資料 3-2
第 12 回科学技術情報整備審議会
令和 2 年 9 月 30 日

令 和 2 年 9 月 30 日
科学技術情報整備審議会
基 本 方 針 検 討 部 会

第五期国立国会図書館科学技術情報整備基本計画策定に向けての提言（素案）
―「人と機械が読む時代」の知識基盤の確立に向けて―

I　本提言の位置付け

本提言は、国立国会図書館における今後 5 年間を目途とした科学技術情報の整備の在り方についての基本方針を提言するものである。しかしながら、急速な技術の進歩などを想定し、意図的にやや長期的な展望を意識したものとしたものである。デジタル技術が生活のあらゆる局面に浸透し、蓄積・流通するデジタルデータを AI 等により利活用する社会の変革が進みつつある今日、国立国会図書館は、「人」だけでなく、AI 等の「機械」も「読者」とするときを迎えた。デジタルを前提とした社会に向けて、国立国会図書館は、我が国唯一の国立図書館として、あるいは国会に附属する立法補佐機関として、いかに科学技術情報を整備していくべきか、換言すれば、どのようにこれからの時代の知識基盤を確立していくべきかという観点から、本提言をまとめた。

なお、今年改正された科学技術基本法[1]及び国による次期（第六期）科学技術基本計

[1] 科学技術基本法等の一部を改正する法律（令和 2 年法律第 63 号）により令和 3 年 4 月 1 日から科学技術・イノベーション基本法に改められる。

画策定に向けた議論においては、自然科学のみならず、人文学・社会科学を含むおよそあらゆる学問の領域を対象とし、Society5.0 の実現に向けて、人文学・社会科学と自然科学の知見を横断的、総合的に活用して、SDGs や AI の倫理等の現代的諸課題の解決やイノベーションの創出に当たろうとしている。本提言においても、この定義及び方向性を共有している。

II　基本的な視点

1　研究・社会のデジタル・シフト

(1)　データ駆動型研究の進展

近年の「デジタルトランスフォーメーション（DX）」とも呼ばれる、デジタル技術を通じた社会の変革は、図書館を含めた学術情報全般の流通・コミュニケーションの構造変化を引き起こしている。具体的には、国際的にオープンサイエンスの概念が普及し、学術ジャーナルのデジタル化・オープンアクセス化という従来の流れを推し進めただけではなく、在野の研究者等を巻き込んだシチズンサイエンスの広がり、研究サイクル・研究プロセスにおけるデジタル化の一層の浸透や、データを出発点として仮説・モデル・知識を生成する研究、すなわち「データ駆動型研究」の進展をもたらしている。

このデータ駆動型研究はあらゆる分野において展開されており、自然科学はもとより、人文学・社会科学においても研究スタイルを変化させ、新たな研究の可能性を開いた。情報学や統計学の知見により研究手法を革新するとともに、人文学分野で生み出される大規模データを他分野に導入して新たな研究課題や知識を得ようとする近年の「デジタル人文学」はその好例であり、文献等の文化資源の蓄積をいかすこの変革は図書館においても着目される。

(2)　新型コロナウイルス感染症拡大がもたらした「ニュー・ノーマル」

折からの新型コロナウイルス感染症（covid-19）の世界的拡大は、リモートワークや会議・イベント等のオンライン化を推し進めることとなり、学術情報流通・コミュニケーションの基盤のぜい弱性もあぶり出した。国立国会図書館や大学図書館による資料の閲覧・貸出し及び紙の複写物の提供といった、物理的な場所と資料等（フィジカル）に依拠したサービスの休止・縮小により、国内外の専門家や学生、一般市民への直接的な図書館サービスが途絶しただけでなく、我が国全体の教育・研究活動等の停滞を招いたことは、その顕著な事例である。

これは、国立国会図書館を含めた我が国の学術情報機関がこれまで行ってきた資料のデジタル化や IT を活用した遠隔サービスは、新型コロナウイルスが我々にもたらした「ニュー・ノーマル（新たな日常）」においては、デジタルトランスフォーメーションという面では結果的に不十分であったことを示すものと言える。換言すれば、フィジカ

