恒久的保存基盤の構築に向けた技術調査報告書

国立国会図書館

平成 29 年 5 月

電子情報部電子情報サービス課次世代システム開発研究室

国立国会図書館電子情報部電子情報サービス課次世代室システム開発研究室では、電子情報の長期利用保証に係る調査研究の一環として、恒久的保存基盤の構築に向けた調査を実施している。本報告書では、恒久的保存基盤の構築に必要と想定される技術要素について、海外の研究動向を調査した結果を報告する。

（はじめに）

平成 27 年度に公開された、「イノベーションを支える『知識インフラ』の深化のための提言 ～第四期科学技術情報整備基本計画策定に向けて～ 」[1]では、「情報資源の恒久的保存を図り、人類の知的営為を未来に伝え、現在はもちろん、未来の利用者が意思決定や価値の創造に活用できるように整備する」ための「深化型知識インフラ」の実現について提言されている。また、第 186 回国会で可決された「著作権法の一部を改正する法律案」における参議院の附帯決議 [2]では、「我が国の貴重な文化関係資料を次世代に継承し、その活用を図るうえで重要な役割を果たすもの」として、「ナショナルアーカイブ」の構築に向けた取り組みを推進するとされている。

これらの動向を踏まえ、国立国会図書館はデジタルコンテンツについて、国内関係機関とも連携しつつ、恒久的なアクセスを保証する電子情報の保存基盤を実現することを最終的な目的として、「恒久的保存基盤」の構築に向けた電子情報の長期利用保証に関する調査研究を行っている。

本報告書では、恒久的保存基盤構築に必要な技術要素として、電子情報資源の長期保存に関する標準と技術動向、電子情報資源の利用提供に関する技術動向、多様な分野の機関が連携した分散型アーカイブについて、海外の研究動向を調査した結果を報告する。また、末尾の用語集には、本文中に登場する用語・略語について、長期保存に関連するものを簡単にまとめている。

1 http://dl.ndl.go.jp/info:ndljp/pid/9917300

2 http://kokkai.ndl.go.jp/SENTAKU/sangiin/186/0061/18604240061012a.html

目次

1　電子情報の長期利用保証に係るこれまでの調査研究

国立国会図書館では、電子情報の長期利用保証について、所蔵するパッケージ系出版物等を対象に調査研究を行ってきた。

平成 14 年度から 16 年度までに基礎的な調査研究として、国内の状況及び海外の研究・技術動向を把握し、保存の対象とその方法、海外のガイドラインの調査、国立国会図書館が所蔵するパッケージ系電子出版物の実態調査などを行った。

平成 18 年度から 22 年度までは録音・映像資料のデジタル化や、フロッピーディスクのマイグレーション試行、保存システムの構築に係る要素技術に関する調査を行った。

2　電子情報の脆弱性と長期利用保証

(1) 電子情報の脆弱性

電子情報は、紙媒体で出版された図書等と比較して次のような問題を抱えている[3]。

- 電子情報を利用するためには、それに対応する特定の再生機器やソフトウェアなどが必要であるが、これらは絶えず進歩し、古いものは使えなくなることが多い。
- 紙媒体の寿命とくらべて、記録媒体の寿命は著しく短い。
- インターネット上の情報は消失する可能性が高い。
- 情報の複製や改ざんが容易であり、オリジナルであることを保証することが困難である。

特にインターネット上で流通する情報は消失する危険性が高い。日本の府省ウェブサイトでは、5 年で 70%近くが、URL が消えたか又は URL が残っていても内容が全く同じではなくなったという調査報告がされている[4]。こうした初めからデジタル形式で生産された情報資源（いわゆるボーンデジタルコンテンツ）、あるいはアナログ媒体からデジタル化して生産されたデジタルデータ（画像ファイルやテキストデータ）を確実に未来に伝えることは、情報資源の多くがデジタル形式に移行している現代の大きな課題である。

そこで、以下ではデジタル形式の情報やコンテンツ[5]の長期利用保証について説明する。

(2) ビット列の保存と論理的な保存

電子情報の長期利用を保証するには、ビット列の保存と論理的な保存の二つの考え方がある。

ビット列の保存とは、媒体中のビット列が損失しないように、それらの物理媒体を適切な保管環境で保存することや、オリジナルの媒体からそのデジタルデータと全く同じビット列

---

3 http://www.ndl.go.jp/jp/aboutus/dlib/preservation/index.html

4 国立国会図書館. カレントアウェアネス-E. No.296 2016.01.21. E1757 - 日本の府省ウェブサイトの残存率：WARP における調査 http://current.ndl.go.jp/e1757

5 「電子情報」には、本来レーザーディスクなどアナログ形式で保存された情報も含まれるが、本報告書では以後デジタル形式で保存された情報の意味で「電子情報」を使用する。

のまま別の媒体にコピーすることである。しかし、ビット列の保存では再生環境が旧式化した場合には利用が保証されない。ビット列が保存されても、人がその内容を理解できる形式で再生できなければ、長期利用を保証することにならない。

そこで、ビット列を人が理解できる形式で再生可能とする方法についての知識も保存する必要がある。これが論理的な保存である。論理的な保存には「マイグレーション」と「エミュレーション」という方法がある。

マイグレーションは、オリジナルのビット列が再生環境の変化により再生できなくなる前に、ファイルフォーマットを変換する方法である[6]。マイグレーションはデータの改変を伴うため、適切に実施しないとオリジナルのファイルから情報が欠損する可能性があり、ファイルフォーマットに関する専門的な知識が必要である。そこでファイルフォーマットに関するデータベースである「フォーマットレジストリ」を構築し、その知識を共有することが有効と考えられる。

エミュレーションは、エミュレータというソフトウェアを使って旧式化した再生環境を最新の環境上に仮想的に再現し、オリジナルのビット列を再生する方法である。例えば再生に必要な環境（ハードウェア、OS、ソフトウェアに関する情報）を保存し、利用できるようにすることである。エミュレーションではオリジナルのファイルが改変される危険性は少ないが、再生環境を完璧に再現するのは難しい場合がある。エミュレータの開発にはハードウェア、OS、ソフトウェアなどの再生環境に関する専門知識も必要で難易度が高く、開発コストも必要となる。そこで各機関が開発したエミュレータを共有することで、開発コストを節約し、より再現性の高いエミュレータを開発することが可能となると考えられる。

(3) 電子情報の長期保存に関する規格

電子情報資源の長期保存については、長期保存の参照モデルである「Reference Model for an Open Archival Information System（OAIS 参照モデル。以下 OAIS と略記する。）」[28]が国際標準として制定されている。OAIS 自体は、具体的な技術仕様ではなく、概念整理のためのモデルであり、アーカイブが果たすべき責任や機能要素（エンティティ）などを定義しているものである。その中で特徴的なのは、アーカイブがサービスを提供する特定のコミュニティ（Designated Community）を明確にして、対象コミュニティが長期にわたって情報にアクセスし、他者からの手助けなく情報を理解できるように情報を保存すべきという主張がなされている点である。

OAIS では、「コンテンツ」（保存対象とするデータのビット列）とそれを解釈するための「表現情報」を合わせて「内容情報」と定義している。そして、コンテンツの由来を示す来歴、他の情報との関連を示す文脈、コンテンツを識別するための ID 等の参照、内容情報が変更されていないことを示す不変性を記述した「保存記述情報」（PDI:Preservation

6 ビット列をそのまま別の媒体に移行（コピー）することもマイグレーションの一種であるが、本報告書でマイグレーションという場合、ファイルフォーマットの変換を伴うものを指すこととする。

Description Information）を、情報パッケージの構造を記述する「パッケージ情報」とまとめて情報パッケージとしてアーカイブに保存することとしている。パッケージのコンテンツを解釈するための記述情報（図書であれば、タイトルや著者等の書誌情報に相当するメタデータ。記述メタデータとも言う。）は、情報パッケージの外で管理するとしている。情報パッケージの種類としては、保管用情報パッケージ(AIP: Archival Information Package)のほかに、提供者から情報の提供を受ける際の提出用情報パッケージ(SIP: Submission Information Package)や利用者に情報提供を行う際の配布用情報パッケージ(DIP: Dissemination Information Package)なども定義されている。

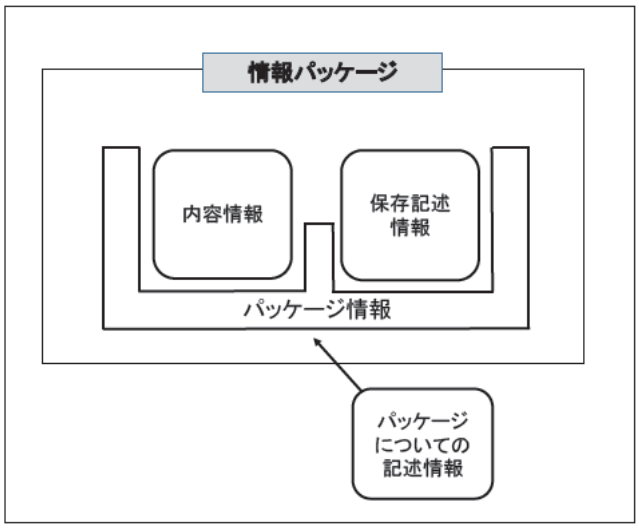

図 1 情報パッケージの概要図 [41]図 3 より

OAIS では、アーカイブが備えるべき機能として、6 つのエンティティを定義している。受入（Ingest)、保管（Archival Storage)、データ管理（Data Management)、保存計画（Preservation Planning)、運用統括（Administration)、アクセス（Access）の 6 つである。OAIS ではこれらの機能が詳細に規定されている。

