THE LINUX FOUNDATION | Research

World of Open Source®

オープンデータへの道筋

2024 World Open Innovation Conference Challenge Session での発見

2025 年 3 月

Anna Hermansen,
The Linux Foundation

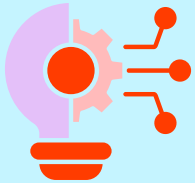
Paul Wiegmann,
Eindhoven University of Technology

序文 Professor Henry Chesbrough,
Luiss University and Haas School of Business at UC Berkeley

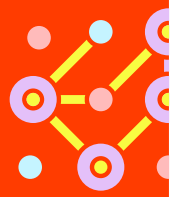
オープンデータへの道筋

データ サイロは研究とイノベーションの妨げとなり、AIモデルのトレーニングに必要なデータの増加に伴い、ますます厄介なものとなっています。

オープンデータは、誰もが自由にアクセスして利用することができ、**イノベーションの新たな道筋、より高い信頼性、信頼の向上**につながります。



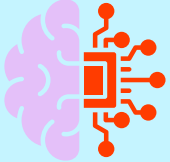
ソフトウェアと比較したデータのユニークな特性（**メンテナンス、品質、プライバシー、ライセンスの多様性**など）により、そのオープン化は困難な課題となっています。



データセットのクリーニング、標準化、維持には、**多大な人的資源**が必要です。

データセットの維持にかかる**金銭的**および**リソースのコストは、データの質とアクセス コストの間にトレードオフ**を生み出します。

標準化が不十分なためデータセットが使用できないものの、AIツールは非構造化データをより適切に管理する機会を提供しています。



データ プライバシーに関する懸念は、GDPRなどの規制への準拠から生じ、**リスク回避の風潮**を生み出しています。

データを独自に管理することで、企業はコンプライアンスと品質に関する確実性を高め、**競争優位性を失う恐れを軽減**することができます。

「セミ オープン」データ プラットフォームでは、コラボレーション参加者が、競争優位性を維持しながら、ベストプラクティスやその他の競争前のデータを共有することができます。

Overture Maps Foundationは、データ所有者やサービス プロバイダーが活用できる、**オープンで、プラットフォームに依存しない、標準化された地理空間データ プラットフォーム**を構築しています。

オープンデータ インフラストラクチャの構築には、**現在のデータ収集および共有プロセスの見直し**が必要です。

オープンデータでは、競争前のレイヤーにおける**コラボレーションを奨励**すると同時に、ガバナンス構造に**抑制と均衡を保つ仕組み**を組み込む必要があります。

目次

序文

デジタル化が進む世界では、私たちは皆、データのユーザーであり、作成者でもあります。しかし、私たちはしばしば、購入を急いだり、物語を読んだり、投稿をしたり、写真に反応したりする中で、この行為の潜在的な影響を無視してしまっています。このことを早くから理解した企業は、非常に大きな価値を獲得し、現在では蓄積したデータへのアクセスを仲介する役割を担っています。オープンデータの概念は、このような状況に対応するためのものです。これが、Linux Foundation のサポートを受けて作成された、オープンデータへの道筋に関するレポートの背景です。

Linux Foundation は、第11回 World Open Innovation Conference で、オープンデータに関するワークショップを開催することを決定しました。このワークショップへの関心と参加を後押ししたのは、人工知能ソフトウェア（AI）の急速な成長でした。AI は、そのアルゴリズムを訓練するために膨大な量のデータを必要とします。

このレポートはワークショップの要点をまとめたものですので、ここでは私が特に有用だと感じたいくつかのポイントを強調してご紹介します。1つは、誰もが自分のデータを保護したいと思う一方で、非常に高性能なアルゴリズムを求めているということです。アルゴリズムを高いレベルで動作させるには、大量のデータが必要です。そのため、非常に大規模な組織を除いて、より優れた AI アルゴリズムを実現するためには、データへのアクセスをオープンにするのが理にかなっています。

また、データは時とともに増え、変化していくという認識も重要です。したがって、オープンデータに移行するために、1回限りの作業や費用が発生することはありません。むしろ、これは継続的な取り組みであり、データの避けられない変化に対応するためのコストと取り組みをサポートする経済モデルを見出すことは、プロジェクトをサポートする経営陣の責任の一部とすべきです。

今後はどうなるでしょう？オープンデータが今後どのように発展していくかについて、少なくとも3つのアイデアが提案されました。1つ目はデータの所有権で、ユーザーは特定の条件下で自分の個人データを提供し、その他の条件下ではそのデータの使用を制限することを選択できるものです。2つ目は、「競争前」のデータセットにデータをコントリビュートする動機付けを創出することでした。これにより、コントリビュートされたデータが、例えば特定の個人を特定するために使用されることを防ぎながら、より一般的な特性の分析を可能にします。重要なことは、この競争前のデータセットが広く利用可能となり、これまで小規模企業や個人にとってはアクセスが極めて困難だった、あるいは単に利用できなかったデータへのアクセスが民主化されることです。

3つ目の重要な考え方は、ガバナンスの概念でした。大量のデータのリポジトリは、どこかに保存しておく必要があります。ハードウェア、ソフトウェア、セキュリティ、メンテナンスに費用がかかります。有用なデータ リポジトリへの幅広いアクセスを維持するためには、何らかの経済モデルが必要です。また、データへのアクセスに関する決定、およびそのアクセスに伴うあらゆる費用については、オープンデータ プロセスをサポートするステークホルダーにとって信頼性の高いガバナンスの仕組みの中で決定されなければなりません。

Henry Chesbrough
LUISS University in Rome, UC Berkeley in USA

はじめに：オープンデータの歴史と未来

私たちのデータは至る所に存在し、あらゆるものを支えています。マーケティング、医療、政府サービス、そしてAIエージェントのプログラミングに至るまで、組織はデータを活用して、効率と効果を最大限に高めています。しかし、データは多くの場合、企業内にサイロ化されており、サードパーティがデータにアクセスするには、技術面、法律面、経済面、運用面、文化面などの多面的で大きな障害を克服する必要があります。[1] データへの依存度が高まる中、これらの障害を評価し、組織がよりオープンで共有しやすい体制へと移行する方法を検討することが求められています。

