科学技術動向研究

オープンサイエンスをめぐる新しい潮流（その4）

研究コミュニティに向けた協働データインフラの開発動向―欧州のEUDATの取組から―

野村 稔

概　要

現在、研究データの共有化、オープンアクセスの必要性が世界中で議論されている。欧州では、研究の遂行において、国境や学術領域を越えて自由にデータの利活用が行える汎用目的のデータサービスが不足しているという認識を持っている。この対応として、EU の FP7 のファンドを受けた EUDAT プロジェクトが 2011 年 10 月に発足している。本プロジェクトの目的は、研究コミュニティの内外において、研究者がデータを共有し、研究活動を効率的に遂行できるようにすることである。そして、13 か国の 26 機関を中心にしたコンソーシアムが構築され、複数の研究コミュニティを対象にした研究データに対する共通のサービスや運用方法の具体化が図られてきている。

内閣府では、研究データを中心としたオープンサイエンスに関する議論を開始しており、今正にオープン化に関わる世界的議論や動向の的確な把握が必要とされている。その実装面での一事例として、EUDAT の取組は我が国としても参考とすべきものである。

キーワード：EUDAT，e-infrastructure，研究コミュニティ，協働データインフラ，オープンサイエンス

1 はじめに

当研究所では、2014 年 4 月－10 月に第 10 回科学技術予測調査（通称：デルファイ調査）を実施した。本調査におけるデータに関係する課題は、全課題中の約 10% の 90 数件が設定されており、様々な分野に分布している。そのうち、データ基盤（データインフラ）やその活用・処理に関する課題は 65% を占めておりその関心の高さを示している。調査結果から実現時期を見ると、技術的実現は 2019～2027 年（中央値 2020 年）、社会的実装は 2020～2032 年（中央値 2025 年）と、比較的近い将来に実現され社会実装されるとの認識である。しかし、ICT・アナリティクス分野での各細目の重要度と国際競争力についてみると、ビッグデータ関連は、重要度は比較的高いが国際競争力は余り高くないという結果がでている。データへの対応の重要性を認識している一方で、その実現に必要とされる競争力は予想外に低いことが浮き彫りになった形である[1)]。

現在、研究データの共有化、オープンアクセスの必要性が世界中で議論されている。その必要性についての背景や動向についての詳細は、最近本誌で発表した記事を参照されたい[2~4)]。

本稿では、研究データの共有化に対する具体的なサービスの実装面に焦点をあて、複数の研究コミュニティへの対応として欧州で推進されている Collaborative Data Infrastructure（CDI：以下、協働データインフラ）の開発プロジェクト EUDAT（European Data Infrastructure）を取り上げ、その具体化に至る過程、提供されつつあるサービスについて紹介し、注目すべき諸点を探る。

2 EUDATプロジェクト

EUDATプロジェクトは、欧州で実施されている研究プロジェクトや研究者のニーズに適合した協働データインフラを提供することで、研究コミュニティの内外において、研究者が地理的及び学術的な境界を越えてデータを共有することにより研究活動を効率的に遂行できるようにすることを目的としている。

2-1 設立の背景

EUDATの起源は、PARADE（Partnership for Accessing Data in Europe）イニシアティブの活動に遡る。PARADEは、2009年10月に欧州のデータインフラ戦略に関するホワイトペーパーを発行した。ここで示された欧州の共有インフラの概念は、多くの政策機関や専門機関によって支援され詳細化が行われた[5]。

これと並行して、2009年後半に欧州連合（EU）の競争力評議会（EU competitiveness council）が、欧州委員会（European Commission：EC）に対し、科学のためのICTをベースとするインフラ（e-infrastructure）に関する今後の課題と対応につき検討を依頼している。これに応えた形で、アカデミア、研究機関、データセンター、産業界などのメンバーからなるハイレベル専門家グループが設けられた。このグループは、ECの要請により、科学データのためのe-infrastructureの展開に向けた「ビジョン2030」を策定し、2010年10月に報告書「Riding the wave」[6]を提出した。ECはこの実現に向けたcall（公募）を実施し、その結果として、PARADEの概念をも包含したEUDATが選定されている。この報告書[6]には、今後の研究データへの取組の方向性が以下のように記載されている。

