化が、情報通信技術の発展によって大きく変わろうとしている。この「オープンサイエンス」の動きは多種多様な研究者、研究領域が生みだした知識を組み合わせ、研究を加速するだけでなく、新たな知識を生む触媒にもなりうる。日本の科学研究を支援するIT基盤を提供してきた国立情報学研究所はこの流れの中でどのような役割を果たそうとしているのか、喜連川優所長に聞いた。

喜連川 優
KITSUREGAWA Masaru

滝　オープンサイエンスという言葉をよく耳にするようになりました。

喜連川　オープンサイエンスには二つの論点があります。「オープンアクセスジャーナル」と「オープンリサーチデータ」です。これまで科学論文を載せた雑誌は、購読者がお金を払って購読していました。論文著者が出版社にお金を払って、無料で広く一般の人が論文を読めるようにするのがオープンアクセスジャーナルで、世界的に広がりを見せています。

一方、オープンリサーチデータは、論文と一緒に論文の根拠となるデータを公開する動きです。データがあれば論文で主張されていることの再現が容易になります。その結果、多くの研究者が論文の結論やデータを早く活用でき、科学の進展やイノベーションを加速できます。再現できない論文を出す行為を減らす効果も期待できます。データが出ていれば次の人はそれを利用して研究するので、単に研究が加速するだけではなく、重複投資を避けられ、研究を効率的に進められます。

オープンアクセスジャーナルの議論はひとまずメドがついた感じなので、いまはオープンリサーチデータがホットな課題になってきました。

滝　データの公開は大切ですが、研究者にはそうするインセンティブがないように思えます。

喜連川　論文を識別する番号「デジタルオブジェクト識別子」(DOI=Digital Object Identifier) がありますが、データにも識別番号を与えて「このデータを使って論文を書きました」とデータを引用する習慣がすでに始まっています。貴重なデー

タを生み出した研究者にリスペクトを示し、データ公開へのインセンティブを生み出す動きです。

ただ論文の評価に比べてデータの評価は難しい。データの精度は利用目的によって水準が異なりますし、データの正しさも使い方の局面が変われば違ってくることもありえます。論文と同じようにはいかないかもしれません。

丹精込めてつくったデータを我が物としたい研究者の気持ちは理解できます。ただ論文の再現可能性を担保するのは科学者の責務です。相手が出すものは自分も出すのが前提です。公平感をいかに実現するかは、広く言えば「外交」の問題といってもいいと思います。日米欧などの研究機関がつくったリサーチデータ・アライアンス（RDA）と呼ぶ組織があり、データの共有化でどんな価値観が新たに生み出せるか議論をしています。

滝　NII はデータのオープン化の流れにどう対応するのですか。

喜連川　NII は、大学など研究機関の学術情報を収集・保存・利用するための「機関リポジトリ」の運用で大学図書館などを支援してきました。共用リポジトリサービス「JAIRO Cloud」を提供し国内 465 の大学や研究機関に利用してもらっています。このサービスを拡張してデータも格納できるようにすれば、大学から喜ばれるのではないかと考えています。

データの性格は天文学や高エネルギー物理学、ゲノム解析、物質材料研究など領域によって異なり、データの扱い方の慣習も違います。データの内容などを示すために付与するメタデータをどうするかなど、領域ごとの研究者と相談して決めていかなければならず、データの格納は論文とは違った難しい面があります。少々時間がかかるかもしれませんが、一歩一歩進める必要があります。

滝　NII がストレージサービスを提供して研究を下支えするわけですね。

喜連川　そうです。ただデータには必ずその解析に用いたプログラムが存在します。解析の再現性を担保するにはプログラムも格納する必要がありますが、これも技術的にはなかなかしんどいことです。みなさんがパソコンで経験されているように、OS のバージョンが変わるとアプリケーションソフトが動かなくなることはよくあることです。

こう考えると、長い目でみて、NII は領域ごとにその領域の研究者のみなさんと、データの格納のやり方とデータを料理するプログラムを載せた研究のプラットフォームを一緒に考え、つくっていくという構図になります。

所謂「Science 2.0」の世界ともいえるかもしれません。

大きな方向感としては、科学研究が IT プラットフォームに載っていく流れです。みなさんはなぜアマゾンのクラウドサービスを利用するのでしょうか。そこへ行けば必要なものがすべてそろっており、アマゾンが提供する環境が便利で豊かだからです。研究を始めるのに、いちいちコンピュータを導入して自分でプログラムを書くのではなく、なるべく先人が開発したソフトウェアなどを利用した方が速いし効率的です。

滝　そうした方向感は科学界で広く共有されていますか。

喜連川　生命科学のゲノム研究ではすでに一般的です。解読された塩基配列データは共有され、研究者は自らの競争力のコアは何かをしっかり認識したうえで、公開されたデータから必要なもの、いいものをどんどんとってきて活用しています。グーグルがディープラーニングのライブラリをオープンにしたのも似ていますが、こちらは最先端の研究者を引き付け、グーグルの方法論を広げようとする意図があります。誰かがライブラリを公開して「さあ、どうぞ使ってください」というやり方もありますが、みんなで同じプラットフォームに載っけ合いましょうという時代が来ると思います。NII はそこを目指します。

滝　研究支援の IT プラットフォームを提供する動きは大手出版など民間にもあります。

喜連川　そうしたサービスが商業的にペイできるのは、産業に近く、研究者間の競争が激しい領域でしょう。研究費が潤沢で商用サービスが成り立つ。私の個人的な思いとしては、商用サービスがあまり興味を示さない領域の支援にも力を入れたい。また異分野の融合領域は研究が「沸騰」しやすい。NII は大学の共同利用機関としては多様なプレイヤーとつきあうことが多いので、その持ち味を生かして融合領域の支援にも努めていきます。

