参考資料２

「AI 時代の知的財産権検討会　中間とりまとめ」（2024 年５月）（抜粋）

（57～59 頁）

（３）学習用データとしてのデジタルアーカイブ[52]整備

ア　具体的な課題

生成 AI の開発・提供・利用の促進には、生成 AI の開発に要する学習データの整備が前提として重要になる。

他方、美術館や博物館等のアーカイブ機関は、既に各種コンテンツのデジタルアーカイブを保有しているため、当該デジタルアーカイブを AI 学習の用に供することが考えられるところ、その意義についてどのように考えるかをまずは整理する必要がある。また、アーカイブ機関が各種コンテンツのデジタルアーカイブを保有していたとしても、当該デジタルアーカイブに係る保有データの権利者ではない場合も多いことから、アーカイブ機関が権利を有していない保有データを AI 開発等のために利用する場合において、アーカイブ機関が知的財産法の観点で留意すべき事項は何かについても検討する必要がある。加えて、アーカイブ機関によって、その保有するデータの技術仕様が異なっている場合もあることから、アーカイブ機関が保有するデータを AI 学習に供するために必要な技術仕様等についても整理する必要がある。

本検討会では、以上の諸点について検討の上、整理を行った。

イ　学習用データとしてのデジタルアーカイブ整備の基本的な考え方

意見募集では、アーカイブ機関の保有するデジタルアーカイブを生成 AI の学習に供することは、文化財の美の価値や精神性を失墜させることになるといった意見や、仮に政府がデジタルアーカイブを整備する場合には、完璧にクリーンなデータ（権利者の許諾を得たデータ）を集めることには限界があるのではないかといった懸念が見られた。

他方で、例えば、古写真の修復や文字からの肖像画の生成、原生生物から絶滅動物の画像化など、学術的な面での活用への期待を述べる意見や、国際競争力を強化するために日本国内での大規模なデータセットを構築することは不可欠であるとする意見、公共的なデジタルアーカイブの整備と適切な価格での頒布により、公共施設の維持管理費に充てることも可能ではないかといった意見など、デジタルアーカイブを整備することに意義を見出す意見も見られた。ただし、デジタルアーカイブ整備に意義を見出す意見の中においても、まずは国や地方公共団体が権利を有する文書等やパブリックドメインのデジタルアーカイブを整備し、公開する（ただし、プライバシーや秘密保持への配慮は必要）ことから始めるべきではないかという意見も見られたところである。

---

[52] 「デジタルアーカイブ」とは、一般的には、博物館・美術館・公文書館や図書館等の収蔵品を始め、有形・無形の学術・文化資源等をデジタル化して記録保存を行うことを指す。

これらの意見募集の結果に鑑みれば、デジタルアーカイブを AI 学習用データして活用することについては、アーカイブ機関が保有するデータの性格を踏まえ、各アーカイブ機関において、まずはパブリックドメインとなっているデータや適正に権利処理が完了しているデータ、国や地方公共団体をはじめとする公的機関が著作権等の権利を有している文書等を中心に据えて、デジタルアーカイブ整備を進めることを当面の基本的な考え方とすることが適当と考えられる。[53]

なお、デジタルアーカイブ整備等に関する主な関連規定は次のとおりであり、これらの規定を遵守しつつ、整備を進めることが肝要である。

【参考】アーカイブ整備等に関する主な関連規定

国立国会図書館法等（アーカイブ化関連）	
国立国会図書館法 24条～25条の4	【国立国会図書館】 →納本制度（24条（国の機関関係）・24条の2（地方公共団体の機関関係）・25条（左記以外）） →インターネット資料等の記録（25条の3（公的機関によるインターネット資料）・25条の4（非公的機関による電子書籍・雑誌等））
公文書管理法 15条～16条	【公文書館】 →特定歴史公文書等の保存等（15条（保存）・16条（利用請求及びその取扱い））
国立公文書館法 11条1項1号	【国立公文書館】 →業務範囲：「特定歴史公文書等を保存し、及び一般の利用に供すること」
博物館法 3条1項3号	【博物館・美術館】 →博物館事業：「博物館資料に係る電磁的記録を作成し、公開すること」
著作権法（権利制限規定：アーカイブ化関連）	
著作権法 31条	【図書館等】（国立国会図書館や公共図書館のほか、博物館・美術館を含む） →欠損・汚損部分の保管や損傷しやすい古書等の保存のための図書館資料の複製（1項2号） →他の図書館等の求めに応じるための絶版等資料の複製（1項3号） →公表された著作物（図書館等資料）について、非営利事業として事前登録者にコピー等制限をつけて行う、一部分の自動公衆送信及びそのため複製（2項）
	【国立国会図書館】 →納本された図書館資料の原本の滅失、損傷、汚損を避けるためのデジタル化による複製（6項） →特定絶版等資料の複製物について、事前登録者にコピー制限をつけて行う自動公衆送信（8項）
同法42条の３	【公文書館】→公文書管理法等に基づき必要な、歴史公文書等の保存のための複製及び必要な利用
同法43条	【国立国会図書館】 →国立国会図書館法に基づき必要な、国等のインターネット資料及び民間により提供されるオンライン資料の収集のために必要な複製
同法47条の5	コンピューター情報処理結果の提供に付随する軽微利用（※サムネイル画像、スニペット表示等）
著作権法（権利制限規定：学習用データセット整備関連）	
著作権法 30条の4	享受を目的としない利用（情報解析等）

ウ　AI 学習実施のために必要な技術仕様

本検討会による検討によれば、AI 学習用データとして利用するために必要なデジタルアーカイブデータの技術仕様は、次のとおりである。

- 開発する AI（何を目的とする AI か）によって、学習するデータの形式は様々で

[53] デジタル庁では、適切な日本語による大規模言語モデル（LLM）の開発促進に向けて、「AI 時代の官民データの整備・連携に向けたアクションプラン」（デジタル庁、令和５年 12 月 20 日）に沿って、生成 AI の学習に寄与する行政保有データのオープン化の検討等を進めることとしている。

ある。

- データセット専用の統一したフォーマットは不要である。ただし、プログラムで読み込めないフォーマットも存在するため、読み込みライブラリが存在するデータが望ましい(少なくとも、フォーマットの仕様は公開されている必要がある。)。

- AI 開発者側で必要なデータ形式に加工するため、学習用データの提供者側は、デジタル化するコンテンツに適正なデータの種類（画像・文章・音声等）のファイル形式で構築すればよい。
  - その際、必要なデータ形式への加工・アクセスを行いやすくするため、例えば、テキストであれば、テキスト対応のファイル形式で保存することや、ファイルの保護措置は解除しておくこと等が望ましい。なお、数式は、LaTeX 形式が望ましい。
  - AI 開発を内製する場合は、内部でのデータ加工は必要である。
- 学習したデータセットの品質により、AI 生成物に差異が生じることから、画像であれば高精細なもの、テキストであれば構造化されたテキストデータ等、リッチなデータとして構築することが望ましい。
  - 品質が劣るものについては、最新の技術動向を踏まえつつ、適宜必要な技術を用いながらデータの品質をリフレッシュしていくことが求められる。
- 学習したデータを判別することも見据えて、メタデータ（サムネイル含む）についても、分野で標準的に広く用いられているメタデータ形式によるメタデータの管理を行うことが望ましい。