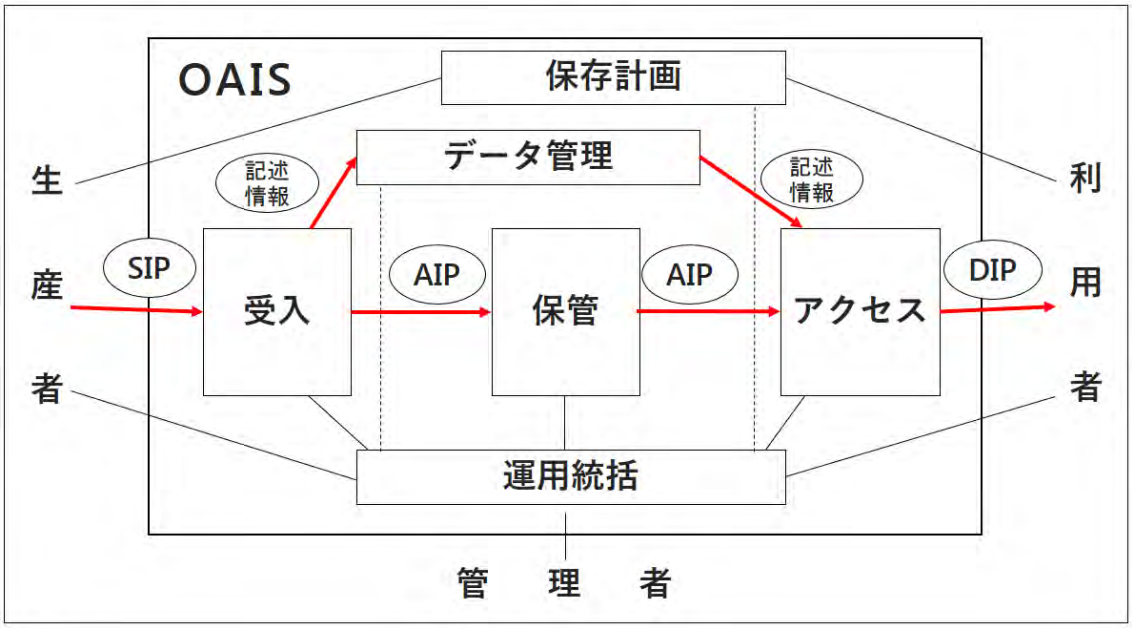

図 2 OAIS のエンティティ概要図　[28]Figure4-1 より

(4) 電子情報の長期利用を保証するための長期保存知識ベースと分散保存

OAIS では、保存されたコンテンツを解釈し、利用可能とするための長期保存知識ベースを持つこととしている。例えば、JPEG2000 というファイルフォーマットの画像を閲覧するためには、JPEG2000 画像を表示するツールやその使用方法が必要である。また、長期的にはツールそのものの保存も考慮する必要がある。JPEG2000 画像の表示ツールのサポートが無くなったときに備え、ツールのソースコード、表示ツールの作成当時に参照していた JPEG2000 の規格書も保存しておく必要があるだろう。

ゲームを長期利用するためには、ゲーム機の保存やゲーム機の使用方法の保存も必要になるが、機械はいつか故障するため、ゲーム機を再現するエミュレーションが必要になる場合がある。研究データもデータだけでは、その意味を理解できないため、データの解釈方法と同時に、そのデータを可視化するアプリケーション等も保存しておく必要があるだろう。ただし、再生アプリケーションを長期利用するためには、それを動かす環境（ハードウエア、OS その他のミドルウェア）も保存しておかなければならない。

ファイルフォーマットに関する知識、エミュレーションに関する技術は、分野を問わず共通するものが多い。そこで各機関が共同で電子情報の長期利用のための知識ベースを構築することが、各機関のアーカイブで保存しているコンテンツの利用保証につながると考えられる。欧米ではこの考えに基づいたフォーマットレジストリの構築について研究されている。フォーマットレジストリやマイグレーション、エミュレーションに関する知識やツールを共有する「長期保存知識ベース」を構築することで、分野を超えた長期保存の取り組みが可能となると考えられる。

デジタル資料の複製を分散して保存することで、一つのアーカイブシステムの運用停止により利用不可となった場合でも、複製を保存している別のアーカイブからコンテンツを利用できるようにすることが可能である。OAIS では、アーカイブ間の連携により、アーカイブシステム間で協力して保存を行う相互運用の方法や戦略についても記述されている。学術文献の一部ではすでにこの取組が行われている。

3　恒久的保存基盤

次世代システム開発研究室では、電子情報の長期利用を保証するためのシステムインフラ（基盤）として、複数の機関が共同して構築する「恒久的保存基盤」を構想した。これは、2章で説明したビット列の保存を確実に行ために、OAIS に記述されている分散保存を実現すること、論理的な保存を図るために OAIS に記述されている知識ベースを「長期保存データベース」として構築しようという構想である。

恒久的保存基盤は、多様な分野の機関が連携した分散型アーカイブにより構成される。恒久的保存基盤は、どこか単一の機関が構築運用するのではなく、各機関が連携したネットワークや分散型アーカイブで構成するものとし、分散型アーカイブは機関単位、又は分野単位で構築されるものとする。図 3 にその構成予想図を示す。

恒久的保存基盤が保存対象とする電子情報資源は、分野や種別を問わない。ただし、電子情報資源なら何でも収集するのではなく、既存の分野、例えば図書に相当する電子情報であれば図書館が、美術作品に相当する電子情報資源であれば美術館が、それぞれの収集方針、選定方針に従って収集するものとする。また、新たな分野の電子情報が現れた場合、その分野に係る機関が収集することとなる。図 3 では、多様な分野の機関が恒久的保存基盤に参加できるものとして、分野 A,B,C のように抽象化して示している。

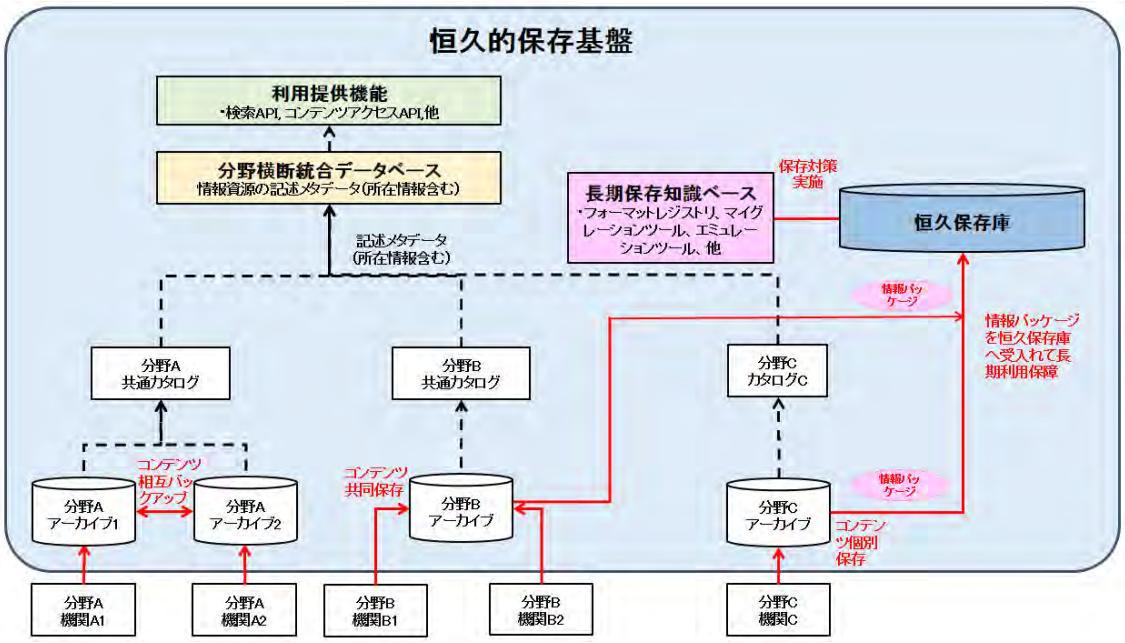

図 3 恒久的保存基盤の構成予想図

恒久的保存基盤では、以下の 4 つの機能の実現を目指す。

(1) 電子情報資源の記述情報と所在情報を集約する分野横断統合データベース

各機関で構築されている電子情報資源（コンテンツ）の記述情報（記述メタデータ）と所在情報を分野横断統合データベースに集約し、分野横断統合データベースを検索することで、国内のコンテンツへの一元的なアクセスを可能とする。図 3 で分野 A は機関毎にアーカイブが運用されている場合であるが、分野 A の記述メタデータを集約した共通カタログを構築し、そこから分野横断統合データベースにコンテンツの所在情報を含む記述メタデータを集約する。分野 B はコンテンツを共同保存するアーカイブが運用されている場合で、この場合は分野 B アーカイブのカタログがそのまま分野 B の共通カタログとなる。その共通カタログから分野 B のメタデータを分野横断統合データベースに集約する。分野 C は複数の機関が存在する大きな分野ではなく一つの機関がアーカイブを運用している場合で、分野 C のアーカイブのカタログを直接分野横断統合メタデータ・データベースに集約する。

各アーカイブから情報資源にアクセスするための記述メタデータを収集する仕組みについては、EU の Europeana[7]や国立国会図書館の国立国会図書館サーチなど、すでにいくつかの取り組みが行われている。

[7] http://www.europeana.eu/portal/en[8] スーパーコンピュータを有する、応用数学とコンピュータサイエンスを専門とする調査機関。

(2) 分散保存

図 3 では複数の機関でコンテンツを分散保存する例を示している。分野 A では、機関 1 がアーカイブ 1 を、機関 2 がアーカイブ 2 を持つが、アーカイブ 1 と 2 で相互にコンテンツのバックアップを取ることで確実な保存を行う。分野 B では機関 B1 と機関 B2 が共同でアーカイブを持つが、分野 B 内にはバックアップが無いため、図 2 右上の「恒久保存庫」にバックアップを保存することで確実な保存を行う。恒久保存庫は分野を問わず、コンテンツを恒久的に保存するためのストレージである。分野 C はアーカイブを運用する機関が一つだけなので相互バックアップができない。そこで恒久保存庫にバックアップを保存することを想定している。

上述のとおり、コンテンツのバックアップは、基本的には各機関間の分散型アーカイブでバックアップを取ることとするが、より確実な保存を実現するため恒久的に電子情報資源を保存するアーカイブである恒久保存庫を構築し、希望する機関からの要望により恒久保存庫にコンテンツを保存できるようにする。恒久保存庫は、標準化された形式の情報パッケージを、標準化された交換プロトコルで受入れ、コンテンツの長期保存と長期利用を保証する。そこで情報パッケージの形式と情報パッケージ受入のための交換プロトコルについて、海外の研究動向を調査した。

(3) 長期保存知識ベース

OAIS の保存計画では、情報資源の長期利用を保証するための計画を策定することを求めている。そのためには、ファイルフォーマットの旧式化検知や再生環境の入手可能性等、継続的なモニタリングが必要である。2(4)で述べたとおり、マイグレーションやエミュレーションには専門的な知識や技術が必要であり、個別の機関が単独で実施するのは困難である。

