ディジタルリソースの保存方式選択のためのガイドライン

白 才恩　　杉本 重雄

筑波大学大学院図書館情報メディア研究科

概要

大量のディジタルリソースが蓄積，発信されている現在，持続的に増加するディジタルリソースを将来に向けて保存，管理することの重要性が広く認められている。一方，ディジタルリソースが将来的に利用可能であり続けさせることの難しさも広く認められている。そこで，本稿ではディジタルリソースを長期保存するため，ディジタルリソースを保存するためのガイドラインを作ることを目的として，保存のためのメタデータに関する検討，ディジタルリソースのタイプによる特徴を検討する。ここでは，完全な保存は困難であるとの前提に立ち，ガイドライン作成の基礎となる，リソースの「エッセンス」をのこすための基準を検討する。

Selection Guidelines for Preservation Method of Digital Resources

Jae Eun Bak　　Shigeo Sugimoto

Graduate School of Library, Information and Media Studies, Univ. of Tsukuba

Abstract

Preservation of digital resources has been widely recognized as an important but difficult issue in the networked information society. Many studies have been done on this issue, and basic issues such as emulation and migration have been raised for digital preservation. This paper is aimed at discussing issues to develop guidelines for digital preservation. It first describes metadata schemas for digital archives, and then describes the characteristics of digital resources. This paper discusses several factors to identify issues to build the guidelines for digital preservation.

1. はじめに

1990年代に始まるインターネットとWebの急速な発展によって，大量かつ多様な情報資源（リソース）が発信され，多くの利用者に受け入れられてきた。現在では，ネットワークなしに我々の日々の活動を行うことは考えられなくなっている。その中で，これまでに大量の情報資源が発信され，蓄積されてきた。その一方，こうした資源を将来に向けて保存，管理するという課題を作り出してきた。

コンピュータとネットワークの発展と普及によって，我々が文書を作り，それを流通させる環境は大きく変わってしまった。従来は，「紙に印刷された文書」や「紙に書いた文書」が情報流通の主たる媒体であったが，現在では，パソコンで文書を作り，そのままネットワーク上に送り出すことが一般化している。すなわち，ディジタル形式で作り，そのままディジタル形式で流通，蓄積することが広く行われている。ディジタルリソース，殊にBorn Digitalなリソースの保存は容易ではないことが広

く知られている。図書館や公文書館のように，資料を将来の利用者のために保存する役割を持つ機関にとって，ディジタルリソースの増加は，単に量的な面で対応が難しいのみならず，保存方法，保存技術の視点からも大きな問題である。

たとえば，20 年後に，現時点で保存したディジタル資料が，現在と同じように再生することは可能であろうか。再生が不可能な場合，どのようにすれば再生できるのであろうか，また，どのようなことができれば再生できたことになるのであろうか。こうしたことがディジタルリソースの保存に関する基本的な視点である。ディジタルリソースは新しいコンピュータ環境に移していかざるを得ない。新しい環境の中で，古いリソースを再生するために，エミュレーションによる方法，マイグレーションによる方法が検討されてきている[1][2]。しかしながら，いずれの場合でも完璧なリソースの保存は不可能である。したがって，「保存すべきものは何であるか」と「保存可能なものは何であるか」を検討した上で，保存の方針，方法を決めなければならない。

紙の資料の場合，「もの」として保存すれば内容はそのまま保存されることになる。これに加えて，保存の経緯や他の資料との関係等，資料の構造に関する情報を加えて保存すればよい。一方，従来，紙の資料をマイクロフィルム化することが広く行われてきている。この場合，紙の資料が持つ機能の一部を失うこと，そして読み取り機械を利用することを前提とした保存であると言える。