（写真＝川本聖哉）

インタビュアーからのひとこと

日米欧などの研究者が協力したヒトゲノム解読計画では読み取った塩基配列データが共有された。巨大加速器や大型天体望遠鏡などの実験・観測データの共有化は以前から進んでいると聞く。データの共有化が大きな潮流であることは間違いない。

他方、すべての領域で野放図に共有化が進むとも思えない。研究者や企業、国家間の競争が激しい領域では話は単純ではない。守るべきデータは存在する。欧米の論文誌に投稿すると査読段階で情報が漏れるとの苦情や不安をしばしば耳にしてきた。似た状況がデータの世界で生ずるのは避けたい。それには日本が公開のルールづくりで積極的に発言し貢献することが必要だろう。

滝 順一 TAKI Junichi

日本経済新聞社 論説委員 兼 経済解説部編集委員。早稲田大学政治経済学部卒業後、日本経済新聞社入社。産業部（現企業報道部）、ワシントン支局、大阪本社経済部編集委員、東京本社科学技術部長などを経て、2009 年 3 月から論説委員。科学技術や環境、医学などを担当する。

オープンデータの最前線

“データのWeb”を実現するLODとDOI

武田英明
［国立情報学研究所 情報学プリンシプル研究系 教授／総合研究大学院大学 複合科学研究科 教授］

世界中に存在する公開可能な論文や研究データを即座に探し出し、自由に連携、利用できるようにするオープンデータ。このオープンデータは、どのような仕組みによって実現されているのか。また、さらなる利便性を確保していくため、現在、どのような取り組みが関係機関によって進められているのか。本研究課題の専門家として海外の状況にも詳しい武田英明教授に聞いた。

“データのWeb”を実現するLOD

オープンサイエンスを推進していくためには、「誰でも自由に使えて再利用もでき、かつ、再配布できるようなデータ」、すなわち“オープンデータ”の仕組みが重要だ。近年、注目を集め、活用され始めているのが「LOD（Linked Open Data）」。武田教授は「LODは、コンピュータ処理を目的に、データをはじめ、公開者や公開日などのメタデータを構造化し、異なるデータが相互に結びつくことを可能にします。いわば“データのWeb”の実現を目指して誕生したものです」と説明する。

LODは、自治体や企業、団体など各情報発信主体が標準フォーマットに従ってデータを公開することでデータを相互にリンクさせ、Web自体を巨大なデータベースとして機能させるという構想のもと、欧米を中心に広がりを見せてきた。バイオサイエンス系研究機関と企業による実験データの共有や、図書館での書誌や典拠のデータベース化、自治体による地域統計情報の提供などを皮切りにさまざまな分野でLOD化が進み、データの利活用が行われている。

武田英明
TAKEDA Hideaki

「端的な例を挙げると、世界中で出版されて図書館に蔵書されている夏目漱石に関する書誌や典拠がすべてつながり、すぐに探して利用できるようになりました」（武田教授）

必要なデータを利用する環境が整備

LODでは、機械が処理可能なWebリソース情報を表すための表現方法「Resource Description Framework（RDF）」、および、検索のためのコンピュータ言語SPARQL（スパークル）が標準化されている。RDFに基づいて世界中のデータベースに登録された情報を、SPARQLで記述されたアプリケーションを用いて取得し、活用するのだ。

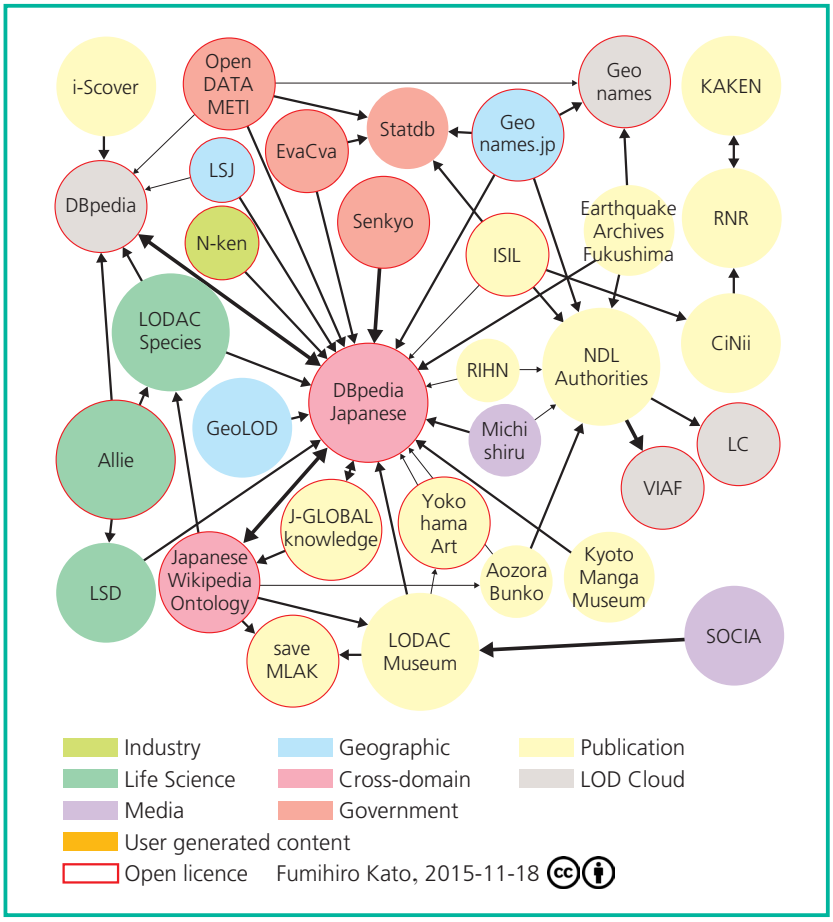
現在、各国の政府や自治体によって立ち上げられたポータルサイトで行政情報や公共データが公開されているほか、「the Datahub」などのWebサイトで世界中のさまざまなデータセットのカタログ化が行われ、データの取得が可能となっている。加えて「Linked Open Vocabularies（LOV）」などのWebサイトでは、RDFに基づいて構成されるデータの項目を定義した「スキーマ」が提供されており、これを用いることで共通化されたデータベースの構築が可能だ。LODを処理するツールやライブラリも広く提供されており、LODの収集から活用までのシステムを比較的容易に組めるようになっている。