恒久的保存基盤では、2(4)で説明したフォーマットレジストリ、マイグレーションツール、エミュレーションツール等、保存計画の実行に必要な知識やツールを分野の壁を超えて集積する長期保存知識ベースを構築する。

長期保存知識ベースは、各機関から利用を可能とするが、恒久保存庫に保存されたコンテンツに対しては旧式化検知を行い、必要に応じてマイグレーションやエミュレーション等の長期保存対策を実施する機能を持たせる。

(4) 利用提供

恒久的保存基盤に保存された電子情報資源にアクセスするには、電子情報資源を検索し、コンテンツそのものにアクセスする機能が必要である。このとき検索やコンテンツアクセスのプロトコルが機関毎にバラバラに実装されていると、アクセス先の機関毎に検索機能・閲覧機能等を開発しなければならず、余分なコストが必要となる。検索やコンテンツアクセスのプロトコルと API を標準化し、どの機関の電子情報資源に対しても同じプロトコルと API

でアクセスできるようにすることで、利用提供機能の開発コストを抑制できる。コンテンツアクセスのプロトコルは、画像、音声など、電子情報資源のフォーマットごとに標準化する。また、権利情報等を踏まえたアクセス制御の機能も実装する必要がある。恒久的保存基盤では、これらのプロトコルと API を標準化し、分野横断統合データベースで提供する。

4　恒久的保存基盤の構築に向けた技術要素

恒久的保存基盤の構築に必要な技術要素をまとめると以下のとおりである。

(1) 電子情報資源の記述メタデータと所在情報の集約
　① 複数のアーカイブシステムからの記述メタデータと所在情報の収集

(2) 分散保存
　① OAIS に準拠したアーカイブシステム
　② 情報パッケージの標準化
　③ アーカイブ間でのコンテンツ交換プロトコル

(3) 長期保存知識ベース
　① 長期保存に必要な保存メタデータ
　② フォーマットレジストリ

(4) 利用提供
　① メタデータ利活用のための API の提供
　② コンテンツアクセスのプロトコルや API の標準化
　③ 権利情報を踏まえたアクセス制御機能

本報告書では、上記のうち、(1)については、国立国会図書館サーチによる実績があるため今回の調査対象から外し、(2)、(3)、(4)に関する海外の研究動向を調査した。

5　海外の研究動向

5.1 OAIS に準拠したアーカイブシステム

OAIS の概要を説明し、実際に OAIS を採用して構築された海外の事例から得られた知見や問題点をまとめる。

Zuse Institute Berlin(ZIB)[8]の Digital Preservation System(DPS)は、収集データの種別に依存しないシステム運用、既存基盤の垂直統合、保管用情報パッケージの相互運用を

8 スーパーコンピュータを有する、応用数学とコンピュータサイエンスを専門とする調査機関。

目指して構成されており、恒久的保存基盤の目指すシステム全体構成の参考となる。

フランス国立図書館(BnF)のリポジトリ(SPAR)での取り組みでは、OAIS における保存計画のシステムへの実装例と、システム化による運用における人的資源の効率化が示されている。

また、分散型アーカイブの OAIS への適用について、Outer OAIS-Inner OAIS(OO-IO:外面-内面 OAIS)モデルについて紹介する。

5.1.1 DPS (ZIB)

ZIB[22]では文化遺産と研究施設をサポートするアーカイブシステムのために、OAIS に示される機能モデルに沿うよう DPS を設計・構築している。

DPS に対する主な要求仕様は、DPS に納められる情報を、Preservation Metadata: Implementation Strategies (PREMIS)[9] [34]で記述される保存メタデータを利用して、他所を参照することなく単独で理解可能な AIP として維持することである。DPS では、必要な機能を一から開発することは現実的でなく、また既存の統合的な長期保存パッケージでは要求される仕様に合わないため、オープンソースである Archivematica[5]、iRODS[38]、Islandora[19]を組み合わせている。オープンソースのツールを利用することで、ツールごとの機能の明確な切り分け、各開発コミュニティによるドキュメントやベストプラクティスの活用ができ、その結果信頼のおけるシステムを構築できるという主張に基づいている。さらに、DPS へのデータ入力においては、全ての図書館、公文書館、美術館からのデータ提供に適合するように努力し、それぞれの組織に向けて Metadata Object Description Schema(MODS)[33]、Encoded Archival Description(EAD)[30]、Lightweight Information Describing Objects(LIDO)[16]に対応している。

DPS の構成では、通常の OAIS の機能エンティティの他に、「事前受入」という機能エンティティがある。これは、入力されるフォーマットに依存せずデータを保存するために、保存のレベルを「受動的な保存」と「能動的な保存」に分けている。受動的な保存に分類されるデータは、データ構造とメタデータに加えてバイナリデータそのものを保存する。一方、能動的な保存に分類されるデータに対しては、オープンな仕様のアーカイブフォーマットに適合するための様々な処理を施す。記述情報が Dublin Core[10]で記録されている場合は、CSV や XML で直接入力することができ、EAD、LIDO、MODS の場合には処理過程においてダブリンコアにマッピングされる。最終的に、全ての情報は同一の構造を持つ SIP に再構築され、受入を容易にしている。

将来的な改善として、新たなデータ提供者が DPS に実装されていないプロトコルを要求する可能性に備えて、SWORDv2[21]、OAI-PMH[25]、S3 クラウド[10]を調査する予定とのこと

[9] 正確には PREMIS Data Dictionary であるが、一般的に PREMIS と記述されるため本報告書でもそのように記述する。

[10] https://aws.amazon.com/jp/documentation/s3/

である。また、他のアーカイブとの連携については、(DIP ではなく)AIP のままお互いのアーカイブでの「受入」が可能であるか試験する予定である。長期保存用のリソースを持たないクライアントに対して、保存データの概要を理解できるように DIP を提供することも予定している。

5.1.2 SPAR (BnF)

SPAR(Scalable Preservation and Archiving Repository)[6]は、フランス国立図書館 (BnF) が 2004 年に検討を開始し、2010 年に稼働開始したアーカイブリポジトリである。稼働後、OAIS の機能エンティティ「受入」、「保管」、「管理」を担うモジュールを中心に開発を進めてきたが、2014 年に「保存計画」を担うモジュールの開発を決定した。

SPAR ではデジタルオブジェクトを 6 種類のトラック（デジタル文書と関連ファイル、音声画像、ARC や WARC ファイル[11]などのウェブ法定納本、電子書籍などの協定済み法定納本、行政記録、第三者機関のアーカイブ）に分けて入力している。また、SPAR におけるチャネル（サブトラック）には、提供者とアーカイブの間で交渉された SLA (Service Level Agreements) により決定される受入フォーマットや最大ファイルサイズなどが記述される。各チャネルは、システムを構成する XML ファイルとして記述される。

SPAR における各処理を可能な限り自動化するために、長期保存のポリシー（フォーマット、エージェント、オントロジー、分類システム、トラックとチャネル）の全ての要素を XML ファイルで記述した参照パッケージが使われる。参照パッケージへのリンク情報と識別子を利用することで、各パッケージ共通情報の参照表現が可能になり、コンテンツそのものだけではなく、OAIS における表現情報と保存記述情報の長期保存における簡素化が図られている。OAIS が規定する全ての要素を自動化することは困難であるため、開発者、長期保存リスク管理者、収集責任者などの人的資源が依然として必要とされる。これは図書館の長年の課題であり現在も解決に取り組んでいる。

SPAR の新しい保存計画モジュールは参照パッケージの策定について、システム管理者や長期保存のエキスパートの共同作業を促進するように設計されている。そのため、図書館全体で共通する知識が取り込まれ、作業が加速し、信頼性が増すと共に、長期保存についての活動がより可視化されるなどのメリットが期待される。さらに、保存計画に関連する全ての決定プロセスが、参照パッケージにおいて文書化されるようになった。

保存計画モジュールが取り込まれた新しい SPAR では、参照パッケージの策定が所定の手順により効果的に検討されるようになった。実際に、文化遺産のデジタル化における TIFF フォーマットから JPEG2000 へのスムーズな移行や、ボーンデジタルのオフィス文書の保存形式に PDF を追加することなどに利用された。

[11] http://warp.da.ndl.go.jp/contents/reccommend/mechanism/mechanism_warc.html

5.1.3 OO-IO モデル

OAIS に基づくアーカイブシステムが増加し、さらなる持続可能性への要求が高まれば、必然的に分散型の長期保存システムが必要とされる。外面-内面 OAIS (OO-IO) モデル[12]は、分散長期保存を可能とする、複数の OAIS 間の共同的相互作用の分析と各 OAIS が担う機能の監査を可能にする。Institution Repository-Bit repository(IR-BR)モデル[13]を前身とする OO-IO モデルでは、まずモデル全体を外面 OAIS で置き換え、さらに外面 OAIS の機能要素のうち、OO-受入、OO-データ管理、OO-保管の各要素について、それらを内面 OAIS でモデル化する。

OO-IO モデルの主要な目的は、複数の組織が参加する分散型アーカイブの活動における組織的（何をするべきか）・技術的（どのようにすればできるか）な課題を単純化することである。単純な OAIS モデルよりも OO-IO モデルが要求されるのは、保管コンポーネントが分離され、複数の OAIS に対応する外部組織によって利用される場合などである。