1）データに対する問題認識

多くの研究コミュニティは、増加の一途をたどるデータに対して、格納場所、検索方法、活用法などに関する課題に直面しており、独自のソリューションを生む傾向にある。結果として各ソリューション間で相互運用性を欠き、分野融合研究を阻害する状況をもたらす。

今までに、欧州では欧州グリッドインフラストラクチャ（EGI）[7]やハイパフォーマンスコンピューティング（HPC）システムの共同利用に関するパートナーシップ（PRACE）[8]によって、研究に必要な計算機リソースやその使用環境は充実してきたが、研究の遂行において国境や学術領域を越えて自由にデータの利活用が行える汎用目的のデータサービスが不足している。

2）描いたビジョン

データのシームレスアクセス、使用、再使用、信頼性をサポートするためのインフラの確立がますます重要である。将来は、データそのものが重要な資産となり、それを活用して様々な科学、技術、経済、社会の進展が可能となる。

3）必要とされる具体的アクションへの提言

緊急に必要とされる具体的なアクションとして、協働データインフラのための国際的フレームワークの開発、e-infrastructureへのファンドの追加、データの価値の測定法やその使用、新世代のデータ科学者の養成と国民の理解の拡大、データインフラを計画するグローバルレベルの高度なグループの設置などを挙げている。

2-2 今までの動き

EUDATプロジェクトは、FP7 e-Infrastructure Call9（WP11）からのファンディングを獲得している。このCall9の目的は、欧州におけるデータインテンシブサイエンスに必要な科学データに対する持続的でロバストなインフラの構築であり、ファンド総額は4,300万ユーロである。

EUDATには、ECからCall9の最大予算である930万ユーロが授与され、ファンディング期間は3年間、その他の出資と合わせて合計予算額は1,630万ユーロとして2011年10月1日に開始している[9]。（2015年以降については後述）

このプロジェクトは、開発・利用側として、EUからファンドを受けた13か国の26の参加機関（EUDATはパートナーと称している）によって構成されており、図表1に示すように各国のデータセンター、HPCセンター、テクノロジプロバイダ、ファンディング提供機関、コミュニティなどが含まれている[10]。さらに、ファンドは受けてないが、このプロジェクトを取り囲む形で、その利用を指向する、あるいはプロジェクトに興味を示す広範な学術分野からの30のコミュニティが別に設定されており、生物医学、環境科学、人文社会科学、物理科学・

メニューへ戻る

図表 1　EUDAT の参加機関（パートナー）

国	パートナー
オーストリア	オーストリア環境庁
チェコ共和国	プラハ・カレル大学
フランス	CINES（高等教育のための国立スーパーコンピューティングセンター） CERFACS（研究機関） Gnubila（企業）
イタリア	CINECA（イタリア計算機センター間大学コンソーシアム） INGV（地球物理学と火山の国立研究所）
スペイン	BSC（バルセロナスーパーコンピューティングセンター） RedIRIS（スペインのアカデミック・研究ネットワーク）
オランダ	SURFsara（国立HPC・e-Scienceセンター）
ポーランド	PSNC（ポズナンスーパーコンピューティング・ネットワーキングセンター）
スウェーデン	SNIC（スウェーデン国立コンピューティングインフラストラクチャ）
ノルウェー	UNINETT（ノルウェー国立研究教育ネットワークのための国有会社）
フィンランド	CSC（国有ITインフラストラクチャの維持・開発を担う国有非営利会社）
英国	STFC（科学技術施設協議会） EPCC（エジンバラ並列コンピューティングセンター） UCL（ユニヴァーシティ・カレッジ・ロンドン） Trust-IT Services Ltd（企業）
スイス	CERN（欧州原子核研究機構）
ドイツ	DKRZ（ドイツ天候コンピューティングセンター） Forschungszentrum Jülich（FZJ:ユーリッヒ研究センター） マックスプランク研究所（Psycholinguistics：心理言語学） マックスプランク研究所（Meteorology：気象学） カールスルーエ工科大学 エバーハルト・カール大学テュービンゲン RZG（Rechenzentrum Garching：ガルヒング計算センター）