LODの活用で先行する欧米では、Wikipediaから情報を抽出してLODとして公開する「DBpedia」と呼ばれるコミュニティプロジェクトが普及してい

る。だが、その情報は英語であり、日本からの登録や利用には障壁もあった。そこで、NIIによって2012年5月に公開されたのが「DBpedia Japanese」だ（図）。

この DBpedia Japanese は武田教授が進める LODAC Project の一つとして行われている。

「DBpedia Japanese の目的は、Wikipedia 日本語版を対象とした DBpedia の提供です。LODAC Project ではこのほか、ばらばらであってもつながるという LOD の特徴を活かして、日本国内の博物館、美術館の収蔵品情報を LOD 化し、日本最大の収蔵品データベースや生物多様性情報のための生物種情報のデータベースを構築しました」と武田教授。そのほかにも、共通語彙基盤上でのデータの収集、公開の仕組みを着々と整えつつある。



図｜日本のリンクトデータクラウド

論文の電子化で生まれたDOI

オープンデータの活用に向けた取り組みには、「デジタルオブジェクト識別子（Digital Object Identifier：DOI）」もある。DOI とは、学術論文に識別子を付加するとともに、論文の URL と、公開日、公開者などが判別できるようなメタデータを登録することでインターネット上のデジタルオブジェクトに持続的にアクセス可能とする技術だ。

「DOI は学術論文誌が電子化され始めた 1990 年代に、出版社によって共同で考案されました。電子化された論文の所在を URL で記した場合、Web サイトのリニューアルなどで URL が変更されるとアクセスできなくなるケースがあります。そこで、URL とは別に論文自体にユニークな ID を付加することで、URL の変更にも対応可能にしたわけです」（武田教授）

DOI を論文に付与することで所在が常にわかるようになったほか、引用文献の同定も容易となる。現在、世界最大の DOI 登録機関である米国の CrossRef では、全世界 7040 万報以上の学術論文に DOI を付与・登録し、引用・被引用文献へのリンクを実施。いまや、研究に不可欠な共通基盤として活用されている。

一方、日本から登録される情報は、言語の障壁などもあり、150 万報程度に留まっていた。そこで、DOI の普及と日本語による学術コンテンツへのアクセスと利便性向上を目指し、国立研究開発法人科学技術振興機構（JST）、国立研究開発法人 物質・材料研究機構（NIMS）、国立国会図書館（NDL）、そして NII によって「ジャパンリンクセンター（JaLC）」が設立された。国内の学術コンテンツを扱う各機関の参加も求め、DOI の普及や国内外情報サービスの利便性向上に向けた取り組みが進められている。

「近年では、DOI を論文だけでなく、研究データにも付加することでオープンサイエンスの有望なインフラにしようという研究が行われています」と武田教授は言う。JaLC では国内研究機関などとともに DOI データの登録実験プロジェクトを実施。今後のシステム構築や運用における課題を抽出し、データ DOI の本格的な活用に向けた取り組みを推進している。

「LOD が誰もが公開可能なデータであるのに対して、DOI はデータの出自が明らかで信頼性がある程度保証されたものです。両者のデータに互換性を与え、自由に連携させれば、さらに研究活動に広がりをもたせられる。実現に向けて、各分野の方々と協調しながら一つひとつ課題を解決していきたいと考えています」と武田教授は話す。

「今後、オープンデータは研究開発のスピードアップにとどまらず、社会の仕組み自体を変えていくでしょう。データのみならず、やがてはデータを生み出した人同士が直接つながってコラボレーションが実現し、新しいイノベーションが創出されるようになる。そうなると、今までの企業や組織という枠組みのあり方も激変していくかもしれません」

（取材・文＝伊藤秀樹　写真＝佐藤祐介）

機関リポジトリから オープンサイエンスへ

山地一禎
［国立情報学研究所 学術リポジトリ推進室／コンテンツ科学研究系 准教授］

×

北本朝展
［国立情報学研究所 コンテンツ科学研究系 准教授／総合研究大学院大学 複合科学研究科 准教授］

教育研究機関の知的生産物を収集・保存して発信するための電子アーカイブシステム「学術機関リポジトリ」や電子リソースを利用する大学などで構成される連合体「学術認証フェデレーション（学認）」の運営に関わる山地一禎准教授と、地球環境データの大規模データベースや国文学研究資料館の古典籍データベース、東洋文庫の貴重書デジタルアーカイブに関する研究プロジェクトを手掛ける北本朝展准教授の２人が、オープンサイエンスの発展について現状と課題、展望について語り合った。