ルのみに依拠したサービスは過去のものとなったのである。

(3) 海外の動向

米国著作権法最終 20 年条項や、欧州におけるデジタルアーカイブ促進のための権利制限規定やオープンデータ及びオープンアクセスの促進のための指令など、欧米諸国のデジタル化やオープン化等に係る法整備等は総じて我が国の先を行くものである。各国の国立図書館では、これらを踏まえつつ、学術リポジトリやオンライン出版プラットフォームの整備といったオープンサイエンスに係る取組や、クラウドソーシングによるデジタル化資料のテキスト化及び全文テキストデータを活用し、その成果を共有することで新たな価値創造を生み出す外部研究者との共同研究等の取組が進められている。

また、近年は、中国や台湾、韓国、シンガポールにおいても大規模な学術リポジトリやデジタルアーカイブが整備されている。例えば、中国では国家図書館が地方公共図書館の所蔵資料のデジタル化・共有を進めるなど、アジア諸国の進境には著しいものがある。

これらの取組も参照しつつ、我が国のあるべき姿を追求していく必要がある。

２ 第四期国立国会図書館科学技術情報整備基本計画の主な成果

国立国会図書館は「第四期国立国会図書館科学技術情報整備基本計画」（以下「第四期計画」という。）において、多種多様な資料・情報への長期的かつ広範なアクセスと利活用を可能とする基盤となる「深化型知識インフラ」の実現を目指し、様々なコンテンツを生み出し蓄積する「恒久的保存のための領域」と、コンテンツをより利活用しやすく整備する「利活用促進のための領域」の二つの領域の充実とこれらをつなげる役割を果たす、としていた。領域別にまとめた主な成果は次のとおりである。

(1) 恒久的保存のための取組

〇デジタル化の推進

平成 29 年度に科学技術関係資料を対象としたデジタル化に係る予算を措置して当該予算を倍増し（約 2.3 億円）、図書、雑誌（和洋の国内学協会誌を含む。）及び博士論文を中心にデジタル化を着実に進めたほか、録音資料（カセットテープ、SP レコード）・映像資料（レーザーディスク）等のデジタル化にも着手した。ただし、デジタル化は所蔵資料のうち和図書・和雑誌等の 5 分の 1 程度にとどまっている。

〇電子情報資源の長期アクセス保証

電子情報資源の長期保存・長期利用保証に係る取組として、NII-ELS（国立情報学研究所電子図書館事業）で維持困難となった学術情報等の保存や、WARP（国立国会図書館インターネット資料収集保存事業）の収集範囲の拡大、パッケージ系電子出版物の長期保存に係る調査、USB・MO 資料等のマイグレーション作業を行った。また、関連し

て、「国立国会図書館東日本大震災アーカイブ」（以下「ひなぎく」という。）において、存続が困難となった各地のアーカイブの承継に着手した。

○文献相当の国内情報資源の網羅的な収集

有償等オンライン資料収集については、関係団体との協議や実証実験を行い一定の進展は見られたものの、制度の確立には至っておらず、オンライン資料の流通増に対して収集できない範囲が広がっている。

(2) 利活用促進のための取組

○デジタル化資料の利活用促進

「図書館向けデジタル化資料送信サービス」の国内参加館を大幅に増加（約 400 館→約 1,200 館）させた。また、令和元年度から海外の図書館等への送信を行うこととなった。ただし、著作権保護期間の延長もあり、インターネット公開点数の伸びは頭打ちとなっている。

○テキストデータの活用

国立国会図書館の開発研究部門において、機械学習を活用したテキスト化精度向上の研究に取り組み、その途中成果として、デジタル化資料の全文検索機能を搭載した「次世代デジタルライブラリー」の公開、テキスト化したデータの学習用データセットとしての公開等を行った。また、日本点字図書館と協力して共同校正システムを用いたテキストデータ化実証実験や、「ひなぎく」での全文検索機能の一部提供を行った。

○多様なコンテンツのメタデータの統合的検索機能の提供

図書館だけでなく、公文書館や美術館、博物館等の多様な分野のデジタルアーカイブの統合ポータル「ジャパンサーチ」を、内閣府をはじめとする関係府省等と協力の上、国立国会図書館の開発研究部門が中心となって開発し、令和 2 年 8 月に正式版を公開した。