オープンデータの概念は、イノベーション、透明性、コラボレーションの促進を目的として、非個人および非商用データが自由に公開される、オープン サイエンスにルーツがあります。[2] このオープンな文化は、収集されたデータが営利の機会とはならない公共財とみなされ、政府や公共部門の情報の透明性が奨励されている公的機関で最も強く見られます。[3] オープン ガバメントは、2000年代にオバマ政権のオープンデータ イニシアチブ（2009年）や欧州の公共部門情報指令（2003年）によって普及し、[1] すぐに多くの政府が、市民が自治体の公共情報にアクセスして分析できるオープンデータ ポータルを開発しました。米国政府のデータ ガバナンス ポータルによると、そのミッションは「政府のオープンデータの力を最大限に活用して、国民や政策決定者の意思決定に情報を提供し、イノベーションと経済活動を促進し、政府機関のミッションを達成し、オープンで透過的な政府の基盤を強化すること」です。[4]

商用および個人データ エコシステムに参入する場合、オープンデータの概念はさらに複雑になります。オープン ガバメントの義務付けがない場合、組織は利益の誘因、プライバシーの懸念、およびコントロールへの期待に苦慮し、その結果、データ所有者の目から見たオープンデータの価値を低下させます。一部の業界では、データ共有は倫理的義務となっています（医療分野など）。一方、他の業界では、サードパーティのデータセットと組み合わせることで価値を高めることができるため、データ共有が奨励されています（マーケティング分野など）。[5,6] しかし、データ アクセスに個人を特定できる情報が含まれる場合、そのデータを保護するプライバシー規制を遵守することが最優先事項となり、データをオープンにすることはリスクを伴います。このプライバシー リスクに加え、データの生成と収集が収益モデルの重要な構成要素となっているため、大企業は自社のデータ周りにウォールド ガーデンを築き、情報の流れを管理しています。[7] 現在のデータ市場モデルは、「データ コモンズ」の所有権を取得する商用企業で構成されています。[9]

ウォールド ガーデンの概念は、さまざまな業界や分野で見られます。たとえば、医療分野では、各病院や診療所が独自の電子記録システムを使用しており、システム間の相互運用性が欠如しているため、データがサイロ化しています。こうしたサイロ化はデータの価値を低下させ、患者、臨床医、研究者に悪影響を及ぼしています。また、標準化が不十分であるため、データが乱雑になり、フラクタル化、さらには使用不能になることもあります。[8] 欧州委員会は、さまざまな国の医療提供者間で電子患者ファイルを転送するためのイニシアチブやプログラムに取り組んできましたが、この相互運用性はまだ始まったばかりで、多くの地域ではまだ標準化されていません。[9] 同様に、エネルギー部門がデジタル トランスフォーメーションと電化を進める中、システムに接続されたすべてのデバイ

オープンデータの定義

本研究では、オープンデータを、誰もが自由にアクセスし、再利用、再配布できるための技術的および法的要件を満たしたデータ インフラストラクチャと定義しました。[16] また、データセンター中心の定義を超えたオープン性についても検討しました。これには、オープン スタンダードの文脈におけるオープン性の側面、例えば、成果物に対するアクセス、管理、および開発コストや、使用する上でのアクセス、管理、および使用コスト、成果物の完全性、成果物の共有、競合システムとのコラボレーションなどが含まれます。[17]

ス間で収集されたデータを共有することは、標準化と相互運用性がない限り困難な課題です。システムのさまざまなポイントで生成されるデータへのアクセスが改善されない限り、事業者やディストリビューターは、需要やグリッドの状態を分析するために必要なインサイトを得ることができません。[10]

データへのアクセスは新しい問題ではありませんが、生成AIのツールが爆発的に普及したことで、モデルのトレーニングに必要なデータ、特にこの種の使用が合法になるようなライセンスが与えられるデータに対する需要が高まっています。企業は、独自のデータを活用してモデルのトレーニングを行うようになってきています。Lawson ら（2024年）が発見したように、企業は、独自モデルと実装しているオープンソース モデルの両方のトレーニングに、自社データの一部を活用しています。[13] 独自のモデルを構築したいという要望は、組織がデータをより細かく制御できることから、非常に強いものです。[11] しかし、独自のデータに完全に依存することは持続可能とはいえず、組織が効果的で堅牢かつ偏りのないモデルを構築するためには、他のソースからの高品質のトレーニング データが必要となっています。[10] この点において、データ ワークフローを品質とコンプライアンスに配慮して責任を持って管理するデータ ガバナンスは、 オープンソースのAIに関するプロジェクトにとって最優先事項となります。[13]

生成型 AI があらゆる業界で重要なツールとなるこの次の時代において、オープンデータの未来は極めて重要になります。2024年11月、本レポートの著者は、カリフォルニア州バークレーで開催された World Open Innovation Conference（WOIC）に参加し、参加者に対して**「オープンでアクセス可能なデータを実現する道筋とは？」**というテーマのセッションを開催しました。データ エコシステムの障害、ニーズ、機会に着目し、参加者に以下の質問について議論してもらいました。

- データへのアクセスや利用にあたって直面する課題にはどのようなものがありますか？
- あなたの組織やプロジェクトは、イノベーションのためにデータをどのように活用していますか？
- あなたの組織は、社内外からのデータ アクセスをどのようにオープンにしていますか？
- 関連するサードパーティのデータにどのようにアクセスしますか？
- データ ニーズに対応するために、テクノロジーをどのように取り入れていますか？
- あなたの組織では、技術以外の分野（文化や方針の変更など）でどのようなソリューションが導入されていますか？
- データをよりオープンにするにはどうすればよいと思いますか？ あなたの経験から、何が必要だと思いますか？