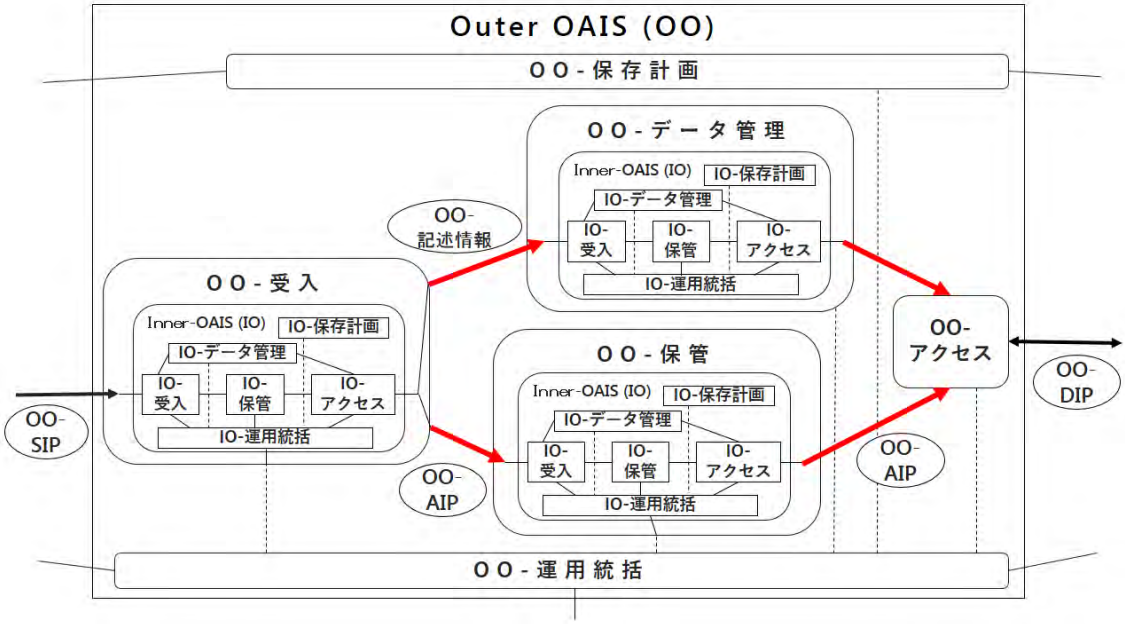

図 4 OO-IO モデル　[12]Figure2 より

OO-保管コンポーネントは、OO-受入コンポーネントからの OO-AIP を受信する。内部的には、OO-AIP が IO-SIP として扱われることを意味する。その後、OO-AIP として OO-アクセスに提供される。続いて、IO-受入から OO-受入コンポーネントに向けて、受入と保管の確認が通知される。

OO-データ管理コンポーネントは、OO-受入コンポーネントが OO-AIP から抽出した OO-記述情報を入力とし、OO-アクセスに処理結果を出力する。OO-データ管理コンポーネントを用いることが合理的である二つの事例が考えられる。一つは、データ管理において非同期のアップデートが生じる時に、一部のデータを分散型のアーカイブシステムで保護するケ

ース。もう一つは、ある電子情報資源に関わる記述メタデータ、フォーマット、環境などの情報が、それぞれ独立に分散した機関で記録されているケースである。

OO-受入コンポーネントは、OO-SIP を入力として、OO-保管コンポーネントに OO-AIP を出力する。同時に、OO-AIP から抽出する OO-記述情報を OO-データ管理に配信する。OO-受入コンポーネントは、OO-保管コンポーネントと OO-データ管理コンポーネントの双方に情報を配信するため、OO-保管コンポーネントよりも内部構造が複雑になる傾向がある。内部処理には、受入データの品質検査、OO-AIP 及び OO-記述情報生成処理が含まれる。OO-受入コンポーネントが必要な事例として、UC3-Merritt[7]や Archivematica のようなマイクロサービスソリューションのための分散された受入や、受入処理の遅延に起因するデータ損失のリスクを低減するための OO-SIP の保存が挙げられる。

分散型アーカイブに対して OO-IO モデルを適用することで、いくつかの有意な結果が得られている。まず、OO-IO モデルでは、Outer と Inner を接頭辞として付けるため、OAIS の機能を明示的に表すことができる。些細な工夫ではあるが、デンマークの事例では特にコミュニケーションの質を向上させ、議論における誤解を避けることに成功している。次に、OO-IO モデルは、Outer と Inner のインターフェースを分析する基盤を提供している。これは、分散アーカイブを検討する上で欠かせない相互運用性についての理解を促進する。最後に、OO-IO モデルは、分散型アーカイブの監査、すなわちアーカイブに要求される機能の検査を実現可能にする。分散型アーカイブ環境における監査の課題は、責任の分担と、複数の OAIS システムを横断し、ISO16363(Audit and certification of trustworthy digital repositories)[27]に定義された、長期保存の要求仕様を満たすことを累積的に実証するエビデンスを集めることである。OO-IO モデルを用いることで、各 OAIS システムの担う機能とその入出力が明確となり、個々の監査を統合することで全体の監査が可能になる。

5.2 情報パッケージの標準化

データへのアクセス性と再利用性を保証するため、長期保存に必要なメタデータの概要を説明する。

5.2.1 長期保存に必要な保存メタデータ

保存期間の長さや組織の変化、文化的又は財務的な優先度が、デジタルコンテンツの継続的なアクセス保証と長期保存に対するリスクを増大させている。デジタルコンテンツを失うリスクを避けるための全てのアーカイブの機能に、保存メタデータが影響を与える。

DIGITAL PRESERVATION METADATA STANDARDS[1]では、デジタルデータの保存に必要とされるメタデータを基本的な保存機能に基づく 4 つのカテゴリに分類している。

① 記述情報メタデータ
② 構造表記メタデータ
③ 物理ファイルの技術的な情報を表すメタデータ
④ 管理情報を表すメタデータ

まず、記述情報メタデータには、著者やタイトルなどのデジタルコンテンツに関わる情報や、来歴情報（スキャンデータであるならば、元になる印刷物の情報）が含まれる。次に、構造表記メタデータは、各データの物理的及び論理的な構造上の関係を表し、ウェブサイトとそれに含まれる画像との対応関係や、電子化された書籍のうち連続するページの対応関係などが記述される。続いて、物理ファイルの技術的な情報を表すメタデータは、全てのコンテンツに共通するソフトウェア及びハードウェアに関する情報（画像ならば水平垂直サイズ、音声ならば再生時間等）や、コンテンツに依存する情報を含む。最後に、管理情報を表すメタデータには、コンテンツの管理者、適用された長期保存の計画を示す情報、権利情報とコンテンツへのアクセス許可情報等が含まれる。この中でも、最後に示した管理情報を表すメタデータは、保存メタデータとして広く知られている。

長期保存に特化したメタデータ仕様の例として、PREMIS と Long-term preservation Metadata for Electronic Resources(LMER)[29]がある。

PREMIS は、長期保存の中心的な機能を提供するために必要な意味単位を規定することを目標としており、特定の技術、構造、コンテンツ種別、若しくは長期保存の戦略を前提としない。そのため「技術的に中立」なデータモデルであると言える。また、メタデータはローカル（アーカイブ内）に保存してもよいし、外部レジストリに保存してもよく、保存場所の自由度が高い。これは導入可能なアーキテクチャが幅広くサポートされ、アーカイブ内部の実装が柔軟になることを意味する。その一方で、アーカイブ間の横断的な情報交換には、相互運用性を改善する目的からより強い制約が課せられる。PREMIS のデータモデルでは、5 つの要素（所蔵対象となるコンテンツの集まりを表す知的エンティティ、長期保存の対象となる情報を表すオブジェクト、人・組織・プログラムなどを表すエージェント、オブジェクトやエージェントに付随する権利、データリポジトリにおける変化を表し、少なくとも一つのオブジェクト又はエージェントへの関与・作用を記述するイベント）が相互に結びついている。

LMER は、PREMIS を代替する解決手段としてドイツ国立図書館が設計したメタデータフォーマットである。PREMIS と同様に、チェックサムやフォーマットのような基本的な技術メタデータを含み、コンテンツ種別に固有のメタデータは、Metadata for Image in XML Schema(MIX)[32]や Technical Metadata for Text(textMD)[35]などの追加スキーマを使用して埋め込み可能である。オブジェクトの修正や変更について、LMER は、その手順などを順次オブジェクトに付随するメタデータとして記述する、プロセスアプローチ方式で記録する。一方 PREMIS は、変更・修正そのものを個別のイベントとして記述して、それぞれをオブジェクトとリンクさせるという、イベントアプローチ方式で記述する。

デジタルオブジェクトに対して長期保存に関する行動が実施される場合、新たな環境で表示や再生をするために、通常は新しいデジタルオブジェクトが生成される。この時、データフォーマットの変換が伴うとデジタルオブジェクトのいくつかの特性が失われたり、

変更されたりするリスクがある。これに対して、欠損や変更が許容できない重要な特性については、保存メタデータにおいて明示的に取り扱うべきである。

長期保存の対象となる情報は、デジタルオブジェクトに関する様々なメタデータで記述される。それらを集約して一つの情報とするために、メタデータコンテナが使われる。メタデータコンテナには、人間が読みやすいだけでなく機械による処理が容易な XML 形式を採用することが望ましい。そのようなコンテナの事例として、Metadata Encoding and Transmission Standard(METS)[31] や MPEG-21 Digital Item Declaration Language(DIDL)[18]がある。

画像、音声、映像、テキストなどのコンテンツタイプに固有な技術的なメタデータ（オーディオデータであればサンプリングレート、画像であれば高さや幅、色深度などの情報）は、メタデータコンテナの拡張スキーマを利用して導入される。これらの情報を、自己記述的にファイルフォーマットの中で扱うこともできるが、明示的に分離して扱うことにより、効率的な処理や技術的なメタデータの個別配信、異なるアクセス権やライセンスの付与が可能になる。コンテンツタイプ固有の技術的なメタデータの例として、ANSI/NISO Z39.87(Technical Metadata for Digital Still Images)[4]と textMD がある。ANSI/NISO Z39.87 は、ビットマップ形式[12]のデジタル画像を記述するための意味単位を規定しており、米国議会図書館が管理する XML スキーマである、MIX としてメタデータをエクスポートできる。一方、textMD は XML スキーマで表される、テキストベースのデジタルオブジェクトについての技術的なメタデータ仕様である。

デジタルコンテンツの長期保存においては、特定の連携機関が保持する外部のリポジトリとのコンテンツ共有が望まれる可能性がある。これは分散的な長期保存の解決策として位置づけることが可能だが、実際には異種混合の長期保存システム間で複合的なデジタルオブジェクトを交換しなければならない。複数のリポジトリ間で正しくデータ交換するためには、互いのデータモデルを正確に理解しなければならないが、リポジトリ間のデータ転送においては、デファクト標準である METS と PREMIS が十分な柔軟性を有することが後述の TIPR[5.3.1 参照]の活動などで証明されている。

現在、電子情報資源の長期保存に寄与する幾つかのメタデータ仕様は、各国の政府機関あるいは国際標準化団体により策定されているが、保存メタデータ基準を確定させるには未だ限られた実績しかない。実用での知見と新たな基礎研究の組み合わせが、長期保存メタデータのさらなる理解に寄与するのである。

