［オープンサイエンスの動向と情報科学の役割］

①オープンサイエンスの成り立ちと学術コミュニケーションの未来

基般

武田英明 国立情報学研究所

オープンサイエンスとは

オープンサイエンスという言葉は2000年前後から頻繁に用いられるようになった．それはインターネットの社会浸透の時期と重なるが，この重なりは偶然ではない．オープンサイエンスといっても新しい別種の科学が生まれたわけではない．オープンサイエンスは新しい科学活動の形であり，近年のディジタル技術およびインターネット技術の発展に伴って，科学活動の実施の方法が変わることによって生まれてきた科学活動の新しい形を指している．

本稿ではまず，インターネットの発展のかかわりからオープンサイエンスとは何であるかを概観し，その上でこれからのオープンサイエンスに必要となる環境について議論する．それは学術研究のための新しい情報環境にほかならない．

オープンサイエンスの出自

オープンサイエンスにおけるオープン性とは

先に述べたように現在言われているオープンサイエンスはインターネットの発展と普及に伴って生まれた科学の在り方を指しており，単純化して言ってしまえば，オープンサイエンスの“オープン性”とはインターネットの“オープン性”に由来する．

ではこれまでの科学のオープン性とは関係ないのだろうか．科学はオープンであるという考え方は近代科学が始まったころからの科学の重要な性質である．科学の成果は広く公表され，その結果，社会に貢献するとともに，科学自体の発展に寄与する．ただ，ここで言う科学のオープン性は基本的には科学者の世界，科学者のコミュニティにおいてオープンであるということを指していて，インターネットのオープン性とは異なっている．実際，インターネットの普及期にはその差異による軋轢が少なからずあった．伝統的な方法によって「公開」された論文や著作物をインターネット（Web）に載せることで，思わぬ副作用が生じた．たとえば，医学系の書籍や論文には病気の症例が患者にかかわる情報とともに記述されることがあったが，このような記述は専門家にとっては有益でも，インターネットでの公表にはそぐわない．これらは科学のオープン性とインターネットのオープン性に差異があることの証左である．

では，今起こりつつあるオープンサイエンスは科学のオープン性とは無関係なのだろうか．実はそうでもない．それはインターネットのオープン性自身の出自にかかわっている．

一般には，インターネットはアメリカ国防総省のプロジェクトによって作られたとされるが，この技術開発にかかわった研究者が学術界（アカデミア）であったことは重要な点である．研究者が即時の商業性ではなく技術の発展可能性からIPを選択したことが今日のインターネットの発展に繋がった．その後もインターネットの開発は大学と研究機関の中で行われた．特に同時期にやはり大学と研究機関で開発が進んでいたUnixとの結合が大きな飛躍になった．このようにインターネットはアカデミアの

中で開発されることで，これまでの培われてきた科学のオープン性を知らず知らずに取り込んで形成されていった．特に Web（World Wide Web, WWW）の発明は顕著で，研究機関である CERN（欧州原子核研究機構）に所属していた Tim Berners-Lee が所内の研究者の間の情報共有を目的に開発したものが Web であり，そこにおいては情報がオープンであることが前提に設計されている．

オープンサイエンスの 4 つの発展形式

このようして生まれたインターネットのオープン性は科学に影響を与えるようになり，オープンサイエンスが生まれたわけだが，そのつながりは単純ではなく，主に 4 つのつながりがある（図 -1 参照）．

科学のインターネット化

コンピュータは当然のごとく，科学研究に活用されてきた．初期には観測記録のディジタル保存といったことから始まり，理論に基づくコンピュータ・シミュレーションが続いた．特に後者を経験科学，理論科学に続く第 3 の科学と呼ぶことがある．さらにはさまざまな観測機器のディジタル化の普及やディジタルデータの処理技術の向上に伴い，科学活動のさまざまな場面でコンピュータが使われるようになった．ここではデータが科学活動の対象であり，科学はデータを生成し，処理，分析することで新しい知見を得るという活動となった．これらは総称してデータ中心科学あるいは e- サイエンスなどと呼ばれ，第 4 の科学とも称される．科学者はインターネットの最初のユーザであり，インターネットの文化形成に影響を与えるとともに，逆に科学もインターネットの文化から影響を受けた．たとえば，インターネットではデータをオープンにすると，もっと容易に科学データの共有ができることが分かり，科学者はこの機能を活用してデータやプログラムの交換をオープンに行うようになった．こうして科学活動のディジタル化・インターネット化は科学自体をこれまでの科学とは違う意味でオープンにしていった．