以下のレポートは、この75分間のセッションのテーマ分析を中心に構成されています。チャタム ハウス ルールに基づき、参加者は受け取った情報を自由に利用することができますが、セッションの参加者を特定することは一切できません。セッションには、オープン イノベーションの分野に精通した、さまざまな業界からの学者や実務者が参加しました。彼らは、理論的な専門知識と、さまざまな状況におけるオープンデータおよびクローズドデータの取り扱いに関する実務経験に基づき、オープンデータの障壁と機会に関するインサイトを共有しました。

オープンデータが重要である理由

起業家、学者、イノベーターなどの聴衆にとって、データへのアクセスはビジネス インテリジェンスとイノベーションの重要な要素です。ある参加者は、投資に関する公開データ分析は、クライアントのビジネス インテリジェンスにとって重要な要素であると述べています。別の参加者は、個人レベルの社内従業員データの価値と、このレベルのデータの透過性を求めるクライアントの要望について論じました。彼らは、「チームリーダーはデータを見て、「これはできる」と判断でき、自信を与えるでしょう。」とコメントしています。公開データと社内データを使用して特定の成果を導き出すことができることは、データのオープン性とアクセス可能性の妥当性を示す証拠となります。Ambiel (2024) の研究によると、サードパーティのデータと社内の専有データを組み合わせることは、「大規模な AI モデルのトレーニング、研究の検証、市場機会の発見に不可欠」です。[7]

セッションの参加者は、現在のデータ アクセス状況では、こうした機会や検証のポイントを必ずしも確保できるとは限らないことを指摘しました。例えば、ある参加者は「誰が何を研究しているかを把握することが難しい」と説明しました。同様に、別の参加者は、あるデータセットに関する学者間のコラボレーションの欠如について、次のように述べています。「もし、良いデータ ソースを手に入れられたら、それは宝の山です。しかし、そのデータをすべて分析することは不可能であり、他の科学分野にとってどのような意味を持つかを理解する余裕もありません。例えば、工学に役立つかもしれないし、社会科学に役立つかもしれないし、化学に役立つかもしれない...しかし、どうやってそれを知ることができるでしょうか?」これは機会損失であり、そのデータセットは他の研究チームにとって有用である可能性がありながら、それらのグループには発見できないという状況です。

こうした機会損失は、イノベーションの実現を妨げることを意味します。スポーツ業界のある参加者は、「スポーツ データは無料であるべきです...チームが競い合う場所は、通常、データに付加価値をつける商業的に価値のある製品を見つけること、つまり、予測分析です。」とコメントしています。競合他社間でもデータの共有は可能であり、それにより組織はより迅速にイノベーションを進め、共有データに基づいて製品を開発することができます。このコラボレーションにより、アナリストがデータポイントを代表する個人から抽象化することが可能となり、プライバシーに関する懸念を軽減するビッグデータが生み出されます。ある参加者は、通行量データという例を挙げました。このデータでは、個人の位置情報それ自体は非常に個人的なものであるが、他のデータポイントと集計して、その場所に何人の人が訪れたかを示すと「単なる一般的なヒストグラムになり、より個人的な情報ではないものに抽象化することができます。」と説明しました。 別の参加者は、「ビッグ データをミクロ レベルまで拡大しても、それはまったく価値のないものになります。重要なのは、その傾向と分析結果だけです。ビッグ データは、十分な量になった瞬間に初めて有用なものになるのです。」と述べています。データセットが大きければ大きいほど価値が高まることを知っている参加者は、有効なデータセットを作成するためにこれは必要なデータへのコントリビューションの動機付けになるべきだと主張しました。

共有データセットの分析的価値に加え、この活動は信頼の向上にもつながります。ある参加者は、データ共有および分析プラットフォームに関するコンソーシアムの例を挙げ、「参加者間に信頼関係が生まれている...これは、パートナーシップの構築について考える上で、非常に説得力があるあります。」と述べています。このコラボレーションという社会的契約により、信頼に依存する共同作業が可能になります。別の参加者は、「私自身は「データ」を見ることができないかもしれない...でも、パートナーに頼めば代わりにやってくれます。そして、私は彼らを信頼しなければいけません。でも、私たちは同じグループの一員だから信頼しています。」と述べています。データをオープンにすることで、イノベーションの新たな道が開かれ、データセットの有効性と信頼性が向上し、チーム間の信頼関係も深まります。

オープンデータの現在の課題

前述の通り、オープンデータ プラットフォームは、技術面、規制面、経済面、文化面での数え切れないほどの課題に直面しています。課題セッションの前半では、これらの課題の一部を紹介し、参加者に、業務でそれらに直面した経験について振り返ってもらいました。

データの独自性

オープンデータが直面する課題を考える際には、ソフトウェアなどの他のコンテンツと比較して、データ特融の特性を考慮することが重要です。Overture Maps FoundationのMarc Prioleau, the executive directorは、自身のブログ記事で、オープンデータがオープンコードと異なる6つの特性を挙げています。

- データの正確な出所
- ナビゲートするための、データ ライセンスのパッチワーク
- データの収集、ホスティング、およびメンテナンスの規模とコスト
- データの継続的な作成に必要なワークフロー
- データの正確性および品質の保証
- 個人を特定できる情報の保護[12]

Bennetら (2024) は、データ集約型アプリケーション (この例では AI アプリケーション) が直面する固有の課題、特に同意の違反の可能性や、さまざまなオープン ライセンスのデータセットの管理について指摘しています。[13] 参加者は、チャレンジ セッションでこれらの課題やその他の課題を取り上げました。

コストと品質のトレードオフ

セッションで何度も取り上げられたテーマのひとつは、コストと品質のトレードオフという考え方でした。一部の参加者は、オープンデータよりも質が高いと評価しているため、データに料金を支払っている、と述べています。「私は、プライベートデータにはお金を払います...なぜなら、そのデータはより質が高く、厳選されているからです。」と、ある参加者は述べています。しかし、これは「当社のビジネス モデルでは持続不可能」な高価なオプションであるため、無料と有料のソースを組み合わせて使用しています。このトレードオフは、別の参加者も「無料のデータは有料のソースで補強しているが、お金に制限がなければ、プライベート データにお金を払うでしょう。なぜなら、そのデータはより質が高く、厳選されているからです。」と述べています。