評価基盤のカギは信頼

――オープンサイエンスは非常に広い概念。

北本　同床異夢と言うべきか、オープンサイエンスは人によって見方が違います。現状では、市民科学やオープンアクセス、オープンデータ、コラボレーションやクラウドファンディング、これらすべてがオープンサイエンスと呼ばれています。私は、オープンサイエンスとは研究に関する研究、メタ研究だと思っています。人によってさまざまなオープンの仕方があって、なぜオープンにするのか目的が違う。

山地　そう、現状はまだ定義づけできていないんです。皆、「何かある」と思っているけれど、大成功した人もいない。私自身はオープンリサーチデータの観点から、NII において「学術機関リポジトリ」のシステムをオープンソースで開発し、そのクラウドサービスを各機関に提供しています。日本は、機関リポジトリの数は 500 以上あり世界第１位を誇りますが、そのうちの200以上が NII のリポジトリモジュール「WEKO」※を使っている。今まで学術機関は論文メインで運用してきましたが、論文は最終的なアウトプット、あるいは次の研究のきっかけでしかなかった。ところが、オープンサイエンスではプロセスを自分の研究に取り込むことがより簡単にできるようになる。ポイントは、データをオープンにするだけでなく、データを誰が何に活用したのかというクローズな部分まで、どのようにインフラとしてサポートしていくかです。

　そうした中で、難しいインフラ構築に取り組めるのは NII ならでは。しかも、中立的な公的機関が作っていて、使い勝手がいいからこそ評価も得られる。実際には要望や問題点の指摘も多いのですが、それだけコンタクトが多いということでもある。そうした利用者とのコミュニケーションはさらに信頼を生んで不可欠なインフラとなり、ブランドになっています。研究所が運営を行うメリットは、運用しながら常に新しいものを開発できることですね。

北本　一方で、基盤に対する貢献をどう評価するかが課題です。私が運用している「デジタル台風」という気象データベースでも、長期的に更新を続けていることが利用者からの信頼につながっている。ところが、もし論文を書くことだけが目的なら更新を続けることは評価の対象にはならない、という問題があります。データ基盤をオープンに運用して研究コミュニティに貢献する活動をどう評価するかは、オープンサイエンスにおける一つの重要な課題です。

　また、オープンとクローズをどう組み合わせて価値を生み出すのかも課題で

山地一禎
YAMAJI Kazutsuna

す。研究成果を共有することでコラボレーションが生まれ、データのリユースにより研究コストも下がる。いずれも、オンラインでつながるということが前提になっています。

山地　IT環境で、どう簡単、便利に研究を加速化できるか。それこそが、NIIの役割の一つです。とくに、クラウド環境は人文・社会科学系にはまだ普及していません。その利便性を広めたい。

北本　人文系では個人研究が多いという点も理由でしょう。そもそもコラボレーションして研究することが今までは少なかった。でもこれからは人文系でも、個人研究に限界が生じることが増えるでしょう。ましてや、デジタル時代の人文学のあり方を研究する研究領域「デジタル人文学」ではコラボレーションが必須です。

山地　オープンにすると他の人の目に触れます。枠組みがあればデータのリユースから新しい研究のサイクルが始まる。新たな研究の発火材料にもなるのです。

予算はどこから捻出すべきか

――マイナス面は？

山地　データを出した人が正しく評価されるのかという問題はあります。

北本　デジタルオブジェクト識別子（DOI）とその登録機関であるジャパンリンクセンター（JaLC）では、研究データに識別子を付与する活動を推進しています。データに識別子が付与されてデータの作者が明示されれば、その貢献を評価することも可能になります。これは論文における著者役割の明示という話題とも関係するでしょう。今までは論文に「著者」というカテゴリーしかなかったため、多くの人々が関与するビッグサイエンスでは著者が1000人も並ぶことがありました。最近は単に論文著者というのではなく、もっと研究への貢献を細分化して明示する方向に進んでいます。ただ、研究者の貢献が計測可能となると、それが一人歩きして意図しない使われ方をする危険もありますが……。

山地　そうした歪みが、むしろ、ドライビングフォースになるかもしれません。今は公開することで透明性を担保する時代です。

北本　そこは意見が分かれるところかもしれません。データを保全して不正をチェックできるようにすることは切実な課題ですし、予算を出す側にもわかりやすい。ですが、オープンサイエンスの目的を不正防止としてしまうと、あまり価値を生まないのでは？

山地　僕の見方は逆です。インフラは核であり、費用がかかる。透明性の担保が出資者に響くのであれば、その点をうまくアピールすればいい。

北本　オープンサイエンスにはいろいろな対立軸がある、ということですね。だからこそ、全体像を踏まえた議論が必要なのです。

北本朝展
KITAMOTO Asanobu

定着と信頼の輪

――今後の展望は。

北本　今の研究のやり方にはいろいろな問題が生じています。それをよい方向に変えていくためのドライバーとなるのが、インターネットです。オープンサイエンスという考え方が出てきた背景には、インターネットの活用によるオープン性の追究が不十分ではないか、という考えがあるのだと思います。

山地　ボトムアップで研究者が何を出していけるかがカギです。我々としては、研究データやラボノートを公開することが得になる環境を作ることが使命と思っています。素材自体はそろっています。データやクラウド基盤をつなげる仕組み、認証の仕組みもリポジトリもある。つなげば、何か生まれるかもしれない。