ディジタルリソースは，ディジタル化して作ったリソースであれ，Born Digital リソースであれ，物理的な実体ではないので，リソースを格納した「もの」を保存することは意味を持たない。リソースの内容を人間に可読な形式にするための情報，リソースの構造や他のリソースとの関連・コンテキスト等がないとリソースの内容の再生に大きな支障をきたすことがある。ディジタルリソースはパソコンのような機材を利用して読まねばならない。すなわち，リソース自体に加えて，それを再現するために必要な機材，ソフトウェアも保存されなければならないことを意味する。しかし，機材やソフトウェア，さらには技術そのものを長期にわたって保存し続けることは現実的ではない。そのため，リソースを新しい環境に移植するか，古い環境を生かし続けることの必要性が生じる。しかしながら，これらも完璧な解ではない。リソースを安全に保存するには，広く用いられている標準規格を利用することが望ましい。ところが，ディジタルリソースはそうした標準規格を用いたものばかりではない。紙の資料を保存する際にマイクロフィルムを用いることが認められたように，ディジタルリソースの機能をある程度犠牲にしても，リソースの本質的な内容を残すことが求められる。報告書[3]では，このことを文書の「エッセンス」を残すとあらわしている。現実には，「エッセンス」がなにであるかのコンセンサスなしにはディジタルリソースの機能制限は行いにくい。

ディジタルリソースの保存のしやすさや保存方法はリソースの形式に依存する。一般には，特定のソフトウェアに依存しないテキスト形式は保存に対しては強く，リソース自身の中に動作記述のあるものは保存に対して弱いといえる。しかし，テキスト形式であっても，外字を含むような場合もあり，また，文字コード体系とそれに関連付けられた字体の標準規格は時とともに変更されるため，単にテキストデータを残すだけでは十分ではない。

本稿では，ディジタルリソースの保存に関し，アーカイブと保存のためのメタデータ，文書のタイプに応じた基本的な保存の方法や課題を述べ，保存方法のガイドライン作りを提案する。

2. ディジタルリソースの保存のためのメタデータ

前に述べたように，ディジタルリソースには，非ディジタル資料をディジタル化して作ったもの（ディジタル化リソース）と，もともとディジタル形式で作られたもの（Born Digital リソース）がある。

リソースの保存には，そうした資料の特性，記述の目的にあったメタデータスキーマを準備する必要がある。リソースの記述は，アーカイブされるリソースの集まりの構造と個々のリソースの書誌的内容に関する記述，リソースの内容，特にファイル間の関係を含むディジタルリソースの構造記述，そしてディジタルリソースの保存のための記述というように異なる目的，異なる視点からの記述が必要である。アーカイブのために用いられるメタデータスキーマには，公文書等のアーカイブのためのISAD(G) [4]，ディジタルリソースのために作られた EAD[5]，ディジタルリソースの保存のためのMETS[6]やPREMIS[7]といったメタデータスキーマがある。以下ではこれらを簡単に説明する。

2.1 アーカイブのためのメタデータスキーマ　－　ISAD(G), EAD, PREMIS, METS

(1) ISAD(G) (General International Standard Archival Description)

ISAD(G)は，もともと公文書館等における文書のアーカイブのために提案されたものであり，ディジタルリソースに特化したものではない。アーカイブされたリソースの全体的な記述のために用いられる。その記述には，リソースがどんな形態であり，どこの機関で作られ，どうやって保管されていたかを記述する。例えば，リソースの履歴，入手の情報，内容，評価・廃棄・保存の期間，利用及び複写条件，リソースのコンテキストに関した記述要素がある。

(2) EAD (Encoded Archival Description)

EADは，ISAD(G)との互換性を考慮して定義されているディジタルリソース向きのメタデータスキーまで，Making of America 等でディジタルリソースの記述に用いられてきた。「必要な要素は何か」に対する意味の内容を具体的に配列・構成することである[8]。EADはディジタルリソースの内容記述に加えて構造表現のための記述要素を持つ。

(3) METS (Metadata Encoding and Transmission Standard)

METSはEADでの経験をもとに作られたもので，ディジタルリソースの構造と内容を表現するメタデータスキーマである。ディジタルリソースの保存のための参照システムを決めるOpen Archival System (OAIS)[9]の情報パッケージとして利用することもできる。