データのコストについて考えたとき、参加者はデータの管理に費用がかかることを指摘し、アクセスに料金を課さないとそうした管理を行うインセンティブが欠如してしまうと述べました。経済的なモデルがなければ、オープンデータのデータセットは、信頼性が低く、標準化もされていないボランティアのコントリビューションに依存しています。ある参加者は「データを収集する人々が標準的な方法でそれを実行してくれればいいのですが、問題は、彼らにはあまり利点がないことです...利点がない限り、彼らは『なぜそれをしなければならないのか?』と言うでしょう。実際にそれを実行しなければならない人は、動機がないか、実行を強制されていないかのどちらかです。」と述べています。

質の高いデータセットを維持する上でのもう一つの問題は、データの可変性です。データが収集される対象が変更されると、データセットの信頼性が低下します。ある参加者は、マッピング データを例に挙げ、次のように述べました。「マッピングの難しいところは、現実の世界を反映しているため、現実の世界が変化すると、マッピングしたデータも変更しなければならないことです。」一部の分野では、こうした人工物が変化する速度が非常に速いため、データが現実を確実に反映するよう、継続的なフィードバック ループが必要となっています。

データセットの継続的な更新とメンテナンスを行わないと、データの品質に懸念が生じます。これには、データが最新であるかどうか、データがさまざまな人口や地理をどの程度よく表しているかも含まれます。ある参加者は、自身の業務における例を挙げ、クライアントに世界中の情報へのアクセスを提供したいと説明したが、実際にこのようなケースは滅多にありません。「世界中とは、北米、ブラジルからのデータがあればブラジル、中央および西ヨーロッパを意味しますが、東ヨーロッパは含まれな

いかもしれません。つまり、さまざまな地理的な制限があるということです。」こうした制限のため、データセットが最新かつ完全であるという主張には懐疑的な見方が生じます。これは、オープンデータセットに依存している人々にとってプレッシャーとなります。なぜなら、サードパーティのデータ ストリームは彼らの管理外であるにもかかわらず、顧客にとって「真実の源」となるからです。

オープンデータ セットを作成する労力

上述の通り、データ管理には費用がかかります。これは、データの収集からメンテナンス、品質管理に至るまで、さまざまなワークフローを管理するために必要な労力が主な要因です。ある参加者は「私は現在、多くのデータ クリーニングを行っています。それには近道はないと思います。回避策はなく、それはもちろん研究の一部です。」と述べました。しかし、組織のインテリジェンスにサードパーティのオープンデータを統合する際には、重要なリソースに関する考慮事項もあります。参加者は、オープンデータ セットの利用における品質管理の取り組みについて説明しました。「最新情報を入手することがいかに難しいか、言葉では言い表せません。結局は、その人に電話をかけなければならず、手間がかかりますし、効率的ではありません。これらのデータベースは出発点としては良いのですが、それが唯一の真実のソースや検索の最終目的地となることはほとんどありません。」

これらのデータセットは不完全であることが多く、信頼性に欠ける可能性もあるため、それらをどのように使用するかについて決定を下さなければなりません。ある参加者は「良いデータでも悪いデータでも、データをインポートし、手作業で作業を行う必要があります。そして、ここで問題となるのは、手間をかけて作業を行うか、それとも、自分でgoogle検索する手間を省くために、4分の1だけ記入して残りは記入しないままにするか、ということです。その結果、不完全なデータセットとなり、使用できなくなるという最終的な問題に直面します。」と述べました。

もちろん、データがクリーン アップされた後も、その後の作業には依然として多大な人的リソースが必要となります。ある参加者は、「何年もかかる」と述べています。「クリーン アップに費やす時間だけでなく、分析、調査、出版、レビューに費やす時間も考慮すべきかもしれません。」

標準化

データ クリーニングにおける重要な要素のひとつは、標準化です。データセットが標準化されていない場合、データの信頼性と他のデータとの比較における有用性に影響を及ぼします。これは、ある参加者が次のように述べました。「誰もが何でも自由に記入できると、その部門や研究者グループの能力など、全体像を把握できる標準化されたデータを得ることができません。そうなると、その問題に適任な人物を見つけたと誤認する可能性が出てきます。これが、現在、私たちが抱えている主な問題です。」標準化のプロセスは、業界によって複雑さが異なり、ステークホルダーがサードパーティのデータを読み取り、使用する能力に影響を与えます。「エンジニアリング データには、番号と単位があり、それだけです。それにタイムスタンプが付いている場合もあります。」と、ある参加者は述べました。「しかし、医療データを解釈するには、使用された方法、測定された条件、測定された時期など、さまざまな情報が必要です。」これにより、業界間のデータ共有はさらに複雑になります。

標準化が大きな懸念事項として挙げられましたが、一部では、AIの進歩もあり、将来的にはそれほど問題ではなくなる可能性があるとの意見も出ました。ある参加者は次のように述べています。「私は、2つの点から、標準化が時間とともに容易になるだろうと予想しています。1つ目は、10年前と比べてデフォルトで注釈が付けられ、より構造化されたデータがますます増えていることです。2つ目は、非構造化データに対して構造化処理を行うアルゴリズムがますます増えていることです。したがって、技術的な観点からは、この問題は自ずと解決されるだろうと非常に楽観視しています。」標準化の必要性に関する透明性、および非構造化データの利用性を高めるために利用可能なツールの利点は、非標準化がデータ共有の実務に与える影響を軽減する可能性があります。

データ プライバシー

データの品質に加え、プライバシーやビジネス上の理由によるデータの保護も、オープンデータの大きな障壁のひとつと考えられています。個人識別情報 (PII) などの機密データを公開する可能性があるため、一部の参加者はデータのオープンにためらいを見せました。「当社のデータは非公開であるため、オープンソースやオープンデータを利用することはでき