オープンアクセス

第 2 のルートは大学や研究所の図書館を通じてのオープンサイエンスへの接近である．オープンサイエンスという言葉よりも先にオープンアクセスという言葉が浸透してきた．

科学者は研究も行うが，研究成果の発表も行う．これは主に学術雑誌への論文の発表という形をとる．第二次世界大戦後，学術雑誌の種類は右肩上がりに増え，また雑誌あたりの費用も同様に右肩上がりであった．学術雑誌は購入先のほとんどが大学・研究図書館で閉じた世界であり，いわゆる市場原理が働きづらい．一方，大学・研究図書館の予算というのはそれほど増えず，このため必要な学術雑誌を図書館が買えなくなるという事態が出てきた．これが 1980 年代にアメリカで顕著になり，シリアルズ・クライシスと呼ばれた．このような状

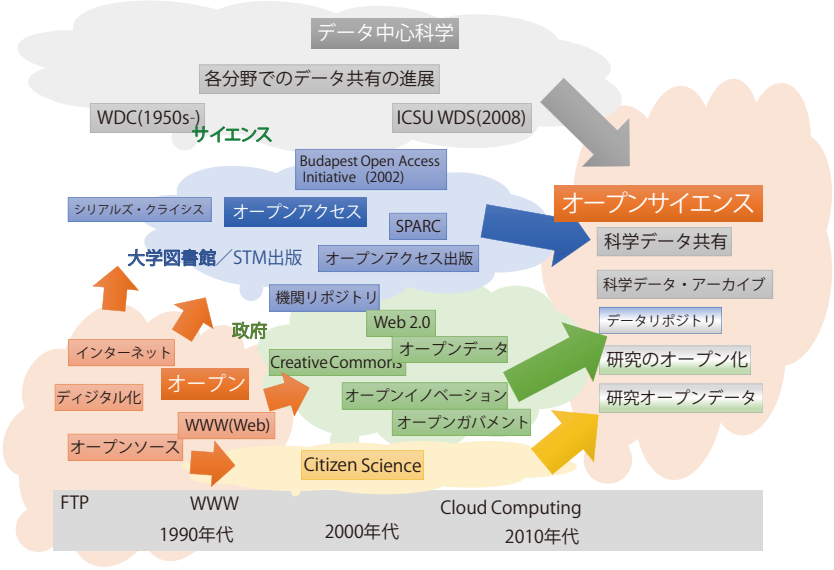

図 -1　4 つのつながり

況下で，インターネット／Webにおけるオープンソースソフトウェアにおけるオープンの概念を利用して，オープンアクセスという概念が生まれた．オープンアクセスは2002年にBudapest Open Access Initiativeの中で明確に定義されている．そのようなオープンアクセスの実現の一手段として，機関リポジトリが提案された．機関リポジトリとは，“大学がその構成員に提供する，大学やその構成員により作成されたディジタル資料を管理し発信するための一連のサービス”と位置付けられる．オープンアクセスの当初の目標は論文のオープン化であった．しかし，同じ仕組みは研究データのオープン化にも利用可能である．既存の機関リポジトリでデータ公開を行う，あるいは研究データ専用の機関リポジトリによるデータ公開がいま始まっている．