北本　実際にデータを作って公開すると、意外なところからコンタクトがあります。長期的な投資と考えればメリットはある。ただし、長期的にデータを作ることができる立場の人や機関がやらないと、投資を回収するのは難しいかもしれない。

山地　でも、データを公開すれば、誰かが見つけてくれる。

北本　時間はかかりますけどね。

山地　だからこそ、長期的にインフラ構築や運用ができる組織としてのNIIの意味は大きいということでしょう。

北本　5～10年続けないと、信頼は得られません。基盤というのはそういうものです。

山地　本当にサービスしている人じゃないと、この面白さは見えてこないんだろうなぁ(笑)。苦労はありますが、サービスが全国に広がっていく快感もあるのです。

北本　ぜひこの面白さを、NIIとともに体験してほしいですね。

（構成＝森山和道　写真＝土佐麻理子）

※WEKO
学術成果を保存・公開することを目的にNIIが開発しているNetCommons2上で動作するリポジトリシステム。「WEKO」はスワヒリ語でリポジトリ（貯蔵庫）のこと。

実環境データを情報学研究に活かす

「データセット共同利用研究開発センター（DSC）」の役割

大山敬三

［国立情報学研究所 データセット共同利用研究開発センター長・コンテンツ科学研究系 教授／
総合研究大学院大学 複合科学研究科 教授・情報学専攻長］

ディープラーニングなどの人工知能技術やビッグデータ処理技術は近年、産業応用が加速している。その一方で学術研究にも早期実用化や産業への応用が社会的に強く求められるようになり、実社会で生まれるデータを用いることがより重要になっている。そこで実環境で蓄積された大規模データを情報学研究に活かす使命を担って設立されたのが、NII の「データセット共同利用研究開発センター（DSC）」だ。センター長の大山敬三教授が、研究者と産業界をつなぎ、知的財産やプライバシーの保護という課題にも取り組みながら、研究資源であるデータの受け入れと提供を行うセンターの意義について語った。

大山敬三
OYAMA Keizo

研究者と提供企業双方にメリット

近年、大規模データ処理技術は、新しいビジネス創出やサービスの高度化に欠かせないものとして、その研究成果の早期実用化が強く求められています。とくに統計的機械学習やビッグデータ解析などの研究分野には、大きな期待が寄せられています。この社会的要請に対応するためには、従来のような研究者による研究用の手作りデータでは不十分であり、実社会から得られた大規模な実データの入手が不可欠です。

一方、産業界ではネットビジネス企業が本格的な研究組織をスタートさせたり、先端技術を持つベンチャー企業が市場に足場を確保したりしているように、最新・最適な技術を採り入れることが競争力の源泉になっています。しかし、自組織内の研究だけでは不十分であり、自社のデータを提供してでも大学などの公的研究機関と共同研究したいという企業が増えてきました。専門的な研究を行う大学院生らにデータを提供することで、自社への関心を高めてもらい、優秀な人材の確保につなげたいという狙いもあります。

このように研究者と企業の利害が一致する面はあるのですが、実際に研究者が企業と個別に交渉するのは難しい。また、企業のデータには機密情報、著作権、プライバシー保護など多くの制約があり、個別に複雑な利用条件を調整してデータを準備しなければなりません。そこでDSCが双方の間に立ち、企業などからデータを受け入れて、一定のルールに基づいて研究者に提供する役割を果たしているのです。

続々と出てきた研究成果

DSC設立のそもそものきっかけは、NIIが1997年末にスタートさせた評価型ワークショップである「NTCIR（NII Testbeds and Community for Information access Research、エンティサイル）」のために提供されたヤフー株式会社の「Yahoo！知恵袋データセット」です。このデータはワークショップ以外の研究目的でも多くの大学や企業の研究機関に提供され、さまざまな研究が行われました。この成果が注目され、NIIを通じて研究者にデータを提供したいと手を挙げる企業が徐々に増えてきました。そこでNIIは2010年にデータの受け入れと提供を行う窓口として「情報学研究データリポジトリ（IDR）」を設け、さらにデータの共有と活用を進めるために、2015年4月にDSCを設置しました。DSCはNTCIRの運営、IDRの窓口、およびNIIの「音声資源コンソーシアム（SRC）」の活動を統合し、研究資源としてのデータを核としたオープンサイエンスの推進を目指しています。

DSCでは現在までに民間企業6社から14種、国文学研究資料館から1種の

表1 | DSCが提供しているデータセット

提供組織	提供データセット
ヤフー株式会社	Yahoo！知恵袋データ（第２版）
楽天株式会社	楽天市場の全商品データ、レビューデータ 楽天トラベルの施設データ、レビューデータ 楽天ゴルフの施設データ、レビューデータ 楽天レシピのレシピ情報、レシピ画像 楽天オークションの評価コメント情報、取引情報 アノテーション付きデータ 楽天 Viki のビデオ情報、ユーザ情報
株式会社ドワンゴおよび有限会社未来検索ブラジル	ニコニコ動画コメント等データ ニコニコ大百科データ
株式会社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
クックパッド株式会社	レシピデータ 献立データ
株式会社ネクスト	賃貸物件データ、画像データ（HOME'S サイトのデータ）
人間文化研究機構 国文学研究資料館	古典籍データ（書誌・画像・タグ・本文テキスト）
NII	NTCIR テストコレクション 音声コーパス 会話コーパス（準備中。音声・映像データを含む）