ません。そのため、社内の AI および高度な分析グループが、すべてを監視するための独自のオンプレミス ツールを開発しています。」主に、オンプレミスのストレージのみを使用し、クラウドストレージを使用しないという規制への準拠、および国境を越えた準拠に関する懸念が寄せられました。「ガバナンスの仕組みや政策は、データをどの程度オープンにできるかという点において、非常にデリケートです。なぜなら、それは実際には各国の事情に依存するからです。」また、参加者はAI モデルについて、「各国が独自の AI 法を制定しているかどうか、そしてデータが国境を越えてグローバル化すると、その取り扱いが非常に困難になる。」という懸念も述べました。活動に適用される規制の複雑さ、特に国境を越えて行われる活動を考慮すると、理解できるためらいが生じます。

データ共有に関する規制の影響について議論する際、多くの参加者は、一般データ保護規則（GDPR）を参考例として挙げました。ある参加者は、クライアントのために完了した仕事について、「その会社は、生産プロセスの作業者の動きをマッピングして、作業者のリスクを回避したいと考えていました。しかし、個人データを収集していることから、GDPR の問題が発生することが判明しました。」という事例を共有しました。これは、参加者が有意義な結果を出すことを妨げ、興味深いことに、クライアントと労働者が望んでいたこととは反対の結果となりました。GDPR の規定では、5人未満のデータセットの収集や共有は禁止されていますが、参加者は以下のように述べています。「実際には、従業員はそれを望んでいます。なぜなら、彼らにとっては、自分のデータを操作し、その内容を確認するための優れたツールだからです。その観点から、従業員は、自分のデータを見たいと、従業員代表委員会と私たちと対立している状況です。」

ある参加者によると、GDPRがデータ共有に悪影響を及ぼしていることは珍しくなく、「GDPRを本来の目的のために使用するのではなく、その可能性を皆に恐れるように仕向けているだけです。最初の5年間は、GDPR は物事を阻止するためにのみ使用されていましたが、実際には、GDPRは多くのことを可能にするものです。問題はありません。しかし、GDPRで何ができるか知らない人たちは、安全な側に立ち、もう何もできないと主張するのです。」と述べていました。これは、別の参加者によって指摘された非対称リスクを生み出します。弁護士が特定のデータ共有活動を承認し、その判断が誤っていた場合、彼らは職を失うリスクを負うことになります。そのため、不確実性がある場合には否定することが容易になります。「このデータを共有してもいいですか？と尋ねると、最終的には、「いいえ」と答えるように指示されている人物に回されます。つまり、非対称的なリスクを解決しなければならないのです。」これが、データセットをオープン化に対する強い抵抗を生み出しています。

データの管理

規制面およびビジネス面の観点で、管理の必要性があることが、データオープン化に対する抵抗の理由でした。これは社内の監視という形で現れ、ある参加者は「IT部門はデータへのオープン アクセスを提供したがらない。」と述べました。前述の通り、組織の法務チームもデータへのアクセスを妨げる可能性があります。「データ共有やデータ受信は、誰かが「ちょっと、X氏に確認しなければならないんだけど、X氏は法律ガイドのようなものだから…」となり停止されます。たとえオープンソースのコードや、利用可能なオープンソース ソフトウェアを使用する必要があったとしても、IT 部門に「許可がない」と断られるのです。そのため、規制とぶつかる場合もあります。」最後に、もう1つの内部統制の形態について、ある参加者が説明しました。「第三者の血圧測定値に頼るのではなく、自分で測定し直す方がはるかに安全です。自分で管理しているからです。」より安全で簡単な代替策は、組織がデータをクローズドにしておくことです。

データの非公開を維持するプライバシーや品質に関する懸念に加え、データを共有することで競争優位性を失う可能性があるという意見も参加者は述べました。ある学術関係の参加者は、論文の発表と同時に収集したデータセットの文脈において、次のように自問自答する例を挙げました。「他の研究者のために（データセットを）事前に公開すべきでしょうか？戦略の問題であり、率直に言えば、同じ研究テーマで他の研究者が私たちよりも早くデータセットを利用することを恐れているという理由があります。」組織のデータを一般公開すると、その潜在能力が損なわれるのではないかという懸念がありました。

これは、一部のデータはオープンにすべきではないのかという疑問につながります。ある参加者は、「もちろん、すべてのデータがオープンにすべきというわけではない。」と述べ、この考えを次のように表現しました。Ambiel（2024）が指摘するように、金融サービスや医療などの業界の一部の企業や組織では、そのデータは価値が高すぎる、あるいは機密性が高すぎるため、新たなリスクを導入することはできず、その結果、データの活用と使用を制限せざるを得ない場面があります。しかし、ある参加者は、考慮すべき場合があると述べました。「すべてのデータは高度な機密情報ですか？実際にあなたとあなたの競争優位性に関する機密情報であるデータはどれで、それほど機密情報ではないデータはどれですか？」自社の競争優位性を構成するデータ、あるいは他者と共有することでより価値が高まるデータと比較して機密性の高いデータについて考慮することが重要です。

オープンデータの成功と機会

後半のチャレンジ セッションでは、オープンデータの未来に焦点を当て、参加者はオープンデータ エコシステムに関するケース スタディやアイデアについて議論しました。ある参加者は、コラボレーターがスタートアップ企業との協働から得たベスト プラクティスや成果を共有する資金調達プラットフォームの例を挙げました。このシステムでは、「オープンデータは、スタートアップのすべての特性、つまり彼らが誰と仕事をしたか、タイムライン、実験が行われた時期、規模などに関するデータです。ですから、何かをした、あるいは何かを達成できなかったと主張するには、十分な量のデータをコントリビュートしなければなりません。しかし、寄付者の名前を明かすほどの情報ではありません。つまり、セミオープンデータです。」この競争前の層を構築することで、エコシステム内のさまざまなプレーヤーが、個々の組織の優位性を損なうことなく、すべてのグループに利益をもたらす形で非独占的なデータを共有する機会が生まれます。

学術的な観点では、データセットを公開することに関する懸念から、潜在的なクレジット付与とライセンス戦略についての議論が起こりました。ある参加者がグループに尋ねました。「データセットを公表前に利用可能にしたいが、他の研究者に先を越されることを恐れています。データ収集だけで、とにかくクレジットを得られるならどうでしょうか？ 彼らは研究をより早く進めていますが、それでもあなたはデータをライセンス化することでクレジットを得られます。これは私のデータであり、研究に利用する場合は、私がそのクレジットを得ます。」ライセンスによるデータ収集および使用に関するこの監査証跡は、特に学術分野において、競争上の優位性に関する懸念において潜在的な解決策となります。