政府のオープンデータ

3番目のルートは政府のオープン化，オープンガバメントを通じた接近である．

Tim O'Reillyは2010年に「政府2.0」というキャッチフレーズを考案し，政府はインターネットへのサービス提供は極力行わず，データのみを提供し，民間がそのデータを活用することでサービスが実現するべきという方向性をまとめ上げた．ここにおいて公共部門情報の問題がインターネットのオープン性と積極的に交わり，その結果としてオープンデータという言葉が広く言及されるようになった．「オープンデータ」という言葉が最初に定着したのがこの政府のデータのオープン化であり，今でも狭義では政府のオープンデータを指す．2009年1月に米国大統領に就任したBarack Obama氏は就任直後にオープンガバメントの基本方針を提唱した．ここでは透明性，参加，協働の3つの原則が掲げられた．透明性はデータがオープンであることを求めている．これはまさにオープンデータのことである．この原則は元々は行政にかかわるデータやプロセスに関してであったが，政府が運営している研究機関や政府の資金を得て実施している研究プロジェクトも同じようにオープンであるべきだという形に展開していった．

インターネット化した市民科学

最後のルートは市民科学（citizen science）からの接近である．現代の科学は高度に専門化され，長い教育期間と専門的な研究経験が必要な分野が多い．とはいえ，科学は職業的科学者だけが行うものではない．生物分類学，天文学や歴史学等など，非職業的科学者いわゆるアマチュア科学者が活躍する分野もある．さらに，発掘調査や翻刻，翻訳など一部の分野では限られた知識と経験でも参加する科学活動があった．この場合でも参加できる活動分野が限定されており，また実際に参加するには時間や地域的制約があった．インターネットの発達によって，この制約が大きく減少して，新しい市民参加の道が開けてきた．これまでの市民参加と異なる点は，非常に多数の参加による集合知型の活動であることと，参加の形態もきわめて小さなタスクの実施でよいことが挙げられる．たとえば，Galaxy Zoo Projectではハッブル望遠鏡で撮影した銀河の画像からその回転の向きといった銀河の分類を市民が行うということを行っている．これらの場合，データもオープンであるが，誰でも参加できるという点でプロセスがオープンになっていることが重要である．

科学者／科学の世界と4つのオープンサイエンスの関係

上記の4つのルートは科学にあったさまざまな境界の壁を低くしてきている．それをもう少し詳しく見てみよう．科学を取り巻く世界は3つの重層的な世界として見ることができる（図-2参照）．1番中心は科学者の世界であり，科学者同士が明示的あるいは暗黙的な関係で繋がってできている世界である．その外には科学を支える組織があり，科学者の科学活動を支えている．大学や研究機関，学術出版社，学会，研究資金を提供する政府や資金提供団体である．これを科学の世界と呼ぼう．最後にそれを含む社会全体である．

1番目のインターネット化する科学は科学者の科学活動に関するもので，1番内側の科学者の世

界にかかわり，科学者間の境界や，所属組織の境界，科学分野の境界の壁といったものを低くしている．2番目のオープンアクセスは1番目を含む2番目の世界（科学の世界）にかかわるものである．これは科学活動にかかわる組織間の境界の壁を低くして，科学に関する情報や知識の流通を促進させている．3つ目のオープンガバメントからくるオープンデータと4つ目の市民科学は，最後の科学の世界と社会全体との境界の壁を低くしている．

総じていえば，オープンサイエンスとはさまざまな境界の壁を低くして，シームレスな科学の世界を目指しているものといえる．もちろんその原動力はコンピュータとインターネットの技術であり，これがこれまであった壁を低くするあるいは破壊することを可能にしている．

研究プロセスと研究データのオープン化

オープンサイエンスはさまざまな境界の壁を低くしてシームレスの科学の世界を目指すものであるが，何をシームレスにするのだろうか．それは大きくプロセスとデータに大別される．

科学のプロセスは閉じていた．すなわち，科学のプロセスは職業的科学者が大学などの研究機関で行うものであり，さらには特定のテーマは特定の科学者が特定の機関や学会で行うものであった．オープンサイエンスではこのような科学のプロセスにかかわる人や機関に対する制約が解かれ，さまざまな人や機関が参画でき，また科学プロセスもさまざまに組み合わされることが可能になる．この顕著な方向が市民科学への接近である（本特集江渡氏の記事参照）．この面でのオープン化はまだ始まったばかりである．