○メタデータのオープンライセンス化・標準化の推進

国立国会図書館作成書誌データの無償化（CC BY 互換）を行い、メタデータのオープンデータセットの提供を開始した。また、「ジャパンサーチ」の「共通メタデータフォーマット」を策定するとともに、メタデータのオープンライセンス化（原則 CC0）及びメタデータ API 機能による利活用促進を行った。このほか、国立国会図書館が所蔵資料をデジタル化したものへの DOI の付与、DC-NDL（RDF）の維持・普及を行った。

以上をまとめると、第四期計画期間中に「深化型知識インフラ」の構築に向けた取組は着実に前進していると評価できる。一方で、今後の課題としては、全文テキスト化を含めたデジタル化の推進や、ジャパンサーチの利活用促進、オンライン資料の収集範囲拡大、データのオープン化の一層の推進、電子情報の長期保存の本格的な実施等が挙げられる。これらを踏まえつつ、国立国会図書館は、次期計画において前節で述べたよう

な状況に対応する知識基盤の整備に取り組むべきである。

III 「人と機械が読む時代」の知識基盤の確立に向けた取組の方向性

1 全体の方向性

国立国会図書館は、その使命を果たすため、デジタル技術を活用し、また、ポスト・コロナのデジタルを前提とした新しい社会に適合した方法により、科学技術情報の整備を促進していくべきである。デジタルトランスフォーメーションによる社会変革を後押しし、少子高齢化や地方創生といった SDGs にもつながる我が国の課題の解決に貢献していくために、オープンで広く信頼され、大規模災害や感染症流行といった非常事態に対するレジリエンス（しなやかさ）を備えた国の知識基盤の整備に取り組まねばならない。この取組を推進することにおいてこそ、これからの社会において、人々に広く開かれ、国の知識基盤の一翼を担うという国立国会図書館の本来的な役割をよりよく果たすことができるのである。

その際、冒頭に述べた「人」と「機械」という二つの「読者」（利活用のチャネル）から逆算して、求められる国立国会図書館の取組を整理していくことを提案したい。具体的には、国内外の専門家や学生、一般市民といったあらゆる人々が読める（オンラインによるアクセスが可能な）環境を整備して調査、研究、教育等の場面で利活用してもらうという方向性と、国立国会図書館が潜在的・顕在的に保有するデータを AI 等の機械が読める（利活用可能な）形式で提供することでデータ駆動型研究に貢献するという方向性を、まずは提示する。この二つの方向性を実現するための、利活用促進と恒久的保存のための基盤の整備が、引き続き国立国会図書館の取組の核心となる。その際、これらの取組が、国立国会図書館の中心的任務の一つである立法補佐機能の深化・高度化にも資するものとすべきである。

利活用促進のための領域においては、全文テキスト化等を射程に入れた、国会情報を含む資料のデジタル化を戦略的に推進するとともに、メタデータを含むそれらのデータの組織化・オープン化に取り組むべきである。また、著作権処理の加速、「図書館向けデジタル化資料送信サービス」の拡大及び「ジャパンサーチ」等の拡充による情報アクセス環境の改善、並びに教育シーンでの利活用モデルの提示等にも取り組むべきである。これらの実現のためには、特に制度・技術面においては、外部の知見を取り込みつつ取り組むことも必要である。

恒久的保存のための領域においては、従来の資料収集の強化継続に加え、未収資料について、メタデータのみならず原資料をデジタルデータで収集して恒久的に保存するとともに、存続が困難となったデータベースやデジタルアーカイブ、分野横断的な研究データ等の承継にも引き続き取り組む必要がある。また、図書館関係のみならず様々な分野のデジタルアーカイブのメタデータの収集に、「ジャパンサーチ」等を通じ

て取り組むべきである。その際には、関係諸機関と分担しつつも、国立国会図書館が日本及び日本語に係る全体の知識基盤整備において主導的な立場を果たしていかねばならないだろう。