出典：参考資料 10 を基に科学技術動向研究センターにて作成

工学、材料科学、エネルギーなどの分野から各複数コミュニティの参加がある[11]。

そしてこれらを共に合わせてコンソーシアムを形成し、フィンランドの CSC-IT Center for Science（略、CSC）が主導している。

開発スケジュールとしては、2012 年に最初のサービス、2013 年にコミュニティ横断のサービス、2014 年に完全なサービスの配備をマイルストンとして位置付けていた[9]。プロジェクトは、7 つのワークパッケージに細分化され、相互に連携をとりながら推進された。

プロジェクトは、既に 3 年を経過しており、複数の研究コミュニティを対象に、調査、試行を経て、共通サービスと運用法の具体化が図られてきている。そして、現在、対象とする研究コミュニティを増やしながら、試行やトレーニングを実施し、より多くの研究活動に適合するサービス内容の充実に努めている。

2-3　具体化したサービス内容

以下に、サービスの開発過程で考慮された諸施策、開発したサービス、運用形態などについて示す[12]。

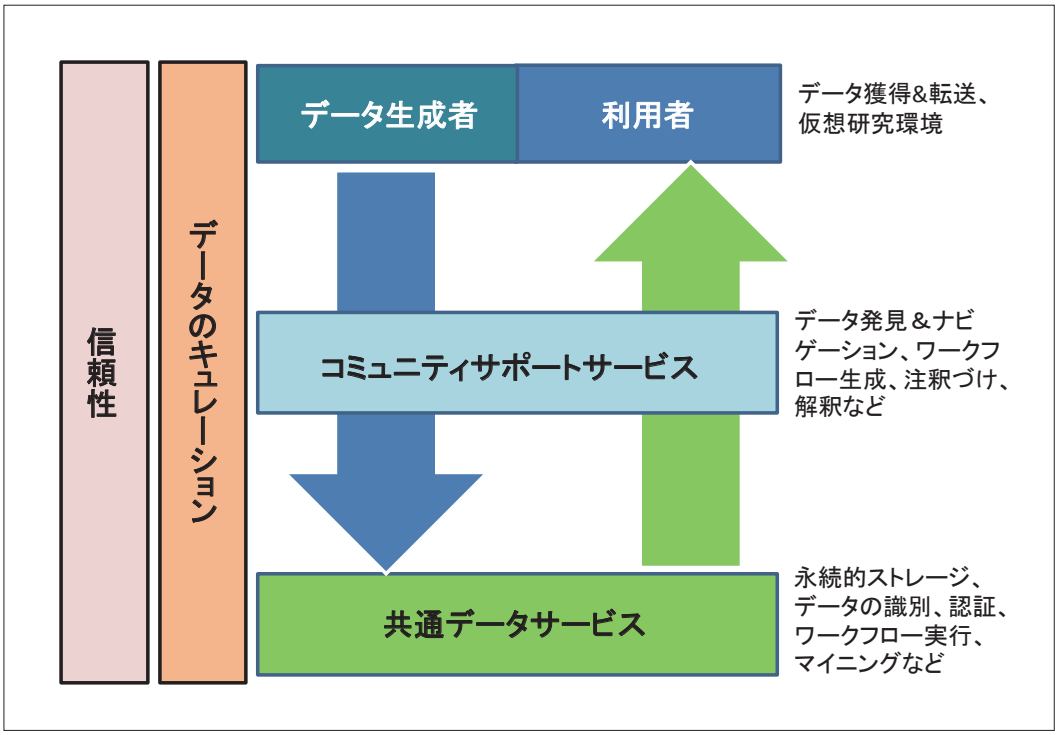
1）複数コミュニティに共通なデータサービス

図表 2 は、「Riding the wave」[6] で示された協働データインフラに対する考え方であり、これが EUDAT で目指す姿となっている。

ここでは、データ生成者と利用者は、データの獲得、転送、処理などを、所属するコミュニティが提供するサポートサービスを利用して行い、それらのコミュニティサポートサービスは、異なった分野間で横断的に使用可能な共通データサービスに依存するという階層関係をとっている。そして、全体を一つの系と考えており、この系全体にわたってデータのキュレーション（収集した情報を特定のテーマに沿って編集し、新たな意味や価値を付与する）や信頼性の確保を必要としている。すなわち、各階層での活動主体（アクター）間において必要とされるあるべき協働の形ともいえる。