一方，データのオープン化は今，大きく動いている．そこで，以降では研究データのオープン化を中心に見ていくことにする．

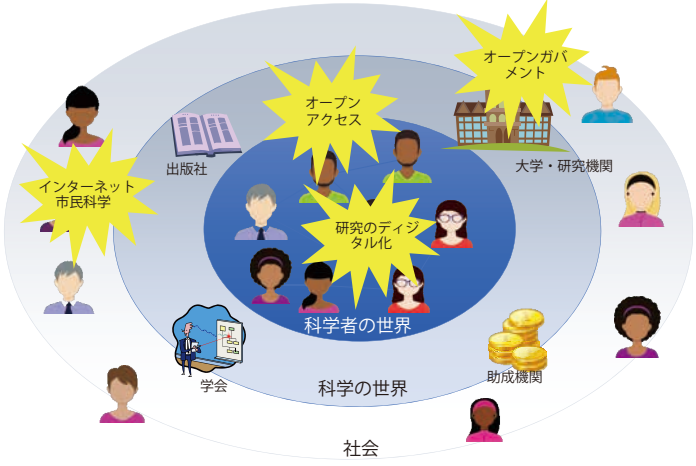

図-2　研究と社会とオープンサイエンス

オープンサイエンスのプラットフォーム化

前章での述べたように，オープンサイエンスとはさまざまな境界を低めるものであり，オープンサイエンスによって，これまでは個別に存在していた科学分野でのさまざまな活動がインターネット上で統合されようとしている．しかしインターネットおよびWebそのものがオープンサイエンスの活動に適した機能を提供しているわけではない．このため，インターネットを基盤として，その上にオープンサイエンスのためのプラットフォームが構築されつつある．

ジャーナルから脱却する学術コミュニケーション

20世紀における学術コミュニケーションは，ピア査読によって採否が決まる学術ジャーナルによって実現されてきた☆1．研究者・学者が学術成果があったとき，当然，その成果を（1）後世に記録したいし，（2）公表して関係者に広く知ってもらいたいし，（3）関係者に適切に評価されたい．学術

☆1 学術ジャーナルの誕生は，1660年代にフランスでのJournal des sçavans，イギリスでのThe Philosophical Transactions of the Royal Societyの発刊を端緒とするが，いわゆるピア査読による学術ジャーナルが一般化するのは遅く，18世紀半ばから編集者による査読が始まり，いまよくあるような外部査読者による査読は20世紀半ばからにすぎない．

ジャーナルはこの（1）記録，（2）公表，（3）評価，の3つの機能を持った便利かつ唯一の仕組みであった．

ところが，インターネットの普及は，（1）を容易にするとともに，（2）を大幅に代替する手段を提供した．学術ジャーナルは唯一の学術コミュニケーション手段ではなくなっている☆2．紙による公表からディジタルによる公表に変わることで，文章を基本とする論文以外のさまざまな研究成果，データ，プログラムコードといったものも公表可能になった．これは学術コミュニケーションの幅を大きく広げるものである．

一方，単に論文やデータをインターネットにおくだけで学術ジャーナルの代わりになるわけではない．学術ジャーナルにおいては，必要な情報の構造化がなされ，一定の内容のコントロールがなされ，適切な読者へ半永久的に届く仕組みを提供していた．そういった機能は新たに用意する必要がある．それがオープンサイエンスのためのデータ公開の原則であり，データ公開プラットフォームである．

データ公開の原則

ジャーナルに代わり，論文を含むすべての学術成果がインターネットを通じて公開共有されるようになることが期待されているが，それは単にデータをインターネットにおけばよいというわけではない．広く科学コミュニティや社会に利用可能でないといけない．これには現在，FAIR原則☆3というものが広く受け入れられている．その概要は以下の通りである．

