学術研究におけるデータ倫理と利活用の両立を支援するための次世代データプラットフォーム

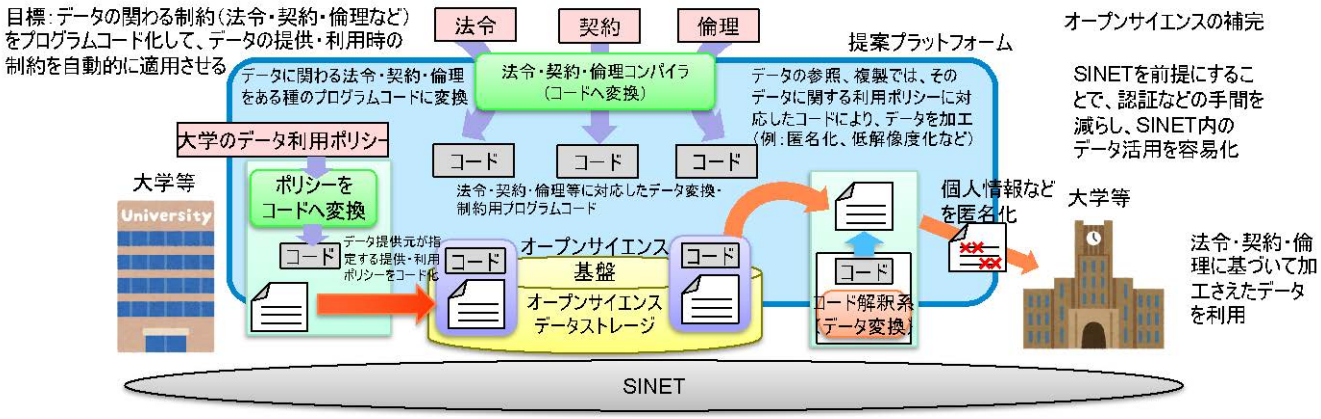
①　計画の概要

データ主導の学術研究、これまでの医学・理学・工学だけでなく、今後は人文学を含む広範な学術研究分野においても研究データの重要性と需要性が高まることが予想される。一方で研究データには、機密保持性に加えて、個人情報やプライバシー、知的財産を含む法令的及び倫理的な制約が課せられており、その制約を満足するように研究データを取扱うことは極めて煩雑であり、研究者の大きな負担となっている。また、オープンサイエンスなどにより、研究データの共用が期待されているが、これらの制約は研究データ共用の深刻な足枷となると予想される。

そこで本提案では、我が国では学術機関の大半が、国立情報学研究所が構築・運営する学術情報基盤となる SINET を利用していることを鑑みながら、SINET の接続機関において取得・保持されるデータに対して、法令的及び倫理的な制約を低減するデータプラットフォームを構築することで、前述の問題の解消を目指す。具体的には、研究データを格納するリポジトリやオープンサイエンス基盤に対して、データの保存範囲・期間の設定、データの改変・共用に関わるアクセス制御、統計及び匿名化などの加工の仕組みを実現するゲートウェイを付加する。また、共用されるデータに対する計算処理では、共用する側は計算処理を定義し、その結果を得られるが、処理対象のデータそのものへのアクセスを制限する。また、個人情報を含む研究データに関しては個人本人同意の代行する仕組みや、知的財産性を含むデータは引用に関わる情報や間接参照のリンク情報等を提供する仕組みを作り、さらに学術研究における著作物の利用状況をトレースできるようにして、著作物の不正な拡散を抑止し、利用許諾請求などを自動化していく。なお、本提案は SINET 及びオープンサイエンスの枠組みに付加かつ補完的な位置づけであると同時にオープンサイエンスの先を狙うものである。

②　学術的な意義

本提案は特定分野に特化せずに、データの利用における、データに関わる法令的及び契約、倫理上の制約による研究者への負担を軽減することを目的としている。特に今後、オープンサイエンスなどへの流れにより、研究データの第三者利用が求められるが、法令的及び契約、倫理上の制約が大きな足枷になることが予想され、その対策は学術界全体の課題となる。本提案そのものの学術的な貢献は、(1)データ利用に関わる法令や契約、倫理からコード化技術とそのコード化を行い、自動化が行えるようにする。自然言語かつ曖昧性を許容する法規定との齟齬が生じる可能性があるが、デジタルファーストに代表されるように、今後、各種手続きにおいてデジタル化が進展することが予想される状況では、コードとして法令や契約、倫理をコンピュータによる自動処理を実現することは、学術の範囲を超えて、社会全体にとっても大きな意味をもつ。(2)通信と計算、そしてデータを融合した次世代プラットフォームを実現する。既存の SINET は通信に特化し、リポジトリやオープンサイエンスはデータに格納に特化し、スパコンなどは計算に特化しているが、データの円滑な利用はそれぞれがシームレスに接続される必要がある。本提案は機関内及び機関間のデータのアクセス制御及び加工などを行う仕組みが、その仕組みによりデータごとに通信範囲を限定し、さらに計算範囲や手法などを統合的に管理できるようになる。これはオープンサイエンスの先にある研究基盤として重要な位置づけとなる。(3)現在、論文などの大きなコンテンツ単位に ID が割り振られているが、部分的な文章や図に対しても ID を割り当てる技術を提供するとともに、SINET 内におけるデータに関してその複製や引用をトレースする仕組みを構築していく。これは論文などにおける剽窃の研究不正対策としても有用となる。