（注）2016 年 1 月 12 日現在　（出典）http://www.nii.ac.jp/dsc/idr/datalist.html

データセットの提供を受け、NTCIR のテストコレクションや SRC の音声コーパスなど数十種を加えて、情報学や関連諸分野の研究者に無償で提供しています（表 1、図 1）。データはテキストのほか画像、音声、映像も含んでおり、一部を除きインターネットからダウンロードして利用できます。

データセットの利用状況を見ると、個別にデータを提供していた 2007 年から利用者数は順調に伸びています（図 2）。また、データセットを利用した研究成果の論文は 2014 年末時点で 350 本、利用研究室は 2015 年 11 月末時点で 468 と、どちらも増加傾向が加速しています。

研究成果は実に多様で、たとえば、料理レシピデータを自動的に解釈し、複数の作業を並行して行うフロー図を作成する研究、Q&A データから問題に対する最適解決策を求める研究、動画へのコメントデータから楽曲のサビを推測する研究などが出てきています。

情報の保護をクラウド化で解決

これからますます実用性の高い研究が出てくることを期待していますが、そのためには一層多様なデータが必要になります。そこでの一番大きな課題は、データに含まれている可能性がある潜在的なプライバシーや機密情報をどう保護するかという点です。たとえば、提供データセット自体には個人情報が含まれていなくても、他のデータと突き合わせると探り当てられてしまうことがあります。実際に、米国のAOLが検索クエリデータを公開したところ、利用者が特定されてプライバシーが暴露されるという事件が起こりました。これが今でも企業にデータ提供をためらわせる一因となっています。

IDR では現在のところ、「覚書」を交わすなどによって決められた利用上の条件を守ってもらうようにしていますが、いずれはクラウド上で安全にデータを利用できる仕組みを導入したいと考えています。提供方法としては、ダウンロード禁止などの利用制限つきで提供する、API を通して統計処理した結果だけを返す、利用者がプログラムを作成・登録してクラウド上で実行することによりプログラムからのみデータアクセスできるようにするなど、さまざまなバリエーションを検討しています。

データを守りつつ活用を促進するためには、クラウド利用を前提とすることによって可能となる技術を使い、企業が求める安全性と研究者が望む利用方法との現実的な折り合いをつけることが不可欠でしょう。それに向けて現在は企業との共同研究を進めている最中で、来年度中には具体的な結果を出していきたいと考えています。

（構成・文＝土肥正弘　写真＝佐藤祐介）

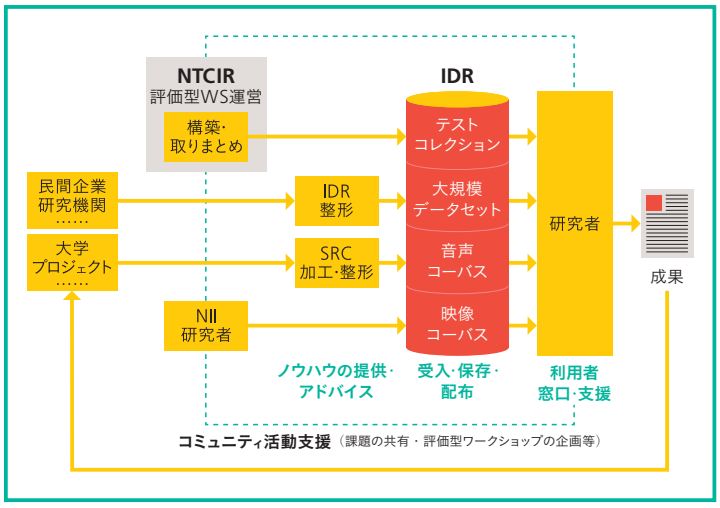


図1 | DSCのデータセット提供にかかわる活動

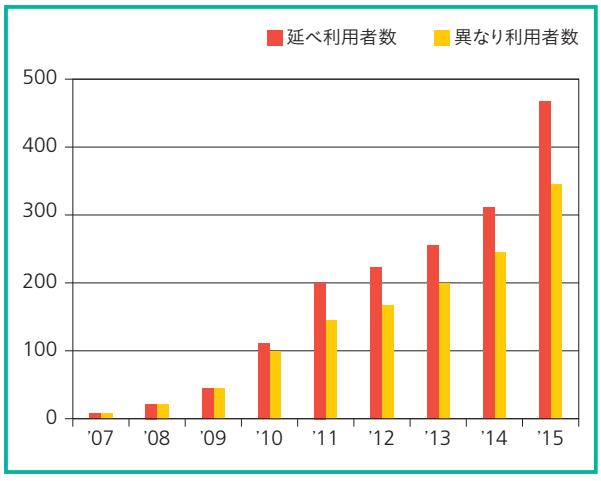


図2 | データセットの累積利用者数の推移
（民間企業提供データセット。ただしニコニコデータセットを除く）

News 1 オープンデータめぐり議論 ～情報・システム研究機構がシンポジウム開催

情報・システム研究機構はシンポジウム「オープンサイエンスにおける研究データのオープン化」を２月８日に開催しました。

内閣府「オープンサイエンスに関する検討会」の有川節夫座長（前九州大学総長）は基調講演「オープン化による新たなサイエンスの展開」で、国際的な潮流であるオープンサイエンスの必要性や課題を論じました。そして、「自分の論文や使用データは今日からでも公開できる。まずは大学から始めよう」と呼びかけ、共有リポジトリサービス JAIRO Cloud の利用を提案しました。

続いて、情報通信研究機構統合データシステム研究開発室の村山泰啓室長が「極域科学とオープンデータ」、東京大学の高木利久教授が「生命科学とオープンデータ」をテーマにそれぞれ講演しました。また、Nature Publishing Group オープンリサーチマーケティングマネージャーの新谷洋子氏は研究データ共有をサービスに展開した「Scientific Data」を紹介し、データ公開に対するインセンティブ強化の重要性を訴えました。