EUDAT は、「この協働データインフラは、科学コミュニティへ一般的なサービスを提供することで、それらのコミュニティの学術分野に固有なサービスへの取組に、より多くの時間や投資を集中することを可能にする。また、個々の研究者、小さなコミュニティ、そして目的にかなったデータ管理が不足しているプロジェクトに、共通データサービスを提供し、そのインフラ開発に要する設備投資の必要性を取り除く」としている。

図表2　協働データインフラの姿



出典：参考資料6他を参考に科学技術動向研究センターにて作成

異なった学術分野の研究コミュニティは、データ構造やコンテンツが異なるため、固有の対処法をとっているのが一般であるが、同時に多くの基礎的なサービス要件も共有しており、この共通的な性質が複数の研究コミュニティのサポートに向けた共通データサービスの構築を可能にするとEUDATは述べる。

2）コミュニティとの連携作業

EUDATは、協働データインフラに求められる要件の明確化に向け、幅広い分野の研究コミュニティと連携して作業をしている。最初は、プロジェクトパートナーであるCLARIN（言語関連）、EPOS（固体地球科学）、ENES（気候科学）、LIFEWATCH（環境科学）、そしてVPH（生物医学）などのコミュニティを対象にし、それらのコミュニティで採られているアプローチとサービス要件を調査することから開始している。具体的には、コミュニティの代表者とのインタビューや頻繁なやりとりを通し、数か月後に優先的に開発すべきサービスとして、①サイトからサイトへのデータレプリケーション、②HPC施設へのデータステージング、③メタデータの整備、④使用容易なストレージ、の4つを特定している。

また、協働データインフラの充実に向け、他の多くのコミュニティとも連携している。

3）サービス内容

協働データインフラを構築するコアサービスとして、シングルサインオン、永続識別子（persistent identifier：PID）サービス、ウェブ実行・ワークフローサービス、モニタリング・アカウンティングサービスほかを要素として設け、それらを包含して図表3に示す5つのサービスを開発した。これらのサービスは、現在、運用中であり、さらなる機能強化も計画されている。

また、コミュニティによって用意され、EUDATとして提供される拡張コアサービスとして、共同メタデータサービス、共同データマイニングサービスなどが予定されている。

4）運用形態

①　運用リソース

EUDATに加盟するデータセンターはEUDATノードと呼ばれ、EUDATストレージを提供している。現状でのストレージ量に対する明確な言及はないが、EUDATの活動開始から1年後、実稼働に先だって試行的な運用環境が構築されており、その構成としては、480テラバイトのオンラインストレージと4ペタバイトのニアライン（テープ）ストレージを提供する5サイト（RZG, CINECA, SARA, CSC, FZJ）の記述があり、最初に4利用者コミュニティ（ENES, EPOS, CLARIN, VPH）にむけてサービスされた。実際の運用環境では利用者の必要に応

メニューへ戻る

図表 3　サービス内容

サービス名	機能
B2FIND	EUDATデータセンターと他のリポジトリ内に格納されている研究データを検索するポータル。研究データのメタデータカタログにより科学データを、素早く容易に発見し表示するサービス
B2SAFE	コミュニティや部門のリポジトリに研究データに関するデータ管理ポリシーを実装し、データのレプリケーション（複製）を可能とするサービス
B2STAGE	EUDATストレージと外部HPCシステムの作業領域間で大規模な研究データを転送するサービス
B2SHARE	様々なコンテキストからなる小規模の研究データをストレージへ蓄積、共有させるサービス。データの長期間の持続性を保証して保護
B2DROP	研究データの複数のバージョンの同期化、更新と、研究者間でのデータ交換を可能とするサービス

出典：参考資料 9 を基に科学技術動向研究センターにて作成

じたリソースが準備されると思われる。

② HPC アクセス

多くの研究の遂行では、強力な計算能力をもつ HPC システム上でシミュレーションを実施することがしばしば必要になる。その場合、研究データを HPC システム上で処理できるように移動し、処理結果のデータを移動元に戻すことが必要となる。ここでは、そのことをステージングと呼んでおり、大規模データを EUDAT ストレージと HPC 施設、例えば、欧州の PRACE の HPC システムなどとの間でやり取りするためのサービスである。一連のフローを見ると、ある研究コミュニティからの研究データは、まず EUDAT ノードのストレージにレプリケーションされる。その後、その EUDAT ノードの近隣かリモートの HPC 施設の作業用領域へ移され、HPC 処理後に結果が元の研究コミュニティに戻ることが容易にできるようなサービスが提供されている。このサービスでは、研究データのレプリケーションを伴うが、PID を駆使して全ての複製物の追跡可能性が担保されている。