(4) PREMIS (Preservation Metadata and Implementation Standard)

PREMISは，ディジタルリソースの保存を目的に作られたメタデータスキーまで，図1に示すデータモデルを基礎にしている。リソースの記述内容と構造・技術的内容を表す要素を分離しているほか，リソースにかかわる権利や関係者，イベントも別にとらえている。

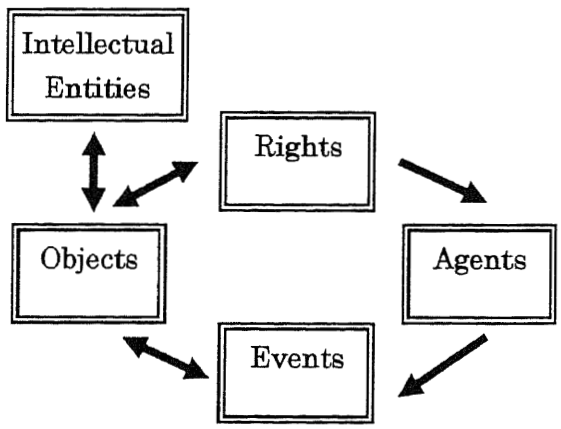

図1. PREMISのデータモデル

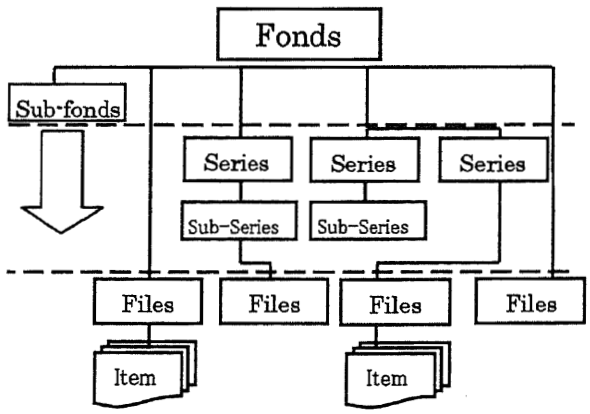

図2. Fondの概念構成図

2.2 アーカイブメタデータを支えるモデル

図2に示すように，ISAD(G)は階層構造（Level of Description）を持つ。この構造は，リソースのコレクションの捉え方を決めており，アーカイブ全体の構造を表しているともいえる。

EADやMETSは基本的にはリソースを対象として記述する。これらは記述対象リソースの内容とその構造に関する記述要素を持っており，いずれもXMLでの記述形式を決めている。METSの場合，内容記述，管理情報記述，リソースの構造記述を陽に区分して構成している。来歴情報の記述要素や，ソフトウェアの動きに関する記述要素など，保存指向の記述要素を持っている。

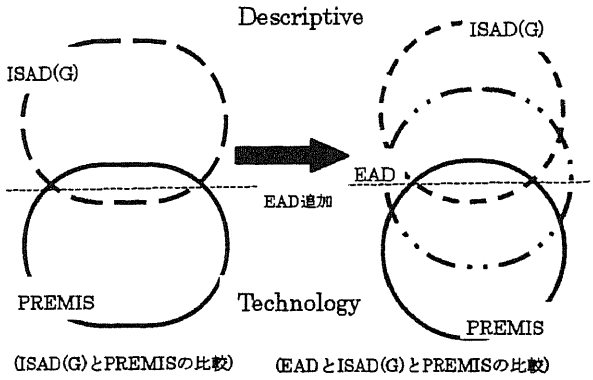

図3. ISAD(G)，EAD，PREMISの特色

PREMISは，図1に示したIntellectual Entities, Objects, Rights, Agents, Eventsの5種類の記述対象実体（エンティティ）からなる単純なモデルを基礎にしている。このようにエンティティをとらえることで，リソースそのものの知的内容と構造的内容をより明確に捉えられるとともに，リソースにかかわる関連実体に対する視点を明確にしている。