後半の討論＝写真＝では統計数理研究所の丸山宏教授と山下智志教授、国立極地研究所の伊村智教授、国立遺伝学研究所の小出剛准教授、ライフサイエンス統合データベースセンターの箕輪真理特任准教授、NII コンテンツ科学研究系の北本朝展准教授が「研究現場におけるオープンデータの進め方」について議論。「論文／データの区別ではなく、サイエンスへの貢献度が評価されるべきだ」「論文のエビデンスにならない失敗データも公開することで、斬新な研究の可能性が生まれる」「公開しやすい環境の整備や意識づけが重要だ」など、今後の方向性を示しました。

News 2 今年度最後の産官学連携塾「質感研究」の最前線伝える

今年度の最終回となる「第５回産官学連携塾」を１月 22 日に開催しました。「質感研究の発展」と題し、コンピュータビジョンを専門とするコンテンツ科学研究系の佐藤いまり教授＝写真＝が講師を務めました。

講義ではまず分光特性の解析による「ものの見え方」を説明。光源の違いや反射のあるなしで見え方がまったく異なることを示し、少ないサンプルで物体の再現を可能にするサンプリング手法を紹介しながら、画像から物体を安定して推定する方法についての研究の進捗状況を解説しました。また CG を例に、質感研究の成果がクリエイティブの分野で活用されている状況も説明しました。

佐藤教授は「産業界の方も可能な範囲で開発研究を共有していただければ、我々研究者も研究テーマが広がり、相互発展できる」と、積極的な情報交換を呼びかけました。

News 3 ビッグデータでバブルの正体探る ～第5回NII湘南会議記念講演会

情報学分野の世界トップレベルの研究者が一堂に会して現在の未解決問題を議論し、解決を図ることで、情報学の進展を目指す「NII 湘南会議」。そのアウトリーチ活動として、「第５回 NII 湘南会議記念講演会」を 12 月 13 日に開催しました。主題は「経済の『今』を知る――ビッグデータで探るバブルの正体」。経済物理学が専門の情報社会相関研究系、水野貴之准教授が講師を務めました。

水野准教授は「バブルのキーワードは“格差”」と指摘し、例として不動産データを挙げました。通常時には物件の広さと価格は概ね比例するのに対して、バブル期には同じ条件でも特定の物件価格が投機目的で高騰して「ばらつき」が出ると説明。株価の場合も、ビッグデータ解析で銘柄間の格差を監視することにより、株価上昇時にそれがバブルなのか持続的経済成長なのかを判断できると述べました。

また、最近の研究としてニュースや Twitter による景気観測、経済ネットワークを介したグローバル金融危機の予測なども紹介。身近な問題を最新の研究で解決しようとする内容に、参加者は耳を傾けていました。

Flash

▶ 第4回 SPARC Japan セミナー 2015

３月９日開催。「研究振興の文脈における大学図書館の機能」がテーマ。星子奈美氏（九州大学附属図書館）の司会で、尾城孝一氏（東京大学附属図書館）、引原隆士氏（京都大学図書館機構長）、真子博氏（内閣府）、有川節夫氏（前九州大学総長）がそれぞれオープンアクセスやオープンサイエンスと大学図書館の役割などについて講演。

後半では市古みどり氏（慶應義塾大学日吉メディアセンター）をモデレーターに、日本の研究力向上に大学図書館がどう寄与できるかを主題にパネルディスカッションを実施。

Topics

産学連携で2研究施設新設、イノベーション広げる拠点に
金融スマートデータとコグニティブ・テクノロジーが主眼

産学連携を推進するNIIは、2月1日付で2研究施設を設置しました。2月9日の記者会見で発表したのは、三井住友アセットマネジメント株式会社（SMAM）と共同で設置した「金融スマートデータ研究センター」。その6日後には「コグニティブ・イノベーションセンター」の新設も発表しました。同センターでの研究は日本アイ・ビー・エム株式会社（日本IBM）の支援を受けます。

研究施設とは特定分野の研究に専念する研究部門で、両センターの設置でNIIの研究施設の数は「11」になりました。ともに目的は研究成果を社会に還元することであり、特定の技術力を強化するのではなく社会におけるイノベーションの根を広げていくことを狙っています。

金融スマートデータ研究センターのセンター長は、喜連川優NII所長。情報・システム研究機構が2月に導入したばかりの、公益性が高い研究部門を民間経費で設置・運用する「共同研究部門制度」を利用しました。NIIが民間経費で研究施設を設置するのは初となります。

「金融スマートデータ」とは、そのままでは巨大で複雑なデータの集積物に過ぎないビッグデータを処理・分析し、新たな価値の創出につながる有益な知識へと変えたものです。本センターでは、金融スマートデータを活用して経済・社会現象の法則の解明に挑み、長期的な「未来予測」の実現、ひいては国内金融市場の活性化や国民の安定的な資産形成といった社会的使命を果たすことを目指します。

金融スマートデータ研究センターの共同設置を発表するSMAMの横山邦男社長（左）と喜連川所長


コグニティブ・イノベーションセンターの新設記者会見で石塚センター長、喜連川所長、日本IBMのキャメロン・アート氏（右から）

一方、コグニティブ・イノベーションセンターのセンター長には、元人工知能学会会長の石塚満氏（早稲田大学教授、東京大学名誉教授）を招聘しました。中心テーマである「コグニティブ・テクノロジー」とは、機械学習や自然言語の処理と理解、ビッグデータや知識ベースの構築と利用など知的情報処理の集合体。ディープラーニングなどの最新の人工知能技術にとどまらず、先端的情報技術を幅広く活用し、ビッグデータから学習して自然なインタラクションの中で人間の認知や判断を支援する面に主眼を置いています。