OVERTURE MAPS

Overture Maps Foundationは、あらゆる地図製品で自由にアクセスできる、信頼性が高く相互運用性に優れたオープンな地図データを作成することで、地図業界に変革をもたらしています。戦略的なコラボレーションを通じて、加盟組織は、コミュニティ、政府、企業などのソースからデータを組み合わせ、標準化されたスキーマとデータセットを開発しています。Overtureは、厳格な検証と標準化によりデータの品質を確保し、オープンデータの利点を維持しながら、商用アプリケーションへの適合性を確保しています。

Overtureは、業界が抱える根本的な課題、すなわち、ライセンス費用を上回る場合も多い、地理空間データの処理および統合のコストと複雑さの増大に対処しています。共有インフラストラクチャと標準化されたデータ パイプラインを構築することで、Overtureは組織全体にわたる重複作業を排除します。重要な革新は、グローバルに安定した一意の識別子をマップ機能に提供するグローバル エンティティ参照システム（GERS）です。GERSは、グローバルでオープン、エンティティ ベースであるという点で特徴的であり、組織が外部データをベース マップに直接リンクし、アプリケーション間の相互運用性を確保することを可能にします。このコラボレーション アプローチにより、組織は、業界全体のイノベーションを加速し、標準化され、継続的に改善されるベース レイヤーを活用しながら、付加価値の高いサービスに注力することができます。

Overtureは2022年12月の設立以来、Facebook、Instagram、Bing/Azure maps、Esri の ArcGIS Living Atlas などのプラットフォームを通じて、何億人もの消費者が利用するアプリケーションにデータを提供し、大きな進歩を遂げてきました。2024年現在、Overture は、23億件の建物のフットプリント、5,400万件の POI、区分、および土地や水に関するデータを含むコンテキスト レイヤーを網羅する、生産準備の整ったデータセットをリリースしています。交通データセットは、世界中の8,600万キロメートルの道路をマッピングし、詳細な交通規則や規制を含んでいます。Overtureは、4人の創設メンバーから始まり、多様な分野にわたる37以上の組織に拡大し、マッピング エコシステム全体のオープンな基盤層としての地位を確立しています。[14]

次のステップ：オープンなデータ構造に必要なもの

オープンデータに関する事例研究と機会についての議論は、よりオープンなデータシステムを構築するための次のステップへの分析につながりました。Majer（2024）が調査した内容と同様に、ウォールド ガーデン アプローチ全体を、新しいガバナンス メカニズム、分散化、コラボレーション、オープンソースによって解体する必要があります。[9] 議論の分析から、データ共有の状況を再構築し、エコシステムをよりオープンなものへと移行する上で役立つ3つの重要なテーマが明らかになりました。

1つ目は、**データ所有権**は、重要な公共の懸念事項として、現在のデータ収集および共有の在り方を再考する上で有用な手掛かりとなります。ある参加者は、現在の権力と支配の力学を「非対称」と表現し、「私たちは、データが主に多額の収益のために収集され、ビジネス モデルを生み出すために使用されていた時代から来ています。ユーザーである私たちは、フィードバックを得ることも、払い戻しを受けることもありませんでした。そして、これが、私たちが多くのことを過剰に規制している理由なのです。」と述べました。彼らの観点からは、この非対称的なパワーバランス、そしてそれに対抗しようとする規制は、使用権の再構成によって解決できると考えています。これは、「私は自分のデータを提供します、あるいは提供しません。私は自分のデータを提供しますが、そのデータについては特定の用途にのみ使用を許可します。私のデータを提供しますが、広告などには使用できません。そのデータは、癌の治療など、その目的のみに使用してください。」というものです。これらの使用権により、データの使用のビジビリティと透過性が向上し、データの公開やプライバシーに関する懸念が軽減され、データの保護よりも共有がより重要となる環境が生まれます。

参加者は、使用権を実際に確立する方法として、技術的なアイデアを共有しました。例えば、ある参加者は、データが特定の方法で使用されることを保証するというアイデアについて議論しました。「その問題にはどのように取り組んでいますか？ブロックチェーンなどを考えてみると、特定の条件や目的のためにデータを共有しようとしている人々に安心を与える上で、テクノロジーが役割を果たすことができると思います。」別の参加者は、「スマート コントラクトやブロックチェーンのようなものを追加することができる」と賛同しました。ブロックチェーン以外にも、自分のデータを再び管理するための方法として、ソリッド スタンダードを提案する参加者もいました。「私にとって、これは大きな変化です。なぜなら、ユーザーである私は自分の意思で共有をオン、オフすることができるようになるからです。つまり、企業が自分の好みに合わなくなったら、私はそれをオフにするだけです。これは、今日ではまだ不可能なことです。」

データの所有権も、データの共有に関する法的フレームワークを提供するコミュニティ データ ライセンス契約（CDLA）などのライセンスを通じて管理する必要があります。最新のリリースである CDLA 2では、データの使用、修正、共有に関する条件の概要が記載されており、データの所有者を保護しながら、データの広範な共有と使用を可能としています。オープンデータ セットに関するガバナンス構造とコラボレーション モデルを構築する場合、この種のライセンスは、データセットの活動に関する構造を提供し、信頼性を高め、データ共有プロセスを効率化します。[15]