図3はこうしたスキーマの間の関係を示す。ISAD(G)はこのリソースの組織化全般をふくむ記述的内容を，EADやMETSはリソースの構造や管理情報と内容記述，PREMISはこのディジタルリソースに対する内容と構造の記述に指向している。

3. ディジタルリソースの保存のためのリソースの特性の分類

ディジタルリソースには多様なリソースのタイプがある。紙，オーディオテープ，ビデオテープ等からディジタル化されたリソース，ワープロ文書，Web-page，動画等のBorn Digitalリソースがある。以下では，リソースをディジタル化リソースとBorn Digitalリソースの二つに分け，それぞれについてリソースのタイプをいくつかに分けて考察する。

3.1 ディジタル化されたリソース

既存の非ディジタル資料をディジタル化するケースである。多くのディジタルライブラリでディジタル化資料が提供されてきている。一般的に，資料の性質にあったディジタル化方法がとられる。原資料が貴重なものである場合は，原資料を保存する一方，電子化資料を閲覧用に用いる場合が多い。また，一般の資料の場合には，閲覧利用に加えて，電子化した資料を保存にも利用する場合がある。

(1) テキスト

本稿では，文字を主体とする文書をテキストと呼ぶ。ごく一般的な図書，雑誌，業務文書，手紙，マニュスクリプトなどを含めてテキストと呼ぶことにする。したがって，ディジタル化されたテキストリソースは冊子体の図書を電子化して作ったもの，業務文書を電子化したもの，歴史資料であるマニュスクリプトを電子化したものなど多様である。OCRや人手でテキストデータ（Plain TextやXMLテキスト）のみを取り出すものもあるが，原資料のイメージを保存するためスキャナやカメラでの電子化が一般的である。イメージデータの作成方法は資料の性質によってさまざまである。いずれの場合にも，基本的にはイメージデータとして作成される。イメージデータのフォーマットとして，TIFF，JPEGなどが用いられ，ページデータとして扱うためにPDFが用いられることもある。

こうして作成したデータに，たとえば図書や業務文書，マニュスクリプトとして構造を反映するための機能やいろいろなユーザインタフェース機能を加えてディジタルリソース化される。こうした作業の中で，ハイパーリンク機能の埋め込みや，特殊なユーザインタフェースを利用した機能の付加もなされる。また，データベース管理機能を用いて電子化データのコレクションを管理することも行われる。作成したリソースのメタデータとして先述の EAD のようなアーカイブ指向のメタデータスキーマを利用してメタデータを作成する。

(2) イメージ（静止画イメージ）

イメージタイプの資料には写真，絵，がある。また，地図や絵巻物などテキストとイメージの両方の性質を持つものもある。また，テキストの中に写真が埋め込まれたものもある。絵巻物の中の絵だけを取り出したり，テキストに埋め込まれた写真を取り出したりしてひとつのイメージとして扱うことも可能であるが，それらはディジタル化の方針に依存する。テキストの場合と同様，イメージデータのフォーマットとして，TIFF，JPEG などが用いられる。また，拡大縮小といった機能がテキストに比べてより強く求められるため，それに適したインタフェースのための形式が選ばれることも多い。メタデータに関してもテキストの場合と同様である。

(3) ビデオ

映画やビデオ映像をディジタル化して蓄積保存する。入力となるデータの形式は，アマチュア用のものからプロ用のものまで，いろいろな種類の映画フィルムや VHS 他のアナログビデオテープがある。ディジタル形式ではあっても DVD や CD，LD などに入れられたパッケージ系のビデオコンテンツも，保存のための変換を必要とするものの場合はアナログ資料と同様である。こうした資料を MPEG などの標準フォーマットでディジタル化する。メタデータに関しても MPEG7 のようなビデオ用メタデータスキーマを利用してコンテンツの詳細な内容を記述することができる。Closed Caption のあるものなど，テキストデータ化の対象となるものも考えられる。