1. 見つけられる（Findable）

データにはデータを説明するメタデータが用意され，グローバルで永続的な識別子で識別可能であるとともに，検索可能な形で公開されていないといけない．

2. アクセスできる（Accessible）

データとメタデータは識別子を利用して，オープンなプロトコルでアクセス可能でないといけない．

3. 相互運用可能である（Interoperable）

データとメタデータは，意味が共有できる言語を使って表現されていることが望ましく，またデータ間に関係がある場合，その関係も記述されるべきである．

4. 再利用できる（Re-usable）

データとメタデータは再利用性を高めるために，精度や関連性といった属性，明確でアクセス可能な利用ライセンス，来歴といった情報を付与されていることが望まれる．

データ公開プラットフォームの枠組み

FAIR原則に基づき，学術成果を適切に公開共有するための基本的な仕組みは図-3のようになる．学術成果は論文やプログラムコードを含み，全体としてデータとして扱われる．

そのデータ（1）にはデータそのものに加えてそのデータのフォーマット（2）が明示されないといけない．さらにこのデータがインターネット上で唯一であることを示す識別子（3）が必要である．さらにこのデータの内容がある程度理解できるためにメタデータ（4）が必要となる．論文であれば，タ

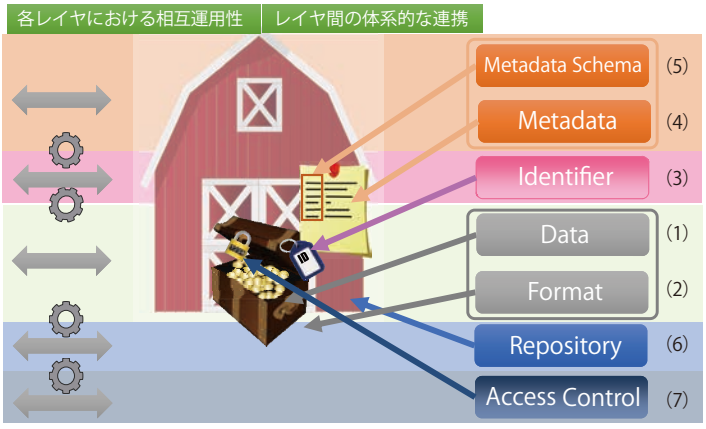

図-3　研究データ共有のアーキテクチャ

☆2　問題は（3）評価の機能である．この機能については F1000（https://f1000.com）のようなオンライン上での新しい評価システムが現在試行されている．

☆3　https://www.force11.org/group/fairgroup/fairprinciples

イトル，著者，所属，ジャーナルタイトル，巻号，ページ，発行年といった書誌情報であるが，データであれば，タイトルや生成者，公開年に加えてデータのサイズや生成年といったメタデータも必要となる．さらにメタデータに含まれる著者や所属といったものも識別可能であるために識別子が必要となる．また，このようなメタデータを収集集約可能にするためにメタデータのスキーマが適切に定義されていないといけない（5）．一方，メタデータを含むデータそのものを保管し公表するリポジトリ（6）も当然必要であり，またそこでは適切なアクセスコントロール（7）が必要である．

それぞれのレイヤにおいて，相互運用性を高めるために共通化や標準化が必要である．メタデータスキーマの共通化は，後述する DOI の登録機関によるメタデータスキーマや schema.org のスキーマなどが広く使われるようになっている．リポジトリは OAI-PMH というプロトコルが標準として使われ，DSpace 等の代表的なソフトウェアが世界中で利用されている．

こういったレイヤの上に検索や発見のサービスが構築される．たとえば，論文検索においてもリポジトリの情報を検索可能な OpenAIRE や Web 上に公開されたデータセットを検索可能な Google Dataset Search など，オープンな検索サービスが作られつつある．

学術データのための永続的識別子

前節で挙げたレイヤの中で相互運用性を高めるために特に重要となるのが識別子である．学術データのための識別子では，永続性が重要となる．

永続性[☆4]とは簡単に消滅したり変更されたりせず，比較的長期に渡って内容を指し続けることである．学術成果は長期に渡って利用可能でないといけないが，インターネットにおける URL は簡単に変わり得るため，そのギャップを埋める必要がある．