二つ目に、**競争前のデータセットに関するコラボレーションを奨励する**ことで、データ共有に関する課題の一部に対処し、よりオープンな文化への移行をサポートすることができます。前述のように、参加者は、業界内の他者とのコラボレーションに役立つデータ層、つまりデータの抽象化というものが存在することを理解しました。これを実現するには、共同データセットにデータをコントリビュートする価値提案を構築することが重要です。ある参加者は、「特定の方法でのデータセット構築は、ある当事者に依存しているものの、別の当事者に利点をもたらす場合が多いです。問題は、そのデータを所有している人々にそれを提供してもらうにはどうすればよいか、ということです。」と述べました。コラボレーションのメリットは、データが共有されることで生じるプラスの外部効果になります。一部の人々にとっては、他の人々とデータセットを構築することは、実際にデータセットを機能させるデータポイントを追加することを意味します。「初期段階では、データが十分でないため、データを共有することが理にかなっている場合もあります。データが十分ではないという、それだけで動機になるかもしれません。私が100件の観測データしか持っていないのに、他の誰かが200件持っていて、さらに100件をどこかで見つけたとしたら、それは私たち全員にとってより有効なデータセットになるかもしれないからです。」このコントリビューションにより、共有データセットの価値が高まり、イノベーションの可能性がさらに広がります。

ユーザーの権利に関する議論で示唆されているように、利他主義も考慮すべき重要な動機付けのひとつです。これは、人命を救うためにデータを共有することがほとんどの人にとって強い動機付けとなる医療分野では特に明白です。しかし、利他的な行動を促すことは、その価値提案の表現方法に依存すると、ある参加者は述べています。「「あなたのデータを医薬品開発に使用してもいいですか？その最終的な成果はファイザー社やX社になります。」と質問すると、多くの研究では、人々は共有を望まないという結論に達しています。しかし、価値提案を「あなたの血液サンプルを使って癌の治療方法を見つけ、それをオープンにし、その治療法が1つの企業だけに独占されないことを保証します。」と言い換えると、多くの人々が自分の医療データをすべて共有しても全く問題ないと考えるようになります。」これは、データ リクエストに対する信頼と、そのエンティティがリクエスト通りに使用するという前提に基づいています。「私はデータを共有するつもりです。なぜなら、私はあれこれと手助けしていてそれがどこかに届くことを願って、その相手がその情報を正しく扱うことを信頼しているからです。」利他主義は、オープンなデータセットへのコントリビューションに関するコラボレーションを促進する重要な行動メカニズムです。

3つ目に、オープンデータ セットの構築における障害となっているのは、コラボレーションと中立性の文化のバランスを取りながら、抑制と均衡を管理するための適切な**ガバナンス体制**を構築することです。少なくともチャレンジ セッションの参加者の中には、オープンデータ セットには所有権がないため、データの品質に影響が出ているという見方があります。ある参加者は、「技術的には誰もデータを所有していないため、データを更新する動機がないです。」と述べました。この問題の解決策を検討するにあたり、ある参加者は「このようなことが実現するには、どのような政府の仕組みや政策が必要でしょうか？ オープン データベースに必要なものは何でしょうか？ 信頼と、ある種の階層構造です」と提案しました。別の参加者によると、階層構造を推奨するということは、ある組織がホスティング、サーバー費用の負担、アクセスとセキュリティを管理すること意味します。「サーバーの費用は誰が負担するのでしょうか？ アクセスに必要なユーザ名は誰が管理するのでしょうか？ セキュリティはどうするのでしょうか？ アクセスしたユーザーとアクセスした日時を記録するログも必要です。その管理も誰かが担当しなければなりません。 どのIT部門が担当するのでしょうか？ 必然的に、階層構造が必要になります。」データの共有と、プロセスへの個人の参加を促進する取り組みをサポートする新しいガバナンスの形態を検討することは、データ環境を変革し、すべての人の利点となるデータの公開を促進する可能性があります。

結論

WOIC Challenge sessionでは、学者と実務家がオープンデータとクローズドデータのトレードオフをどのように考えているかが明らかになり、オープンデータベースに対する現実的な懸念や期待がいくつか指摘されました。このセッションの分析と報告を通じて、オープンデータの重要性を明らかにし、データを扱う人々がデータセットのより良いコラボレーションの方法、共有の促進、よりオープンな文化をサポートする組織文化の再構築について検討していただくことを目指しています。新しいテクノロジー、新しい政府、新しい経済問題により政策や文化が変化する中、逆風であってもオープンな方向性を確立することが重要です。

方法論

この研究で取り上げた調査結果は、2024年11月6日にカリフォルニア州バークレーで開催された「2024 World Open Innovation Conference」の75分間のセッションの記録を基に作成されたものです。著者はセッションの司会を務め、トピックを紹介した後、グループ ディスカッションを進行しました。議論は録音され、Otteraiを使用して文字起こしされました。最初の著者は、トランスクリプトをコード化し、コードのパターンからテーマを導き出し、調査結果を補強するために二次文献を使用してレポートを作成しました。本レポートは、作成前に、第2著者およびその他のステークホルダーによるピア レビューを受けています。

脚注

1 Attard, Judie, Fabrizio Orlandi, Simon Scerri, et al. “A systematic review of open government data initiatives.” Government Information Quarterly, no. 32 (October 2015): 399-418. https://doi.org/10.1016/j.giq.2015.07.006

2 Braunschweig, Katrin, Julian Eberius, Maik Thiele and Wolfgang Lehner. “The State of Open Data: Limits of Current Open Data Platforms.” (2012). https://api.semanticscholar.org/CorpusID:17298359

3 Zuiderwijk, Anneke and Marijn Janssen. “Open data policies, their implementation and impact: A framework for comparison.” Government Information Quarterly, no. 31 (January 2014): 17-29. https://doi.org/10.1016/j.giq.2013.04.003

4 “Data.gov Home.” Data.gov, accessed February 14, 2025. https://data.gov/

5 Ambiel, Suzanne. “The Case for Confidential Computing: Delivering Business Value Through Protected, Confidential Data Processing.” The Linux Foundation. July 2024. https://www.linuxfoundation.org/research/confidential-computing-use-case-study

6 Gaba, Jeanne Fabiola, Maximilian Siebert, Alain Dupuy, et al. “Funders’ data-sharing policies in therapeutic research: A survey of commercial and non-commercial funders.” PLoS ONE, 15(8). https://doi.org/10.1371/journal.pone.0237464

7 Majer, Alan. “Decentralization and AI: The Building Blocks of a Resilient and Open Digital Future.” The Linux Foundation. November 2024. https://www.linuxfoundation.org/research/decentralized-internet