5.3 アーカイブ間でのコンテンツ交換プロトコル

運用が困難になったアーカイブが持つコンテンツの他機関への引継ぎや、災害からのデータ復旧といったイベント発生時に、アーカイブ間での連携やバックアップのためのコンテンツ交換が必要となる。ここでは、コンテンツ交換用に OAIS の情報パッケージの標準化

[12] 色のついた小さな点（ドット）を集めて画像を表現する形式。

を目指したTIPRの取り組みを紹介する。また、連携機関でバックアップを持ち合い、破損データを修復するLOCKSSプロジェクトについても紹介する。

5.3.1 TIPR

アーカイブシステムには、データ保存の優先順位の変更や資金不足などの理由から、既存データの維持や新たなデータの受け入れができなくなるリスクがある。そのためアーカイブシステムは、そのような状況においてもデータを失わないように、アーカイブシステム内のデータを外部機関に移管できなければならならない。これを達成するためには、AIPを交換あるいは転送する必要性がある。

フロリダ図書館自動化センターを中心とした Towards Interoperable Preservation Repositories(TIPR)[20]は、ミュージアム・図書館サービス協会(IMLS: Institute of Museum and Library Services) の資金提供を受け、リポジトリ間の AIP 交換のためのパッケージングフォーマットを作成し、検証した米国でのプロジェクトである。この取り組みの成果は、Repository eXchange Package (RXP)の開発である。軽量なパッケージングである RXP は、RXP 構造をエンコードするために METS を利用し、デジタルオブジェクトの来歴情報を記録するために PREMIS を用いる。RXP は、デジタルオブジェクトの来歴情報の連鎖を破壊することなく、多数の AIP 構造に順応し、また異なる長期保存環境においても簡単に生成と受入が可能である。

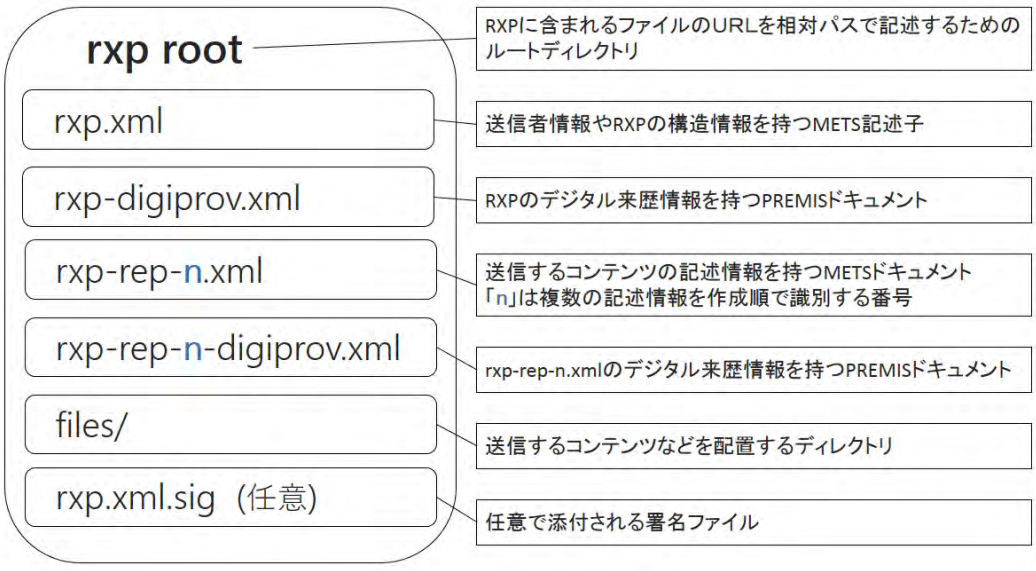

図4 RXP の構造 [20]Figure1 より

AIP の交換において RXP の利用が有効なユースケースとして、アーカイブの継承、災害復旧、ソフトウェアマイグレーション、リスク分散、他機関によるコンテンツ処理がある。

RXP の仕様は完成されているわけではなく、いくつかの点を改良しなければならない。まず、RXP を AIP そのものとして利用できるよう、記述メタデータ(DMD: descriptive metadata)を RXP で扱えるようにする点である。この点はコーネル大学、ニューヨーク大学などの TIPR パートナーらによって検討されている。次に、RXP には転送のためのパッケージフォーマットが規定されているが、それに加えて転送方法と保管についても考慮するべきである。それらが RXP 転送の前にリポジトリ間で合意できればデータ交換がより確実に実行される。合意内容には最低限、RXP 作成の詳細、リポジトリ間の転送計画、RXP 受入後に実行するタスクと RXP の扱い、権利と許諾、資金調達、法的処置を含み、それらは文書化されるべきである。より具体的に説明すると、RXP の作成では、RXP 交換におけるスキーマやパラメータを制御でき、また圧縮の有無や転送成功の条件などを決定するべきである。RXP 受入後には、メタデータのマッピングやフォーマットの変換などが必要であり、また受け入れた RXP の削除の可否なども事前に合意しておくべきである。さらに、パッケージ特有の権利や許諾の確認や、リポジトリの運営、維持に関わるコスト負担、あるいは著作権侵害やリポジトリ交換に関する様々な問題が生じた時の法的処置についても合意がなされるべきである。

5.3.2 LOCKSS 拡張

Lots of Copies Keep Stuff Safe(LOCKSS)プロジェクト[40]は、1998 年にスタンフォード大学においてプロジェクトが開始され，基盤となる LOCKSS ボックスと呼ばれるソフトウェアの開発が進められた。名称のとおり、各機関(ピア)がお互いにローカル・コピーを持ち合うことで、安定的な利用と保存を目指すシステムである。LOCKSS システムは、ピアが構成するネットワークを用いて、電子ジャーナルなどの収集、保存、相互チェック、発信を行う。各ピアの LOCKSS ボックスは、ポーリング機能を用いて異なるピアが持つ同一アーカイブ単位(AU: Archival Unit)でのコンテンツの相違を検出し、損傷している場合はそれらを回復できる。このような特徴を利用する図書館の LOCKSS ボックスは、5 つの機能（ロケーション、検証、承認、検出、保存）を持つ Peer-to-Peer(P2P)ネットワークを形成する。

2012 年にアンドリュー・W・メロン財団からの助成金を獲得し、コンテンツの収集、発信、保存の 3 つの領域で LOCKSS システムの拡張が行われた[9]。拡張前のコンテンツの収集では、初期の静的な Web コンテンツを前提としており、単純なクローリングには限界が生じている。そこで、Web フォームと AJAX(Asynchronous JavaScript And XML)[13]サポートの 2 点において LOCKSS システムの拡張が検討された。Web フォームについては LOCKSS デーモンに実装されたが、AJAX 対応はライセンスとセキュリティ、コスト上の理由から実装されなかった。

コンテンツの発信では、RFC 7089 で規定される Memento[14]を導入し、オリジナルのコンテンツを表す URI と、そのコンテンツのアーカイブを表す URI について、それぞれの URI が

[13] http://www.w3.org/TR/XMLHttpRequest/
[14] http://mementoweb.org/about/

表す情報が時間と共に同一性を失うという課題に対処した。またコンテンツへのアクセスについて、シボレス認証基盤[15]を用いたアクセス制御実験が行われた。

コンテンツの保存では、LOCKSS ボックスの機能における非効率性を解消する 3 つの処理（フィルタリング、対称的ポーリング、所有証明）が追加された。

フィルタリングでは、コンテンツに適用される出版元やダウンロード日時の情報など、動的に生成されるデータの選別を行うことで、機能的な同一性を担保し、不要なデータのコピーを抑制した。

対称的ポーリングでは、ポーリングを呼びかける poller 側が持つコンテンツ破損の情報を、投票に参加する voter 側でも把握できるように工夫することで、従来のポーリング機能にあった情報の非対称性を解消した。これにより、LOCKSS ネットワークにおいて回復プロセスの稼働する期間が大幅に減少することが確認された。

所有証明では、従来のデータ取得可能性の証明（PoR: Proof of Retrievability）に加えて、データの所有証明（PoP: Proof of Possession）の考え方を導入した。LOCKSS ボックスの機能のうち、検出と保存のみに PoR を適用し、PoR よりも処理コストが低い PoP をロケーション、検証、承認に適用することで全体的なリソースコストを削減した。

今回の拡張では実装には至らなかった AJAX 対応とシボレス認証についても、基本的な技術の検証は実施され、今後のユーザーニーズの対応に有益な試行であった。

5.4 フォーマットレジストリ

長期保存知識ベースに必要とされるフォーマットレジストリについて、既存の 2 つのリポジトリから、新たなフォーマットレジストリを構築した UDFR プロジェクトについての文献を紹介する。また、エミュレーションの技術として、コンテンツを複数の機関が保持する場合にも、コンテンツの提供には共通のエミュレーションサービスを利用する EaaS を紹介する。

5.4.1 UDFR

統合デジタルフォーマットレジストリ（UDFR: Unified Digital Format Registry）[37]は、新たにオープンソースで実装され、フォーマットを表現する重要な特性の収集・管理及び保存・照会に対応するプラットフォームである。UDFR は、米国議会図書館から全米デジタル情報基盤整備の保存プログラム（NDIIPP[16]）の一部として資金提供を受け、カリフォルニア大学カリフォルニアデジタルライブラリー（CDL）における、カリフォルニア大学キュレーションセンター(UC3[17]) で開発された。UDFR 以前のプロジェクトとして、2002 年に開発された英国国立公文書館の PRONOM[18]と、2006 年に開発されたハーバード大学の GDFR[19]

[15] http://www.internet2.edu/products-services/trust-identity/shibboleth/
[16] http://www.digitalpreservation.gov/
[17] http://www.cdlib.org/uc3
[18] http://www.nationalarchives.gov.uk/PRONOM

（Global Digital Format Registry）がある。UDFR はこれら 2 つのレジストリソリューション機能と保有データを基に統合され開発された。UDFR の基になった 2 つのサービスは、古いリレーショナル及び XML データベース技術に依存しているが、UDFR では全ての情報が RDF 形式で表現され、Linked Data として公開されている。