DOI

代表的な永続的識別子としては DOI（Digital Object Identifier）がある．これは元々は，学術ジャーナルが電子化したときに，これまで図書館にあった印刷されたジャーナルが長期に渡ってアクセス可能であると同じように，ジャーナル論文がインターネット上で長期に渡ってアクセス可能であるようにするために作られた仕組みである．DOI そのものは DOI Name と呼ばれる識別子（URL として表記される）と実際にその情報がアクセス可能な URL との関係を保持しておき，ユーザは DOI Name さえ知っていれば，たとえ元の情報が移動してもアクセス可能である．この仕組みの実現のためには識別子の登録や変更を受け付ける組織が必要であり，そういった組織の活動の上に成り立っている．

その 1 つである CrossRef は欧米のほとんどの学術ジャーナルが使っており，DOI だけでそれらのジャーナル上の論文にアクセスできる．重要なことは，CrossRef は DOI と同時に書誌情報も収集しており，識別子とメタデータの両方を扱っている．実際，多くの識別子ではメタデータとセットで扱われる．

DOI は現在は論文だけではなくデータにも使われる．DataCite では学術データに対しての DOI 付与のサービスを提供している．

ORCID

DOI とは別に，研究者を識別する識別子として ORCID というものがある．国内では文部科学省が管理している研究者番号があって，国内にいる研究者は容易に同定可能であるが，国際的にはそのような識別子はなかった．ORCID はこの研究者の識別子サービスを提供する非営利組織であり，そこが発行する ID が ORCID ID と呼ばれる．現在，多くの学術ジャーナルにおいて，投稿時に著者名に加えて ORCID ID を入力することで，発行した論文のメタデータにも付与され，論文の著者同定を容易にして

☆4 「永続的」は Persistent の訳で，本来なら「持続的」の方が適当であるが，すでに訳語として使われているため，本稿でも「永続的」と表記する．

いる．また，前述の DataCite においてもデータ公開の際に生成者などに ORCID ID をつけることを可能としている．

ORCID により，世界のどの研究者が何を公開しているかが容易に分かるようになってきている．

その他の識別子

このような識別子はインターネット上に増大する研究成果を適切に発見し，利用するときに有用であることが分かり，現在さまざまな識別子が提案されたり，実装されたりしている．たとえば，The Open Funder Registry は研究費助成機関のための識別子であり，これを研究成果（論文やデータ）のメタデータに含めることで，研究費助成機関は自ら助成した研究がどのような研究成果をもたらしているか容易に知ることができる．

あるいは研究分野特有の対象に識別子を与えることもある．たとえば，International Geo Sample Number（IGSN）は地質標本に対する識別子である．

プラットフォーム化の意味

データ公開のプラットフォーム化は確実に効率良くデータ公開を実現するには必要な仕組みである．しかし，一方で以下のような懸念もある．

(1) 公平な負担：プラットフォーム化はタダではない．運用したり維持したりするのには人的金銭的コストがかかる．それを誰が負担するのか．

(2) 囲い込み：このプラットフォームを使うときに何らかの資格や費用あるいは知識がいるとすると，それによって参加が限定される．

これらはオープンサイエンスの本質であるオープン性を損なう可能性があり，十分な留意が必要だろう．

科学の未来と情報技術

本稿ではオープンサイエンスの始まりから現在の状況までを概観した．ここまでで分かることは，情報技術が科学活動の形を決める大きな要因になっているということである．現在のオープンサイエンスは始めから方向が決まっていたわけではなく，情報の研究者・技術者と科学者が交流することで，現在のような方向に向かうようになった．今後もさらに新しい情報技術が科学のさらなる新しい価値と可能性を発見し得るだろう．これからも情報の研究者・技術者がこの点において科学に貢献することが期待されている．

（2018 年 1 月 24 日受付）

武田英明（正会員） takeda@nii.ac.jp
国立情報学研究所 教授．1991 年東京大学工学系研究科博士課程修了．工学博士．奈良先端科学技術大学院大学大学等を経て，2006 年より現職．人工知能，Web 情報学，学術コミュニケーション等の研究に従事．2018 年本会デジタルプラクティス論文賞受賞．