③ データの可視化と再利用可能性

異なる学術領域の研究データを 1 つの協働データインフラで利用可能にすることは分野融合の研究にとって非常に有益である。そのため、EUDAT は共同のメタデータ（データの説明を施したもの）カタログの開発に取り組んでおり、それを用いることで容易に分野横断的な検索・表示を可能にしている。

5）その他の活動

現在、今後のサービスの拡張に向け、ダイナミックデータやワークフローサポートへの対応などを視野にしたワーキンググループを設置して検討を進めている。また、トレーニングを重要視しており、利用者に対し協働データインフラの最適な使用、操作法の習得を促している。さらに、今後の持続的な運用を目指したコストモデルの検討を重要な要素と位置付けている。

3　注目点

EUDAT で実現されつつあるソリューションのイメージ（著者が理解する範囲で想定）を図表 4 に示す。以下、注目点を述べる。

1）複数のコミュニティへ向けた共通サービス

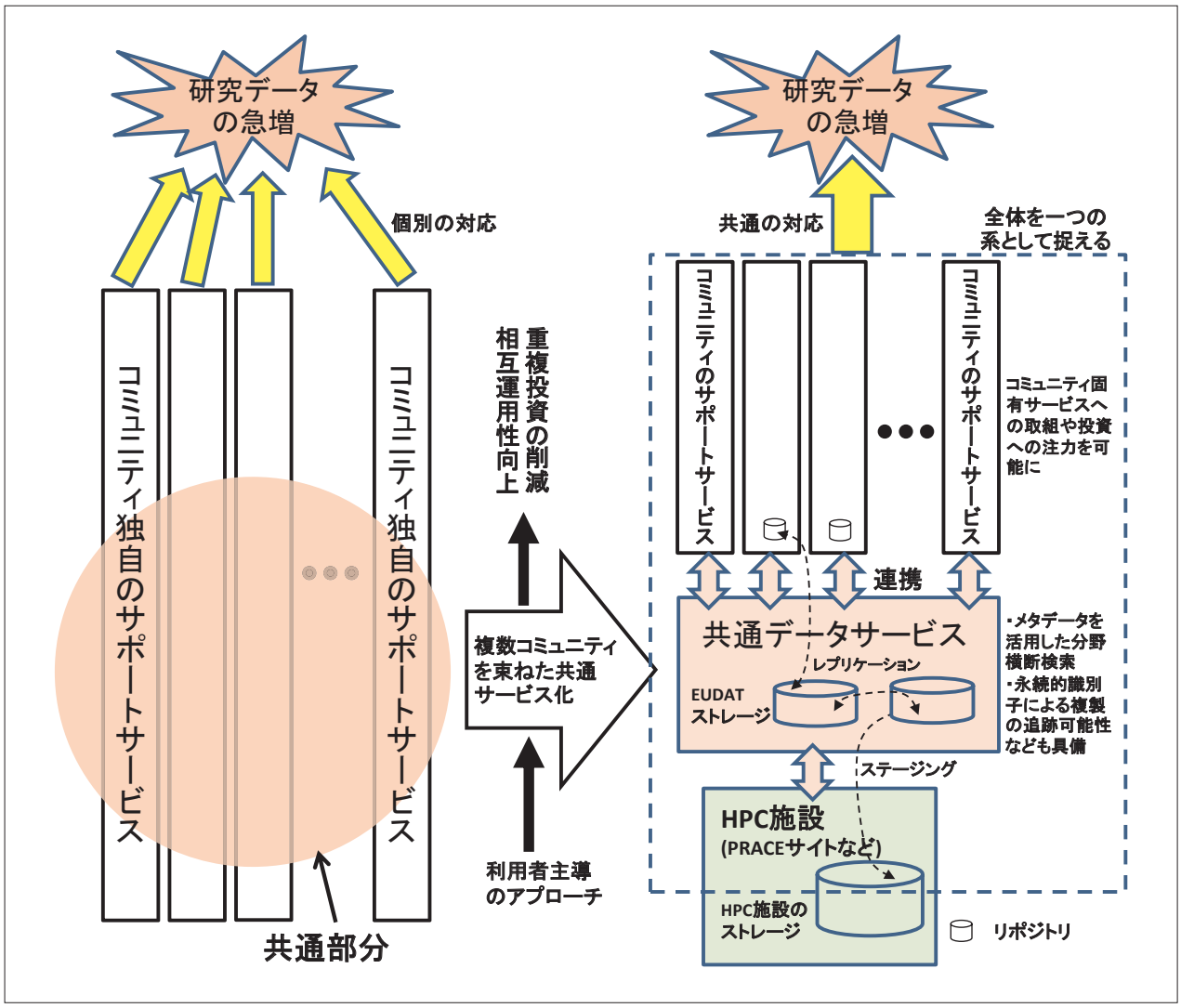
研究データの急増に対し、研究コミュニティでの独自サービスの提供は、重複投資を生むことはもちろん、コミュニティ間での相互運用性に支障をきたし分野融合研究を阻害する一因にもなる。この問題への対応には、複数コミュニティを束ねた共通データサービス化が重要であり、EUDAT の発想はまずここにある。そのために、異分野の複数コミュニティ（6 コミュニティ）をプロジェクト内に最初から巻き込んでサービス要件の抽出を行い、優先度付けを図りながら具体的なサービスの実現に結びつけている。こうした推進法は特に複数コミュニティ向けのサービスの実現では参考にしたい。