本センターの活動には、幅広い業界から日本を代表する多くの企業が参画予定。コグニティブ・テクノロジーの社会応用促進に向けた意識変革、最先端技術と産業の新たな結びつきの発見という二つのイノベーションを起こすことが目的です。

SNS

「これ、いいね！」

Facebook、Twitterアカウントの最も注目を集めた記事（2015年12月～2016年2月）

国立情報学研究所NII（公式） Facebook
www.facebook.com/jouhouken/

［Hi ！ from Bit-kun］LOVE びっと
今日はバレンタインデーということで！
みなさんに、とっても大きな愛をお届けびっと！
（Robert Indiana “LOVE” 1993 @Shinjuku I Land）（2016/02/14）

国立情報学研究所NII（公式） Twitter
@jouhouken

［NII NEWS］秋葉拓哉助教が平成27年度船井研究奨励賞を受賞（2016/02/28）

つぶやくビット君 Twitter
@NII_Bit

元日に放送された#jwave JAM THE WORLD ニューイヤースペシャルで新井紀子教授と津田大介氏 @tsuda が対談した詳細が同氏のメルマガに掲載されていますびっと（2016/01/27）

Essay

科学者がかっこいい社会

相澤彰子 AIZAWA Akiko
［国立情報学研究所
コンテンツ科学研究系　教授］

最近、「好きなSFは何ですか？」という質問を受けた。中学・高校時代に愛読したSFを思い出しながら、そういえば、SFの定義は何だろうか、と考えた。諸説あるようだが、「SF」の「S」はサイエンス（＝科学）であろう。そうすると、「SF」とは科学が進歩した未来世界を描いたフィクションのことだろうか？　あるいは、科学が小道具になっている小説か。このようなあいまいさは、「オープンサイエンス」の定義の難しさにも通じる。科学への期待と夢想が入り混じり、その境界はなかなか明確に定まらない。

オープンサイエンスは広範な概念を含んでいる。その一端を要約するなら、「科学を加速させるための革新的インフラ」となるだろうか。その中には、それぞれの分野が抱える構造的な問題への解決が含まれる。科学の各分野における阻害要因は多種多様であるから、オープンサイエンスが目指すところは、学術雑誌の出版コスト削減から、データへのID付与や引用、永続的アーカイブ構築、市民科学の推進まで多岐にわたる。これらが大きな動きとなり、誰もが参加できる開かれたサイエンスが実現すれば、それが大きなイノベーションへと結びつく。

しかし、オープンサイエンスという言葉は、まだ歩き始めたばかりである。たとえば「オープンサイエンス」で画像検索をして、「ビッグデータ」や「クラウドソーシング」などと結果を見比べると、違いがよくわかる。検索される画像は文字満載のスライドが大半で、共通のビジュアルなイメージというものは見当たらない。確かに、目下のところ、「オープンサイエンス」はニュースに頻出するキーワードではないし、日常生活に密着している感じも少ない。言ってみれば、現状では抽象的かつ特殊なギョーカイ用語である。

SFの話に戻ろう。筆者にとってのSFの定義は、「かっこいい科学者が登場する」ことである。頭脳明晰であり、優れた情報分析力と的確な状況判断で難問に立ち向かうヒーローやヒロインは、心底かっこいい。オープンサイエンスが示すのは、科学者が活躍する世界ではないだろうか。だから、画像検索で出てくるイメージは、颯爽としたかっこいい科学者であって欲しい。

今後の予定

5月25日～27日｜国立情報学研究所 学術情報基盤オープンフォーラム2016＝一橋講堂ほか

5月27日～28日｜国立情報学研究所 オープンハウス2016（研究成果発表・一般公開）＝一橋講堂ほか。詳細や事前登録が必要なイベントへの参加申し込みは、以下のURLで。
http://www.nii.ac.jp/openhouse/

6月22日｜平成28年度 市民講座「情報学最前線」第1回（講師：情報学プリンシプル研究系 秋葉拓哉 助教）＝国立情報学研究所の研究者が情報学の先端を一般向けに解説する年6回のプログラム。日程や各回のテーマなど平成28年度の詳細は、決定次第、以下のURLでお知らせします。
http://www.nii.ac.jp/event/shimin/

表紙の言葉　実験データや画像などの情報を共有することで実現されるオープンサイエンスの世界を、それぞれ離れた場所にいて実験をしているロボットを描くことで表現しました。集合知がもたらす新しい科学のあり方を示唆しています。

情報から知を紡ぎだす。
NII
国立情報学研究所ニュース［NII Today］第71号 平成28年3月
発行｜大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
〒101-8430 東京都千代田区一ツ橋2丁目1番2号 学術総合センター
発行人｜喜連川 優　監修｜佐藤一郎
表紙画｜城谷俊也　編集｜田井中麻都佳
制作｜株式会社マツダオフィス／株式会社アテナ・ブレインズ
本誌についてのお問い合わせ｜総務部企画課 広報チーム
TEL｜03-4212-2164　FAX｜03-4212-2150　e-mail｜kouhou@nii.ac.jp

情報犬ビットくん（NIIキャラクター）

http://www.nii.ac.jp/about/publication/today/