UDFR 構築プロジェクトの要件には、オープンアクセス可能な仕様書の作成、ウェブベースのインターフェース、UDFR 識別子の一意性を保証するための機構などが含まれた。最終的な要件には明示されていないが、複数機関の協力による分散型ネットワークという当初の構想は、今後の発展から排除されるものではない。ソースコードは GPL[20]で公開され、PRONOM からエクスポートされたデータは、OGL ライセンス[21]下で利用可能と定められた。また、外部協力者が提供する情報は、クリエイティブ・コモンズ帰属（CC-BY）のライセンスで利用可能とされた。

UDFR は、PRONOM データモデルと GDFR データモデルを結合したモデルを意図していた。英国国立公文書館が実施していた PRONOM の Linked Data 版の作成作業が UDFR におけるオントロジーの作成作業と並行していたため、UDFR のモデリングはその全てを PRONOM のデータモデルに揃えることができなかった。実際、PRONOM オントロジーは、20 のクラスと 13 のプロパティを規定しているが、UDFR オントロジーは、113 のクラスと 159 のプロパティを規定している。ただし、UDFR オントロジーを策定する過程では、可能な限り PRONOM との一貫性を維持する努力がなされている。相互運用を容易にする目的から、UDFR オントロジーのプロパティはサブタイプ（rdfs:subProperty）を介して Dublin Core(DC)や Friend of a Friend(FOAF)[11]などの既存のオントロジーと関連付けられている。

UDFR には、2 つのソースからデータが入力された。Appspot[22]からエクスポートされた MIME データと PRONOM のデータが登録された。MIME データには、アプリケーション、音声、画像、テキスト、映像など様々な種類が存在し、1127 種類が対象とされた。一方、PRONOM にあるフォーマットデータは、2012 年 2 月にエクスポートされた情報に基づき、ファイル形式、文字エンコーディング、圧縮アルゴリズムなどのオントロジークラスについて 5985 種類が対象とされた。どちらの情報も UDFR へ入力され、42617 個におよぶ RDF トリプルとして登録された。

プロジェクトチームは、情報公開の一環として技術・運用に関わる数多くのプレゼンテーションを実施した。さらに、ワークショップでの議論を通して、PRONOM のデータにある外部シグネチャ（拡張子など）の重複を減らすために、UDFR データモデルの修正案が議論され、システムの改善へ繋がった。ユーザーコミュニティからは、継続的な活動テーマとして、レプリケーション、データソースの追加、オープンレビュー、恒久的な運用拠点とガバナンスの維持などが上げられている。

---

[19] http://gdfr.info/
[20] http://www.ipa.go.jp/osc/license1.html
[21] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[22] http://mediatypes.appspot.com/

5.4.2 EaaS

エミュレーションは、複雑な電子情報資源を保存するためのツールとして近年進化を遂げているが、多くの場合に、それぞれの目的に応じて調整して実装されなければならない問題がある。サービスとしてのエミュレーション(EaaS: Emulation as a Service)[26]は、デジタルオブジェクトの所有者や保存機関など、非技術系ユーザにエミュレーレションサービスを提供することを目的とした概念モデルである。既存の取り組みである KEEP プロジェクト[23]や Olive プラットフォーム[24]の考えを発展させたものである。

EaaS は、あるデジタルオブジェクトを再現するために必要なタスクを分割した分散アーキテクチャとして構築される。全てのコンポーネントを適切に分割することで、各コンポーネントをそれぞれの専門家が個別に管理できるようになる。EaaS モデルは、エミュレーションタスクを扱うサービスと、デジタルオブジェクトを提供するアーカイブの 2 つに分けられる。OS やソフトウェアなどの共通オブジェクトは、共有ストレージに集約して維持費を削減し、保存機関にアーカイブするデジタルオブジェクトは共有せずに、各施設で個別に管理する。

エミュレータ固有の設定や実装に依存しないことを担保し、コンピュータシステムの抽象的かつ包括的な記述を実現するために、EaaS はエミュレーション環境メタデータを導入する。エミュレーション環境メタデータは、技術的なメタデータであり、これを利用するエミュレーションコンポーネントが元の環境を再現できるようにコンピュータ環境を記述する。記述対象には、エミュレーションされるハードウェアアーキテクチャ（プラットフォーム）と、プラットフォームに接続される任意のデバイス（ディスクドライバー、サウンドカード、入力デバイスなど）が含まれる。エミュレーション環境メタデータは、EaaS の各コンポーネントを調整するための基礎的要素である。

EaaS は、接続した外部アーカイブのオブジェクトを安全にエミュレートし、不要なデータ転送と必要なディスクスペースの増加を抑えた効率的な環境管理が可能である。

5.5 メタデータ利活用のための API の提供

分野横断統合データベースに収集された記述メタデータの利活用のために有用と考えられる技術要素として、永続的識別子を用いた相互リンクの実現に向けた取り組み及び OAI-PMH の後継プロトコルとされる ResourceSync について調査した。

5.5.1 相互リンクの実現

ウェブベースの電子情報間の相互関係を確立・維持・支持することの難しさは、1990 年代におけるデジタル図書館戦略以来、継続して議論されている。多くの団体が DOI (Digital

[23] http://cordis.europa.eu/project/rcn/89496_en.html
[24] https://olivearchive.org/

Object Identifier）[3]を用いて、多様な科学的リソースの引用とリンクを増やすことに取り組んできた。さらに、データの識別と引用を容易にすること、また標準化を通して永続的識別子を割り当てるサポートを目的とした「DataCite[25]」などの団体も存在する。しかしながら、文献の引用などは、本質的に外部のリソースに向けた一方向性の関係を表すだけであり、学術研究の多くが入出力関係を伴うバリューチェーンの中に存在することが考慮されていない。この点において、学術リソースが相互にリンクし、コミュニケーションと検索サービスを統合するより優れた仕組みが必要とされる[23]。

電子情報の相互リンクを確立するには多方面に課題があり、その中でも、リンク対象となるオブジェクトの定義と目的、リンク関係が表す意味、相互リンクを実現する技術、ステークホルダとの関係性の構築、拡張性など、これらについて具体的に議論されている。

電子情報の相互リンクの課題に対し、2016 年 1 月 5 日にワシントン DC で開催された「Data & Publication Linking（データと出版物のリンク）」ワークショップにおいて、複数の提言が出されている。

第一の提言は、既存の活動の継続的な進展をサポートすることである。COPDESS[26]などの連合は、出版物とデータ間の関係がどのように統合され、交換されるべきかを調査するための基盤である。同様に、DataCite は広域なコミュニティを形成してきており、一連のツール、メタデータ構造、パートナーシップを有している。

第二の提言は、新規と既存の学術的なコミュニケーション事業において相互リンクを促進することである。データリポジトリ、出版社、機関リポジトリは、リソースに対する機械的なアクセスを可能にするべきである。NISO/OAI ResourceSync[5.5.2 参照]のようなプロトコルは、プログラムでアクセスできるようにリソースを公開する頑健なメカニズムを提供している。また相互リンクの特定と把握については、Open Annotation[27]や OAI-ORE[24]を用いたアノテーションとキュレーション、あるいは REST/HATEOAS アプローチ[15]を介した HTTP リンク情報を利用する手法も考えられる。

第三の提言は、たとえ最適なアプローチが定まっていなくとも、リンクを見つけその関係性を把握する必要性を強調することである。リンクと関係性を作成する試みが、標準化もしくは長期的な解決法に至らないとしても、ツールは構築・検証され、またその取り組みから学んだ事柄は、幅広いコミュニティが次のステップへ進むための情報となるはずである。リンクを見つけ集約することで、異なる相互リンクのアプローチについて、それらの優劣を把握することが可能になる。

5.5.2 ResourceSync

ResourceSync[42]は、2 つのウェブサーバの間でコンテンツの同期を行うためのプロトコ

[25] https://www.datacite.org/
[26] http://www.copdess.org/
[27] http://www.openannotation.org/

ルとその関連規格の総称であり、OAI-PMH の後継規格として 2011 年に検討が始まった。一連の規格のなかでコアとなる ResourceSync Framework Specification (ANSI/NISOZ39.99-2017)[28]のバージョン 1.1 が 2017 年 2 月にリリースされている。

図書館界でメタデータ収集のデファクトスタンダードとして成功している OAI-PMH には限界がいくつか存在する。1 点目は収集の対象がメタデータのみであること。2 点目は図書館界以外では普及していないこと。3 点目は更新と収集のタイミングにタイムラグがあることである。

このような OAI-PMH の限界を乗り越えるべく、ResourceSync は次の特徴を持つものとして開発が始められた。

- 同期の対象は URI を持つウェブ上のあらゆるリソース
- 仕様のベースは検索エンジンにクロール対象を知らせる標準的な仕様である Sitemap プロトコル
- コアの仕様は OAI-PMH と同様プル型であるが、コンテンツの生成などを通知するプッシュ型の仕組みを定めている
- 大規模環境や高頻度で更新を行うコンテンツを想定し、スケーラビリティに配慮した設計を目指している

同期のプロセスは、Sitemap ファイルをベースとした XML ファイルにリソースのリストや更新差分を記述することで、更新の有無が判断できるようになっている。更新が必要であれば、XML ファイルの URI を基に更新を行う。

現状 ResourceSync を実運用として実装したサービスは見られず、今後実際に広く活用されていくかは未知数であるが、いくつかの実験は行われてきている。

5.6 コンテンツアクセスのプロトコルや API の標準化

コンテンツアクセスの標準化に向けての動きとして、デジタル画像へ Web 上でアクセスする際の国際的な枠組みとして利用が広がっている、International Image Interoperability Framework(IIIF)について概要と機能をまとめる。

また、ボーンデジタルコンテンツへのアクセス機能について、検索に有用な共通 DIP フォーマットの策定と、それを用いたアクセスツールを検証した E-ARK プロジェクトの取り組みをまとめる。

5.6.1 IIIF

画像ベースのリソースへのアクセスは文化遺産の調査、教育、継承に必須である。ウェブでやり取りされる画像、本、新聞、地図の大半はデジタル画像である。しかし、インターネット上の画像リソースは機関ごとに配置されており、それぞれ固有のアプリケーションでのアクセスが必要となっている。IIIF[39]は、そのような状況の中で、相互運用可能

[28] http://www.openarchives.org/rs/1.1/resourcesync

な技術とコミュニケーションフレームワークを目指して策定された一連の API の総称である。IIIF を導入すると、共通化された国際的なプロトコルにより、デジタルコンテンツとそれを閲覧するためのビューワを分離でき、デジタルアーカイブを統一的に扱えるようになる。