２　個別の取組の方向性

(1)　データのオープン化と教育等における利活用促進

国立国会図書館が収集・整備・保存するデータ（メタデータや典拠情報を含む。）は、公共財として人も機械も利活用（再利用）できる形式で可能な限りオープンな利用条件で提供されるべきである。仮にオープンにできない場合は、二次利用条件の整備が必須である。提供に際しては、本文検索の実現も含めて各情報資源に適切にユーザをナビゲートするような統合的なオンラインサービスの再構築や検索エンジン等民間によるサービスを通じてのアクセスを視野に入れた、あらゆる「人」に対するインクルーシブなインターフェイスの整備（障害者や高齢者等への対応を含む。）が必要である。同時に、API や機械学習等のためのデータセットとしての提供といった「機械」に対するインターフェイスのための取組の一層の充実も視野に入れるべきである。あわせて、例えば所蔵機関を跨いだ関連資料のひも付けや地理情報等他分野のデータとの連携等、コレクション間の相互連関のための基盤として、関係機関とも調整しながら識別子も含む典拠情報等を戦略的に拡張・整備していく必要がある。

また、物理的・地理的な制約を克服するためにも、著作権処理の推進によるインターネット公開資料の拡大にも一層取り組むべきである。あわせて、図書館資料の複製物のデジタル送信や絶版等資料の利活用（「図書館向けデジタル化資料送信サービス」）の拡大にも、権利者の利益保護にも配慮しつつ、文化庁をはじめ関係諸機関・諸団体とも協議しつつ取り組むべきである。データ駆動型研究等の学術研究を目的とする場合には、権利上制約のあるものも含めたデータの利活用を可能とする枠組みの整備も求められる。

一方で、ただ環境を整備するだけではデータの利活用が進まないのも事実である。これを進めるためには、国立国会図書館においてもフェローシップ制度や共同研究等により、データ駆動型研究の最新動向を踏まえた外部の知見や資源を取り込みつつ、例えば、教育シーン（オンライン教育を含む。）での利活用モデルの収集・提示や、データを用いた研究等を推進することが考えられる。

なお、国内の多様なデジタルアーカイブ資源の利活用という視点では、利活用のプラットフォームとしての「ジャパンサーチ」の更なる連携拡充は極めて重要である。「ジャパンサーチ」による分野や産官学等の垣根を横断した連携により、「国立国会図書館サーチ（又はその後継サービス）」により集約する書籍等分野のデジタルアーカイブ利活用への相乗効果、海外へのコンテンツ発信という視点でも成果を期待できることから、積極的に取り組むべきである。

(2)　資料のデジタル化・全文テキスト化等の推進

(1)で述べたデータの利活用を推し進めるためには、紙等のアナログ媒体に記載された情報のデジタル化が必須であることから、全ての所蔵資料のデジタル化を長期的には実現すべきである。しかし、当面は、国内刊行出版物について刊行年代や分野、資料群に応じた対象の優先順位付けを行いつつ、年代・資料群共に対象をより拡大して進めるべきである。一方で、国会の附属機関として、国会情報についてはより積極的にデジタル化を行い、国会と国民をつなぐ取組に努めるべきである。

しかし、人が読むにせよ、機械が読むにせよ、もはや従来の「画像化」だけでは不十分である。デジタル化資料については、全文テキスト化や画像抽出等の技術を基にした本文検索、画像検索といった所在検索に加えて XML 等による構造化等を行うべきである。あわせて、これらのデータについてデータセットとしての利活用も視野に入れるべきである。その際、OCR やレイアウト認識等の技術開発が課題となるが、開発したプログラムをオープンソースとして公開する等を通じて外部の知見・技術も取り込みながら技術的開発を進めていくための体制・環境の継続的・発展的な整備が必要である。また、関係諸機関・諸団体とも丁寧に協議を行いながら進める必要があるだろう。もとより、この取組は、視覚障害者等のアクセシビリティ改善にも資するものである。

また、デジタル化及び全文テキスト化の対象拡大に伴い、全文検索からのオプトアウト等による個人情報保護やプライバシーへの配慮等も必要となる。透明性が確保された手続と基本的な判断基準が求められよう。