8 Hermansen, Anna. “An Open Architecture for Health Data Interoperability: How Open Source Can Help the Healthcare Sector Overcome the ‘Information Dark Ages.’” The Linux Foundation. October 2024. https://www.linuxfoundation.org/research/health-data-interoperability

9 “Exchange of electronic health records across the EU.” European Commission, accessed February 25, 2025. https://digital-strategy.ec.europa.eu/en/policies/electronic-health-records

10 Dover, Mike. “Open Source and Energy Interoperability: Opportunities for Energy Stakeholders in Canada.” The Linux Foundation. August 2024. https://www.linuxfoundation.org/research/canadian-energy-interoperability

11 Lawson, Adrienn, Stephen Hendrick, Nancy Rausch, et al. “Shaping the Future of Generative AI: The Impact of Open Source Innovation.” The Linux Foundation. November 2024. https://www.linuxfoundation.org/research/gen-ai-2024

12 Prioleau, Marc. “The Unique Challenges of Open Data Projects: Lessons From Overture Maps Foundation.” The Linux Foundation. January 13, 2025. https://www.linuxfoundation.org/blog/the-unique-challenges-of-open-data-projects-lessons-from-overture-maps-foundation

13 Bennet, Karen, Gopi Krishnan Rajbahadur, Arthit Suriyawongkul, et al. “Implementing AI Bill of Materials (AI BOM) with SPDX 3.0: A Comprehensive Guide to Creative AI and Dataset Bill of Materials.” October 2024. https://www.linuxfoundation.org/research/ai-bom

14 “Overture provides free and open map data.” Overture Maps Foundation, accessed February 14, 2025. https://overturemaps.org/?utm_source=LF&utm_id=opendatareport

15 “Open Data Sharing.” Community Data License Agreement, accessed February 28, 2025. https://cdla.dev/

16 “What is open?” Open Knowledge Foundation, accessed February 25, 2025. https://okfn.org/en/library/what-is-open/

17 West, Joel. “The economic realities of open standards: black, white, and many shades of gray.” In: Greenstein S, Stango V, eds. Standards and Public Policy. Cambridge University Press; 2006: 87-122.

謝辞

著者は、シームレスで没入感のあるカンファレンスを開催し、このチャレンジ セッションを議題に取り入れてくださったWorld Open Innovation Conferenceの主催者に感謝いたします。セッションの参加者は、多様で非常に熱心なグループであり、このレポートの基礎となる、適切かつ建設的な見解を述べました。Hilary Carter氏とHenry Chesbrough氏には、原稿の精査に多大なご尽力をいただきました。また、Linux Foundation Creative Services teamと Christina Oliviero氏には、本レポートの作成と発行管理に尽力いただきました。

著者について

Anna Hermansenは、Linux Foundation Researchの研究者兼エコシステム マネージャーとして、Linux Foundationの研究プロジェクトのエンドツーエンドの管理をサポートしています。彼女は、医療におけるデータ共有をより良くサポートするための新しい技術の統合と、医療データ インフラに関する定性的および体系的なレビュー研究を実施し、その研究成果をカンファレンスやワーキング グループで発表しています。彼女の関心は、健康情報学、精密医療、データ共有が重なる点にあります。彼女は、クライアント サービス、プログラムの提供、プロジェクト管理、学術、企業、ウェブユーザー向けの執筆など、幅広い経験を持つジェネラリストです。Linux Foundationの前には、Blockchain Research InstituteとBC Cancer's Research Instituteの 2 つの研究プログラムに携わっていました。 University of British Columbiaで公衆衛生学修士号と国際関係学学士号を取得しました。

Paul Wiegmannは、Eindhoven University of Technology (TU/e)の助教授であり、イノベーションの分野におけるスタンダードおよび標準化について研究および教育を行っています。彼の研究は、経営と政策の交差点にあり、標準化エコシステムにおけるさまざまなステークホルダーが、イノベーションと前向きな社会変化をサポートするために、どのようにスタンダードを策定し、実装できるかを調査しています。Paulの研究は、「Research Policy」、「Academy of Management Annals」、Environmental Innovation and Societal Transitions」などの専門誌や、単著の書籍にも掲載されています。European Academy for Standardisation（EURAS）の会長であり、 University of California, Davis、Yonsei University、Technical University of Berlinの客員研究員を務めた経験があります。TU/eに入社する前は、Erasmus University Rotterdamでイノベーション マネジメントの博士号と修士号を取得し、イギリスのUniversity of Warwickで経営学の学士号を取得しました。

この著作物を参照する場合は、次のように引用してください。Anna Hermansen and Paul Wiegmann, “Pathways to Open Data: Findings from the 2024 World Open Innovation Conference Challenge Session,” foreword by Henry Chesbrough, The Linux Foundation, March 2025.

x.com/linuxfoundation
facebook.com/TheLinuxFoundation
linkedin.com/company/the-linux-foundation
youtube.com/user/TheLinuxFoundation
github.com/LF-Engineering

THE LINUX FOUNDATION | Research

2021年に設立された Linux Foundation Research は、拡大するオープン ソース コラボレーションを調査し、新たな技術トレンド、ベストプラクティス、オープンソース プロジェクトのグローバルな影響に関 する洞察を提供しています。プロジェクトのデータベースやネットワークを活用し、定量的・定性的手法 のベストプラクティスに取り組むことで、世界中の組織にとって有益なオープンソースの知見を提供するライブラリを構築しています。

本訳文について

この日本語文書は、Pathways to Open Data: Findings from the 2024 World Open Innovation Conference Challenge Sessionの参考訳として、The Linux Foundation Japanが便宜上提供するものです。英語版と翻訳版の間で齟齬または矛盾がある場合（翻訳版の提供の遅滞による場合を含むがこれに限らない）、英語版が優先されます。
この日本語文書を引用する際には、下記の一文を記載してください。
引用：Pathways to Open Data: Findings from the 2024 World Open Innovation Conference Challenge Session参考訳（The Linux Foundation Japan提供）

翻訳協力：富田明男、富田佑実