2）利用者主導のアプローチ

EUDAT では、利用者主導によるサービスの実現を目指しており、利用者との接点を多くして、必要とされるサービス要件を抽出することに努力している。そして具体的な開発の後は試行を経てサービスの洗練化を図っている。そのため、非公式な利用者との議論をはじめ、全利用者を対象としたユーザーフォーラムを複数回開催している。

EUDAT の関係者は、ユーザーフォーラムは、コミュニティの構築とステークホルダー間の信頼確立のために不可欠であるとし、「研究コミュニティ

図表4　EUDATのソリューションイメージ



出典：参考資料9～13を参考に科学技術動向研究センターにて作成

との会議や彼らのニーズを聞くことは、EUDATの正に中核である」とも述べている[13]。

3）HPC共通インフラリソースとの整合

欧州での研究インフラストラクチャであるPRACEは、既に世界レベルの性能をもつスーパーコンピュータを6システムも配備しており、その下位レベルのHPCシステムとも合わせて欧州全域でのリソース共有利用が実現されている。このPRACEのHPC施設（スーパーコンピュータを所有するセンター）もEUDATのパートナーの一部を構成しており、前記の運用形態の例でも示したが、PRACEの設備を生かすソリューションがサービスの実装で大きく配慮されている（図表4）。HPCを活用した高度な分析は様々な分野での基礎となりつつあり、その活用を促す使用容易性を確保するサービスを必須の要件としている。

4）持続性のあるオペレーションの追及

永続的な研究データの保管を伴うEUDATでは、持続性は特に重要と位置付けている。ECからのファンディングやメンバー各国からの支援だけでは、持続したサービスは難しい。EUDATは、次に向けての新ファンドを獲得しているが、それとともにサービス収入を得て、継続してサービスの洗練化を図れる好循環モデルを検討中である。

この課題については、出版者を中心としたデータ出版がデータジャーナルの創刊という形で始まったという報告もあるが[4]、正に手探りの発進ともいえ、今後の重要な検討要素である。EUDATの関係者は、今までの大きな成果として、欧州の主要なコミュニティと連動ができてきたことを挙げている。これは、EUDATがコミュニティの信頼を獲得していることを意味しており、この新モデルの実現も期待したい。

メニューへ戻る

4 おわりに

EUDAT の 2015 年以降の活動については、後継プロジェクト（EUDAT2020）が設定されており、Horizon 2020 プログラムからのファンド額として約 1,900 万ユーロ、期間は 3 年間、パートナー数は 33 として、2015 年 3 月 1 日から公式に開始している。