(4) オーディオ

ビデオと同様にオーディオテープ，レコード，CD などからのディジタル化である。AVI，WAV，MP3 等のフォーマットでディジタル化し，保存する。

3.2 Born Digital リソース

はじめからディジタル形式のデータとして作られ，ネットワークやパッケージメディアを利用して提供されるリソースである。元々のディジタル形式が必ずしも保存に適したものとは限らないため，他のディジタルフォーマットへの変換が必要になる場合がある。以下に前節と同様にテキスト，イメージ，ビデオ，オーディオに分けて簡単に考察する。

(1) テキスト

ワープロ文書，表計算ソフト文書，プレゼンテーション，印刷形式文書，Web ページ，Plain Text，XML 等がある。Born Digital リソースの保存には，リソースを格納したファイルに加えて，リソースの再生に必要なハードウェアやソフトウェアも利用可能な状態にしなければならない。テキストタイプのリソースではリンクを持つものがある。リンクは，複数のファイルでできた「ひとつのリソース」の中を結ぶものと，外部のリソースを結ぶものがある。「ひとつのリソース」を構成する複数のファイルを同定することが求められる。

(2)イメージ： 静止画イメージおよびグラフィックスデータ

静止画イメージのファイルフォーマットとしては GIF， JPEG， TIFF， BMP 等がある。静止画イメ

ージに関しては基本的にディジタル化データと同じである。また，コンピュータグラフィックス(CG)ツールを用いて作られたグラフィックスデータは2次元および3次元のものがある。CGによるアニメーションのようにビデオと同様のものもある。いずれにしてもグラフィックス用のソフトウェアへの依存性が高い。

(3) ビデオ（動画)

ディジタル動画であるビデオにはAVI，MOV，WMV，MPEG等がある。原資料の有無を除いて，ディジタル化データと基本的には同じである。

(4) オーディオ

Born Digitalリソースであるオーディオのファイルフォーマットとしては WMV，AIFF，WMA，MP3等がある。また，MIDI形式のように演奏データとして作られるものもある。原資料の有無を除いて，ディジタル化データと基本的には同じである。

4. 保存のためのガイドラインについて

本節では，ディジタルリソースの完全な保存は困難であり，ディジタルリソースが持つ機能をどの程度犠牲にすれば，より確実な保存が可能であるかという立場から考察する。ディジタルリソースの実現に利用される技術を制限することはできない。一方，すべての表現技術が保存に向いているわけではない。そのため，保存のはじめの段階で，できるだけ落ち着き，かつ広く利用されている標準規格の形式に変換して保存すれば，長期の保存が行いやすくなる。しかしながら，リソースの実現形式を変換することは，もとのリソースが持ついろいろな機能や使い勝手の良さを失わせることを意味する。たとえば，ハイパーリンクと動画を含むHTML文書の印刷イメージを印刷して保存するのと同様に，ページイメージデータとして保存すれば，安定した形式での保存が可能になる。しかしながら，これはハイパーリンクや動画，テキストによる検索可能性といったリソースの機能を制限することとを意味する。また，公文書の場合のようにリソースの原本性が問題になるような場合，保存リソースが原リソースと同一のものであると認定するための基準を準備しなければならない。すなわち，ディジタルリソースの保存は一様な方法で行う必要はなく，資料の価値や保存のポリシーに応じた保存方法を選んで適用すべきである。そのため，互いに矛盾する可能性があるいくつかの異なる要件の中で，保存の目的や環境要件に最も適した保存方針を決めることが求められる。

そこで，本稿では，ディジタルリソースの保存のヒントを得るためのガイドライン作成を目的として下に示すいくつかの視点を提案する。保存のガイドラインを決める重要な要素は，保存のためにどこまで難しさに対処するか，言い換えると保存にかけられるコストである。重要なリソースはコストをかけて保存する必要があり，そうでないものはそれなりのコストでの保存を行わざるを得ない。