(3)　多様な文化資源の収集・保存

(1)や(2)で扱うデータをより広く利活用してもらうためには、多様かつ大量のデータの収集・保存（長期アクセス保証）が欠かせない。国立国会図書館が日本及び日本語に係る多様な文化資源の収集を行うに際しては、国立情報学研究所や科学技術振興機構といった国内関係機関と分担しつつ、日本及び日本語に係る知識基盤、すなわち「ナショナルコレクション」を構築するという視点を持つ必要があるだろう。

まず、従来の有体物・オンラインの資料収集・購読については、継続・強化に努めるべきである。とりわけ、かねて課題となっている有償等オンライン資料の収集については、恒久的な収集・保存のための仕組みが存在しない現状のままでは我が国の知識基盤に大きな欠落を残すことになるため、早急な対応が必要である。ただし、当面の進め方として、学術情報について先行的に収集を開始する等、戦略的に推進する等の方向性も考えられる。また、外国刊行資料については、引き続きコアジャーナル中心の電子ジャーナル等の購読及び海外刊行の国内学協会誌の収集に努めるべきであるが、今後はオープンアクセスの進展に応じて、柔軟に見直す視点も求められるだろう。

また、未収資料の収集にも取り組むべきである。例えば、地域資料や海外機関所蔵の

日本関係資料等、国立国会図書館が未収の紙媒体資料については、デジタルでの収集を視野に入れて取り組むべきである（当然ながら、保存・提供も射程に入る）。さらに、研究サイクルのデジタル化やオープンサイエンス、EBPM（証拠に基づく政策立案）の推進を背景に、科学技術情報は、論文だけではなく、その根拠となる研究データも重要な要素となり、大学、研究機関、学協会等のデータプラットフォームの整備が進みつつある。研究データ等の科学技術情報に関して、国立国会図書館は、国立情報学研究所や科学技術振興機構等の関係機関と分担し、国全体として科学技術情報へのアクセスを向上させることが望まれる。この場合、国立国会図書館においては、他ではカバーできない分野、例えば、WARP の枠組みを活用したデータベースやデジタルアーカイブのほか、公的機関の研究データや政策データ、地域資料や特定の研究分野に結び付かない分野横断的なデータの承継等が課題と考えられる。国立国会図書館が、可能なところから承継条件の整理等を行い、アーカイブの支援に取り組むことに期待したい。また、民間ウェブサイトの制度的収集についても、我が国全体の課題として検討を進めるべきである。

メタデータについても、利活用を見据えた識別子の整備と併せて、「ジャパンサーチ」等を通じて引き続き収集すべきである。その際、デジタル化やメタデータの整備が進んでいない機関については、国立国会図書館が整備を支援することも検討すべきである。

その上で、これらの収集したデータへの長期アクセス保証を実現するための持続可能なアーカイブ基盤の整備に努めるべきである。国立国会図書館に対する信頼感は、この点の実行により担保されるものであろう。この分野においても、国立国会図書館がこれまでの取組で得た知見をいかして、長期保存に関する課題等の共有を可能とするコミュニティを醸成する等、各機関（各識別子の国内登録機関を含む。）への支援を検討すべきである。

Ⅳ　おわりに

これまでの図書館や昨今の様々なデジタルアーカイブが、私たちの知的な創造活動のプロセスの基盤となってきたことは、改めて指摘するまでもないだろう。では、「人と機械が読む時代」の知識基盤は私たちに何をもたらすのか。結論を述べれば、私たちの知的な創造活動の可能性を広げるものとして考えたい。

人が通覧することが不可能なほど多様で膨大なデータを扱うためには、AI 等の機械による分析は欠かせない。機械を読者とすることで、日本及び日本語に係るあらゆる分野の知見が集約された大量のデータの中から、人では気付くことが難しいパターンを見いだすことが可能となり、新たな研究、課題解決の可能性が開けていく。そして、見いだされたパターンの持つ意味の分析やそもそもの課題設定といった、機械がカバーできないが、課題解決に直結するような知的活動に、人はより注力することが可能となる。

本提言が、そのための基盤整備の一助となれば幸いである。