IIIF は主に 2 つの API からなる画像表示規格である。1 つは ImageAPI という拡大縮小やトリミング、傾きといった画像表示のパラメータを URI で指定するための規格。もう一つは PresentationAPI という ImageAPI をベースにしつつ、メタデータやアノテーションなどを JSON 形式のマニフェストとしてまとめるための規格である。その他に閲覧の用途には直接関係しないが、認証機能を持った AuthenticationAPI と、アノテーションを検索する SearchAPI がある。

IIIF は他機関との画像共有の規格としてすでに広がっており、Europeana でも参加機関の IIIF 対応のために専門のタスクフォースが立ち上がるなどの動きがある[29]。今後、音声や映像など画像以外の共有にも IIIF の用途が広がる可能性があり、動向を注視する必要がある。

5.6.2 E-ARK

E-ARK(European Archival Records and Knowledge Preservation)[2]は、2014 年 2 月 1 日から 2017 年 1 月 31 日まで実施された、3 年間の国際研究プロジェクトである。

長期保存の最終目標は、データへのアクセス性と再利用性を保証することである。しかし、ボーンデジタルなアーカイブ資料へのアクセスを提供する場合の知見は、まだ限られている。さらに、アクセスに対するユーザーニーズと既存サービスとの間に大きなギャップがあることも確認されている。特にユーザーニーズに関する 4 つの項目が欠落している。

- アクセスツールにおいて、ユーザのデータ利用を支援する機能の欠如
- 検索における包括的かつ定性的なメタデータ[30]の欠如（結果、興味のある情報の発見を困難にしている）
- 柔軟で現代的なアクセスサービスの欠如
- アクセスコンポーネント間の相互運用性の欠如

E-ARK プロジェクトはこれらの課題を解決し、デジタルアーカイブ資料へのアクセス性を改善するソリューションの開発を目的としている。プロジェクトの中心は、標準化された共通の DIP フォーマットとそれを用いた共通のアクセスツールである。

共通 DIP フォーマットは、E-ARK 用プロファイルを持った METS ファイルによって記述され、メタデータは PREMIS と用途に合ったメタデータ標準を併用できる。DIP が持つメタデ

---

29 http://pro.europeana.eu/blogpost/iiif-adoption-by-europeana-future-perspectives-for-the-network-1

30 データベースそのものや、ジオデータ、電子情報資源管理システム(ERMS)データ、OLAP キューブなど、複合的なデータタイプを検索するためのメタデータを意味する。

ータの大部分はSIPとAIPから継承するが、DIP固有のメタデータとして、アクセス権、ユーザ制御、補足的な信頼性情報など複数のカテゴリに分類されるメタデータを内包している。

共通アクセスツールは、複雑化しているコンテンツタイプの課題に焦点を当てている。アクセスフローは、検索とリクエストの管理、DIP準備、DIP配信、DIP管理の4つのステップで構成され、各ステップはさらに細分化されている。共通アクセスツールは参照実装であり、OAIS 準拠のアーカイブを構築するために必要な全てのコンポーネントを提供するように開発されている。

広範囲におよぶ検証プロジェクトにおいて、E-ARKプロジェクトの成果が利害関係者や潜在的ユーザのニーズを満たすかについて検証された。7 つのパイロットサイトにおいて、E-ARK コンポーネントが試験的に実装され、2015年後半から2016年まで、E-ARKプロジェクトの約3分の1の期間を費やして検証された。パイロットプロジェクトについての最終報告[31]が2017年2月に公表され、E-ARK General Model[32]についてもウェブで公開された。

5.7 その他、電子情報の長期保存に関する技術や標準

上記で取り上げた技術のほかに、米国議会図書館によって資金援助された長期保存プロジェクトである Chronopolis と、ニュージーランド政府機関が構築したクラウドストレージを利用した長期保存システムについて、恒久的保存基盤に資すると考えられる点をまとめる。

5.7.1 Chronopolis

米国では1990年代後半から、科学分野、教育コミュニティ、及び連邦政府機関の知的資産などの電子情報資源について、資金援助不足や、ストレージシステム、アクセス機構、エンコーディング形式の技術進化に起因するデジタルデータへのアクセスが失われる事態が懸念されていた。このような流れの中で、米国議会図書館の全米デジタル情報基盤整備によって資金援助されたプロジェクトである Chronopolis[8]は、幅広い領域のアーカイブニーズを満たすマルチメンバーパートナーシップとして始まった。

Chronopolis モデルは、 最も価値の高いデジタルデータの可用性、アクセス性を確保するために必要となるデータ管理及び保存インフラストラクチャのモデルの提供を目指している。これは、Chronopolisの根底にあるコンセプトが、時間の経過とともに拡大、縮小し、拡張できる長期保存サイバーインフラストラクチャ開発の段階的なアプローチであることに基づいている。Chronopolisのアーキテクチャは、受入れたデジタルコレクションの長期保存だけでなく、Chronopolis 機能自体のライフサイクル管理（評価、加入、記述、配置、ストレージ、保存、アクセス）も提供するように設計されている。

---

31 http://www.eark-project.com/resources/project-deliverables/97-d25-1

32 http://www.eark-project.com/resources/general-model

Chronopolis はいくつかのコア技術で構成されており、それらは地理的に異なる場所にコピーされたコンテンツのシームレスな保存環境を提供する目的で連携するように設計されている。主な Chronopolis ツールは、大規模コレクションの転送においても自動的な整合性チェックと様々なファイルを単一パッケージに集約する機能を有する BagIt 転送フォーマット[17]、データグリッド内に格納されたデータコレクションへ統一アクセスを提供する Storage Resource Broker(SRB)、SRB 内でコレクションのコピーを監視するために開発されたレプリケーションモニタ、統合型監視プラットフォームである Auditing Control Environment(ACE)である。

Chronopolis は Chronopolis ツールを使用し、一連のデータコピーと保存サービスを提供する。受入とレプリケーションからモニタリングと管理まで、デジタルオブジェクトのライフサイクルを支える 5 つのサービスが存在する。コレクションの受入サービスは、Chronopolis への登録、品質保証/品質管理などの複数の段階で構成されている。レプリケーションサービスでは、SRB レプリケーションモニタが利用され、マスターサイトへのデータ転送に用いる読み取り専用アカウントの設定や、アクセス権の移譲が適切に実施された後、完全なコレクションのコピーがパートナーサイトに格納される。監査サービスでは、コピーファイルの整合性が定期的に監視される。デフォルトのポリシーでは、監視は 30 日ごとに実施される。メタデータサービスは、Chronopolis システムにおけるメタデータの作成と使用状況を分析し、新たなモデルを開発中である。最後にアクセスポータルサービスでは、疎結合された複数のソフトウェアシステムについて、理解しやすい単一のポータルを提供する。

5.7.2 クラウドコンピューティングの利用

国際的な流れとして、多くの公共サービス団体が、外部委託する情報技術の要件として、優れた経済性とより効率的で効果的な公共サービスの提供を可能にするクラウドコンピューティングの可能性を認識している。文化遺産機関も例外ではないが、その意思決定に資するアドバイスや経験が不足している。このような状況において、ニュージーランドの国立図書館は、国家デジタル遺産アーカイブ（NDHA: the National Library of New Zealand's National Digital Heritage Archive）[33]のデジタルコレクションの保管委託を 2013 年に決定した。結果として、NDHA は外部委託したストレージを有効活用でき、長期保存の対象となるファイル総数とデータ総量が飛躍的に増加し、長期保存に関わるコストの透明性も向上した。これは同様の決定を行う世界中の他の機関を支援するような経験的証拠になるものである[14]。

ストレージの外部委託に関する最初の報告書は、英国で公表されている。2014 年に英国

33
https://web.archive.org/web/20120523225615/http://www.natlib.govt.nz/about-us/current-initiatives/ndha/

国立公文書館は、外部委託に踏み出すことを検討しているアーカイブのために、4 件のケーススタディを記載したガイドライン[36]を発表した。まず注意する点として以下の 3 点が挙げられている。1 点目は、外部委託するサービスや技術のライフスパンは、アーカイブに期待される長期保存の期間より短いこと。2 点目は、データの損失といった重大なリスクに対するアプローチが、保存という目的とは相容れない、単なる金銭補償とは異なるものであること。3 点目は、外部委託をやめる際の出口戦略が策定されていることである。提示されている外部委託の利点の中で、デジタルアーカイブの観点で最も重要なのは、長期保存における性能改善の可能性である。これについては、複数拠点での自動化されたレプリケーションの実現可能性と、デジタルストレージの整合性チェックに関するベンダーの専門知識によって、デジタルデータの完全性を持続させるビットレベルでの長期保存において、性能の改善が可能と考えられる。

6　今後の課題

海外文献の調査の過程で、電子情報の長期利用に関する国内の有識者である筑波大学図書館情報メディア系杉本重雄教授、京都大学東南アジア地域研究研究所副所長原正一郎教授に意見聴取を行い、以下の意見をいただいた。

- どこに何が保存されているのか、所在情報の共有・把握自体が課題である。図 3 の分野横断統合データベースはそのためのものであるが、そこに記録する記述メタデータは時間的に変化していく。図書館資料の書誌の項目は安定しているが、研究データ等は変化が激しい。分野横断統合データベースの実現にはメタデータの変化に対応できるようにメタデータの来歴を管理する仕組みが必要である。
- 電子情報資源を恒久的に保存するには、図 3 の恒久保存庫の機能として、ビットレベルで長期間保存する高信頼性のリポジトリを考えることもできる。その場合、個々の情報資源の内容並びにその保存に関わるメタデータは、ビットレベルリポジトリの外部で持つということも考えることができる。
- 日本では電子情報の長期利用保証に関する取り組みが進んでいない。原因の一つとして OAIS や METS などの基本文書の日本語訳が存在していないことが考えられる。この分野の基本的な知識を国内で普及させるために、基本文書の日本語訳とその公開が望まれる。
- デジタル形式での保存と紙媒体での保存のコスト比較を示すなどして、保存のための基準となるポリシーを示す必要がある。また、恒久的保存基盤の利活用を想定した具体的なユースケースを示す必要がある。

本報告書では、電子情報の長期保存に有用な技術要素の観点から海外文献調査を行ったが、今後は以上のご意見を踏まえ、技術要素の調査に加えて「保存のためのポリシー」や

「恒久的保存基盤の利活用の具体的なユースケース」についても調査を実施したい。

7 References

[1] A. Dappert. DIGITAL PRESERVATION METADATA STANDARDS. 2010, ISQ Vol22, Issue2. https://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf, (accessed 2017-03-23).