図：学術研究におけるデータ倫理と利活用の両立を支援するための次世代データプラットフォーム

③　国内外の動向と当該研究計画の位置づけ

研究データの保存・利用は重要となるために、従前より研究機関により研究リポジトリが提供され、また、昨今はオープンサイエンスに動きにより、広範なデータが特定分野及び分野横断的に共用するためのデータプラットフォームが、国立情報学研究所などから提案・構築が進んでいる。ただし、リポジトリ及びオープンサイエンスは、研究データの格納とメタデータな

どによる検索に主眼を置いているのに対して、本提案は保存すべきデータと共用されるデータについて、法規制や倫理、契約からアクセス制限や加工することを目的にしており、リポジトリ及びオープンサイエンスなどと重複せず、補完的な位置づけとなる。また、コンピュータによる法規制などの支援に関しては、自然言語処理を駆使した法律や判例整理が進んでいるが、本提案は法制度や倫理、契約をコード化して、コンピュータが直接的に扱い、法制度そのもののデジタル化を狙うものとなる。知財に関してはコンテンツに対するアクセス制御、つまり参照・変更などの可否だけでなく、本提案は統計や匿名化などの加工を含めて取り扱うという差異がある。

④ 実施機関と実施体制

日本学術会議「情報学委員会 IT の生む諸課題検討分科会」との協力の下で勧める。また実施に当たっては国立情報学研究所が中心になって行うことを想定している。他機関の体制であるが、法制度に関しては法学関係者の協力が必要であり、東京大学法学部教授（憲法）、慶應義塾大学法学部教授（法哲学）を介してそれぞれの学部と協議中である。また、データプラットフォームに関しては、SINET の接続先である商業クラウドコンピューティング環境を提供するベンダーとも協力しながらすすめることを想定している。なお、本提案に関わるメンバーは、国立情報学研究所に設置された改正個人情報保護法で導入された匿名加工情報の加工基準の策定 WG が主体となっているが、同 WG による提案は、個人情報保護法を所掌する個人情報保護委員会の事務局レポートとして政府方針となっており、情報に関わる法制度には実績がある。

⑤ 所要経費

データのアクセス制御のためのサーバ、コード化した法令や契約、倫理の解釈実行するサーバ、SINET 及びオープンサイエンスの接続基盤、また一時的にデータを保持するためのストレージが中心となる。サーバなどの設備品に約８億円（サーバｘ200 台及びネットワーク機器）、接続基盤に約３億円（サーバ×10 台及びネットワーク機器類）、システム開発に約２億円、研究費（人件費他）に４億円を相当している。

⑥ 年次計画

１年目はデータプラットフォームに関わる要件定義を行う。本提案は複数分野の研究データを扱うことから、人に関するデータの取得が予想される分野、例えば社会学などの科学者コミュニティに対してヒアリングを行う。また知財に関しては論文などへの知財利用とオンライン授業における著作物引用を例に要件を調査する。初年度は要件整理及びヒアリングのための人員として８人の人材を登用予定である。なお、ヒアリングは民間組織の委託も検討する。

２年目は前年にまとめた要件に従ってデータプラットフォームの基本設計を始めるとともに、法令や契約、倫理をコード化するための記述言語体系の設計を法学研究者の協力を得ながら行う。これは機械的な解釈・生成ができる契約となる。設計に関わる人材を登用及び外部委託を行う。

３年目はデータプラットフォームの実装を進めて、一部の学術分野に関しては当該分野の科学者によるモデル的利用を開始する。また、法令や契約、倫理のコード化に関しては、提案したコード記述言語により、個人情報保護法及びそのガイドラインや著作権法に関してコードして、その実行、つまり法制度などの解釈を行う。３年目と４年目は開発・実装のための人材が必要となる。

４年目はデータプラットフォームの評価を行う。複数分野の科学者コミュニティに利用を依頼して、整備を進める。また法令や契約、倫理に対応したコードの実行基盤を構築して、データのアクセス制御や加工を評価する。

５年目は複数分野の科学者コミュニティに利用を行い、知的財産に関するトレースなどについても開発と評価を行うとともに、対応できるデータの種類やコード化する対象となる法令や契約、倫理を広げる。コード化及び多様なデータに対応する人材が必要となる。

なお、共同利用においても運用などで人材が必要であり、要件定義や設計、開発に関わる人材の一部は実施後も雇用を予定している。

⑦ 社会的価値

本提案の受益者は科学者コミュニティとなるが、研究データに関わる法制度や倫理を含む社会的要請を満足させる科学技術を推進することから社会的価値が大きいといえる。特に人に関わる実験などで被験者がその研究データによって、法的及びプライバシー的な権利利益の侵害を最小化し、知財に関しても本プラットフォームにより著作権の侵害などを未然に防ぐ仕組みとなりえる。今後、デジタルファーストの推進により、社会のデジタル化が進むが、そのとき本提案は法令や契約、倫理をコードとして表現できるようにするが、それは法令や契約、倫理を従来の自然言語として表現し、それを人間が理解するという間接的な形態から、情報システムそのものが法令や契約、倫理に従って実現され、同時に契約などはシステムそのものが生成し、他のシステムとの連携することが可能になることから、社会全体の在り方をかえることも期待できる。

⑧ 本計画に関する連絡先

佐藤　一郎（国立情報学研究所）