EUDAT は、現在、研究データの共有、利活用を先導するイニシアティブである RDA（Research Data Alliance）を支援しており、今までに開発したサービスを具体的な実装事例として貢献をはかることを志向している。EUDAT の関係者は、学術刊行物と比較するとき、研究データのオープンアクセスは発展途上であり、その実装が難しいことを挙げており、個々の活動ではなく全体システムとして支える必要性があるとも述べている[14]。また、2－1 で示した報告書の後継として新しい報告書が作成されており、データへの対応に向けた EC の力強い支援を再度要求している[15]。

本稿では、研究データを複数の研究コミュニティ間で、いかに共有、管理するかの課題に向けた実装例として EUDAT プロジェクトを紹介した。内閣府では、研究データを中心としたオープンサイエンスに関する議論を開始しており[16]、今正にオープン化に関わる世界的議論や動向の的確な把握が必要とされている。本稿で紹介した動きが今後の一助となれば幸いである。

参考文献

1）「第 10 回科学技術予測調査結果速報 全体概要」、科学技術・学術政策研究所、2014 年 11 月：http://www.nistep.go.jp/archives/1874
2）村山 泰啓、林 和弘「オープンサイエンスをめぐる新しい潮流（その 1）科学技術・学術情報共有の枠組みの国際動向と研究のオープンデータ」科学技術動向 , No.146, 2014 年 9 月 , p12-17：http://hdl.handle.net/11035/2972
3）村山 泰啓、林 和弘「オープンサイエンスをめぐる新しい潮流（その 2）オープンデータのためのデータ保存・管理体制」科学技術動向 , No.147, 2014 年 11 月 , p16-22：http://hdl.handle.net/11035/2990
4）林 和弘、村山 泰啓「オープンサイエンスをめぐる新しい潮流（その 3）研究データ出版の動向と論文の根拠データの公開促進に向けて」科学技術動向 , No.148, 2015 年 1 月 , p4-9：http://hdl.handle.net/11035/2999
5）Damien Lecarpentier「The EUDAT Project Towards a European Collaborative Data Infrastructure」, Oct. 3, 2011：http://www.verce.eu/Kickoff/Session1/VERCE-EUDAT.pdf
6）「Riding the wave –How Europe can gain from the rising tide of scientific data」, Oct. 2010：http://ec.europa.eu/information_society/newsroom/cf/newsletter-item-detail.cfm?item_id=6204
7）「European Grid Infrastructure」：http://www.egi.eu/
8）「PRACE」：http://www.prace-ri.eu/
9）Kimmo Koski「EUDAT, BoF Session on e - Infrastructure for science in Europe」, ISC' 11 21 June 2011
10）「EUDAT」：http://www.eudat.eu/
11）Damien Lecarpentier「EUDAT Data Services, Tools &Knowledge 」, Nov. 11. 2014：http://e-irg.eu/documents/10920/248525/EUDAT+Workshop+Rome+2014/3e756ce6-669b-41f2-b75b-afde20f3709e
12）Damien Lecarpentier「EUDAT: A New Cross-Disciplinary Data Infrastructure for Science」、The International Journal of Digital Curation Volume 8, Issue 1, 2013
13）「Creating a pan-European data infrastructure」, July 23, 2014：http://www.isgtw.org/feature/creating-pan-european-data-infrastructure
14）「Open data – What do EUDAT communities really think about it?」Jan. 5, 2015：http://www.eudat.eu/news/open-data-%E2%80%93-what-do-eudat-communities-really-think-about-it
15）「The Data Harvest: How sharing research data can yield knowledge, jobs and growth」、RDA Europe Report, Dec. 2014：http://europe.rd-alliance.org/sites/default/files/report/TheDataHarvestReport_%20Final.pdf
16）「国際的動向を踏まえたオープンサイエンスに関する検討会」：http://www8.cao.go.jp/cstp/sonota/openscience/index.html

執筆者プロフィール

野村　稔

科学技術動向研究センター　客員研究官

企業にてコンピュータ設計用 CAD の研究開発、ハイパフォーマンス・コンピューティング領域、ユビキタス領域のビジネス開発に従事後、現職。スーパーコンピュータ、ビッグデータ、半導体技術、LSI 設計技術等の科学技術動向に興味を持つ。

メニューへ戻る