[2] A. Thirifays, K. Johansen. Towards a Common Approach for Access to Digital Archival Records in Europe. 2015, iPres 2015 conference proceedings. https://phaidra.univie.ac.at/view/o:429524, (accessed 2017-03-23).

[3] ANSI/NISO Z39.84-2005 (R2010) Syntax for the Digital Object Identifier. http://www.niso.org/apps/group_public/project/details.php?project_id=62, (accessed 2017-03-27).

[4] ANSI/NISO Z39.87 -2006 (R2011) Data Dictionary - Technical Metadata for Digital Still Images. http://www.niso.org/apps/group_public/project/details.php?project_id=69, (accessed 2017-03-27).

[5] Artefactual Systems Inc. Archivematica. https://www.archivematica.org/en/, (accessed 2017-03-27).

[6] B. Caron, et al. Experiment, Document & Decide: a Collaborative Approach to Preservation Planning at the BnF. 2015, iPres 2015 conference proceedings. https://phaidra.univie.ac.at/view/o:429524, (accessed 2017-03-23).

[7] California Digital Library. UC3-Merritt. https://merritt.cdlib.org/, (accessed 2017-03-27).

[8] D. Minor, D. Sutton, et al. Chronopolis Digital Preservation Network. 2010, The International Journal of Digital Curation. http://library.ucsd.edu/chronopolis/_files/publications/chronopolis_dcc_revised.pdf, (accessed 2017-03-23).

[9] D. Rosenthal, et al. Enhancing the LOCKSS Digital Preservation Technology. 2015, D-Lib Magazine, Vol. 21, No. 9/10. http://www.dlib.org/dlib/september15/rosenthal/09rosenthal.html, (accessed 2017-03-23).

[10]DCMI Usage Board. DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/, (accessed 2017-03-27).

[11]D. Brickley, L. Miller. The FOAF Project. http://www.foaf-project.org/, (accessed 2017-03-27).

[12]E. Zierau, N. McGovern. Supporting the Analysis and Audit of Collaborative OAIS's Using an Outer OAIS-Inner OAIS (OO-IO) Model. 2014, iPres 2014 conference proceedings.
https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf, (accessed 2017-03-23).

[13]E. Zierau, U. Kejser. Cross Institutional Cooperation on a Shared Bit Repository. 2013, Journal of the World Digital Libraries, vol.6, no.1.

[14]G. Oliver, S. Knight. Storage is a Strategic Issue Digital Preservation in the Cloud. 2015, D-Lib Magazine, Vol. 21, No. 3/4.
http://www.dlib.org/dlib/march15/oliver/03oliver.html, (accessed 2017-03-23).

[15]H. Sompel, M. Nelson. Reminiscing About 15 Years of Interoperability Efforts. 2015, D-Lib Magazine, Vol. 21, No. 11/12.
http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html,
(accessed 2017-03-27).

[16]ICOM. The LIDO Working Group.
http://network.icom.museum/cidoc/working-groups/lido/, (accessed 2017-03-27).

[17]IETF. The BagIt File Packaging Format (V0.97).
https://tools.ietf.org/html/draft-kunze-bagit-14, (accessed 2017-03-27).

[18]ISO/IEC 21000-2:2003 (MPEG-21 DIDL).
https://www.iso.org/standard/35366.html, (accessed 2017-03-27).

[19]Islandora Foundation. Islandora. https://islandora.ca/, (accessed 2017-03-27).

[20]J. Pawletko, P. Caplan. Towards Interoperable Preservation Repositories Repository Exchange Package Use Cases and Best Practices. 2011,
https://libraries.flvc.org/documents/181844/502298/Interoperable+Preservation/015eb2e3-5625-43b9-ba94-c6d2c4d1bb04, (accessed 2017-03-23).

[21]JISC. SWORD v2 (Simple Web-service Offering Repository Deposit version 2).
http://swordapp.org/sword-v2/, (accessed 2017-03-27).

[22]M. Klindt, K. Amrhein. One Core Preservation System for All your Data. No Exceptions!. 2015, iPres 2015 conference proceedings.
https://phaidra.univie.ac.at/view/o:429524, (accessed 2017-03-23).

[23]M. Mayernik, et al. Linking Publications and Data: Challenges, Trends, and Opportunities. 2016, D-Lib Magazine, Vol. 22, No. 5/6.
http://www.dlib.org/dlib/may16/mayernik/05mayernik.html,
(accessed 2017-03-23).

[24]Open Archives Initiative. Open Archives Initiative - Object Exchange and Reuse.
http://www.openarchives.org/ore/, (accessed 2017-03-27)

[25]Open Archives Initiative. Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/pmh/, (accessed 2017-03-27).

[26]T. Liebetraut, K. Rechert. Management and Orchestration of Distributed Data Sources to Simplify Access to Emulation-as-a-Service. 2014, iPres 2014 conference proceedings.
https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf, (accessed 2017-03-23).

[27]The Consultative Committee for Space Data Systems. Audit and Certification of Ttustworthy Digital Repositories. 2011, Recommended Practice CCSDS 652.0-M-1.
https://public.ccsds.org/pubs/652x0m1.pdf, (accessed 2017-03-27).

[28]The Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). 2012, Recommended Practice CCSDS 650.0-M-2.
https://public.ccsds.org/pubs/650x0m2.pdf, (accessed 2017-03-27).

[29]The German National Library. LMER (Long-term preservation Metadata for Electronic Resources).http://www.dnb.de/EN/Standardisierung/LMER/lmer_node.html,(accessed 2017-03-27).

[30]The Library of Congress. EAD: Encoded Archival Description.
https://www.loc.gov/ead/, (accessed 2017-03-27).

[31]The Library of Congress. METS (Metadata Encoding and Transmission Standard).
http://www.loc.gov/standards/mets/, (accessed 2017-03-27).

[32]The Library of Congress. MIX (Metadata for Image in XML Schema).
http://www.loc.gov/standards/mix/, (accessed 2017-03-27).

[33]The Library of Congress. Metadata Object Description Schema: MODS.
http://www.loc.gov/standards/mods/, (accessed 2017-03-27).

[34]The Library of Congress. PREMIS: Preservation Metadata Maintenance Activity.
http://www.loc.gov/standards/premis/, (accessed 2017-03-27).

[35]The Library of Congress. textMD (Technical Metadata for Text).
https://www.loc.gov/standards/textMD/, (accessed 2017-03-27).

[36]The National Archives. How Cloud Storage can address the needs of public archives in the UK. 2015.
http://www.nationalarchives.gov.uk/documents/archives/Preserving-Digital-CloudStorage-Guidance_March-2015.pdf, (accessed 2017-03-27).

[37]UC Curation Center, California Digital Library, University of California, Office of the President. Unified digital Format Registry (UDFR) Final Report. 2012.
http://www.udfr.org/project/UDFR-final-report.pdf, (accessed 2017-03-23).

[38]iRODS Consortium. iRODS. https://irods.org/, (accessed 2017-03-27).

[39]永崎研宣. 国際的な画像の相互運用の枠組み IIIF について. 2016, デジタルアーカイブの連携に関する実務者協議会（第５回）.
http://www.kantei.go.jp/jp/singi/titeki2/digitalarchive_kyougikai/jitumu/dai5/siryou1_1.pdf, (accessed 2017-03-23).

[40]細川聖二. グローバルなダーク・アーカイブ CLOCKSS：学術コミュニティーによる電子ジャーナルの長期的保存への取り組み. 情報管理. 2016, vol. 59, no. 3, p. 156-164.
https://www.jstage.jst.go.jp/article/johokanri/59/3/59_156/_article/-char/ja/,(accessed 2017-03-23).

[41]木目沢司. 「国立国会図書館デジタルコレクション」の OAIS 参照モデルへの準拠状況「近代デジタルライブラリー」からの転換. 情報管理. 2015, vol. 58, no. 9, p. 683-693.
https://www.jstage.jst.go.jp/article/johokanri/58/9/58_683/_article/-char/ja/,(accessed 2017-03-23).

[42]林豊. ResourceSync: OAI-PMH の後継規格. カレントアウェアネス. 2015, no. 323, p. 17-21. http://current.ndl.go.jp/ca1845, (accessed 2017-03-23).

8 用語集

メタデータ関連用語

Dublin Core	メタデータ語彙の共通化を意図して設計された、基本語彙セットの国際標準。
EAD	目録、索引などの検索手段やその他資料に関する情報を符号化するための国際規格。
LIDO	アーカイブ資料をウェブ上で公開・利用することを意図したメタデータスキーマ。 博物館資料のメタデータを記述できる特色を持つ。
LMER	ドイツ国立図書館が開発した長期保存のためのメタデータスキーマ。
METS	XML ベースで記述されたメタデータを交換するためのメタデータコンテナ。
MIX	デジタル画像に用いられる技術データを記述するためのメタデータスキーマ。
MODS	書誌情報のためのメタデータスキーマ。
PREMIS	通常は、保存メタデータを規定したデータ辞書とその XML スキーマを指す。 本来はデータ辞書を作成した国際ワーキンググループの名称である。
textMD	テキストファイルの技術メタデータを記述するためのメタデータスキーマ

その他の用語

OAIS	長期保存システム参照モデルの国際標準。
SIP	OAIS での提出(受入)用情報パッケージ。

AIP	OAIS での保管用情報パッケージ。
DIP	OAIS での配布用情報パッケージ。
DOI	デジタルオブジェクトのための識別子。
IIIF	画像データを相互運用可能にするための一連の API の総称
PDI	内容情報（コンテンツ）の由来、他の情報との関係を示す文脈、内容情報を同定するための ID 等の参照、内容情報が改変されていないことを示す固定性からなる。
ResourceSync	OAI-PMH の後継規格として検討されている、コンテンツの同期を行うためのプロトコルとその関連規格の総称。

【執筆者一覧】

電子情報サービス課　木目沢　司
次世代システム開発研究室　原田　久義
里見　航

【責任編集】

次世代システム開発研究室
電子情報サービス課