(1) 保存にかかるコスト
  (a) 保存のための初期変換コスト
  (b) メインテナンス（マイグレーションやエミュレーションを含む）のコスト
    (i) ファイル形式の種類とその多さによるコストの違い

(2) 経年後に求められる要件
  (a) 原本性（保存時の内容と同一であること）
  (b) 見読性（人間が内容を読めること）
  (c) 検索性（内容を何らかの手段で探せること）
  (d) 機能性（リソースに組み込まれた機能）

(i) Look-and-Feel，ハイパーリンク機能，プログラムやマクロ等の機能

(ii) 編集機能　（ワープロ文書等を保存する場合）

(3) 機能制限とそのポリシー（原資料と保存形式の差異の許容範囲）

(a) 初期変換による機能低減とその許容範囲

(b) 環境の変化による機能低減とその許容範囲

(4) リソースの保存に関する情報（保存のためのメタデータ）の蓄積，利用

(a) 保存の過程とその作業にかかわる情報を蓄積すること（OAIS の保存記述情報[9]）

(b) 保存したリソースのアクセス許可対象等の情報の蓄積

ディジタルリソースにはいろいろな種類に応じた保存が求められる。データベースやプログラムとして実現されるリソースの保存は重要な課題であるが，本稿では扱わないことにする。これらは，動作するシステムとして維持管理が続く限り内容は保存されることになる。単純なイメージデータ，ビデオデータ，オーディオデータは，できるだけ広く用いられる標準規格の形式で保存すればよいと考えられる。そこで，ここではテキスト資料についてのみ考察する。ただし，下記のようなごく基盤的な標準に関して，適切な情報が残されることも仮定する。

(1) 文字コードと対応するフォント

(a) 利用される文字コードやそのフォントに関する情報，外字対応のための情報

(2) イメージ等のデータ形式

(3) XML および HTML

(a) タグの定義（XML），HTML のバージョン

テキスト資料については，一般的なワープロ文書やプレゼンテーション文書のようにひとつのファイルとして実現されるものと，Web ページのようにいくつかのファイルによってひとつの文書が組み立てられるものがある。いずれの場合も，ひとまとまりのリソースとしての保存を考える。リソースの保存形態として大雑把には下のように分けて捉えることができる

(1) 元のファイル形式

(a) 何も手を加えないもの

(b) インラインイメージや他のリソースへのリンクなど，保存対象リソースができる限り一貫性を失わないようにするための処理を施したもの

(c) 対象リソースの利用環境要件に含まれるソフトウェア（たとえば，プラグインソフト）を，保存に適したものに変更するなどの処理したもの

(2) 印刷イメージ（スナップショット）形式

(a) イメージデータのみ

(b) イメージデータに内容のテキストを加えたもの。

(c) 上に加え，他のリソースとの関連（たとえばリンク）の一貫性のための処理をしたもの

ごく簡単なシナリオをいくつか示す。

- 原資料の見読性と原本性が強く要求される場合：　単純に印刷イメージ（画像データあるいは PDF のような印刷向きの形式）で保存し，それに適切なメタデータを加えればよいであろう。ただし，この場合，印刷イメージとして作られる資料を，もとの資料が持つ機能とは無関係に，保存のための原本として認める必要がある。ワープロ文書のような場合は，機能低下が目立たないと思われるが，その他の場合は利用者にかなり大きな機能低下を感じさせることになる。
- 安いコストでできるだけ多くの HTML 文書を保存する Web アーカイブの場合：　HTML 文書をそのま

ま保存することになるが，収集保存日時に応じて文書間の参照を無矛盾にするための仕組みを必要とする。

- 保存するリソースの提供可能範囲を限定されたリソースの保存：（見読性や原本性，機能的制約に関する要件を満たす以外に）リソースの提供者からアクセスを制限するための情報を収集し，メタデータとして情報を残していかねばならない。アクセス制限は時間経過とともに変化する可能性があり，それに適したメタデータ記述と保存が求められる。

5. おわりに

本研究の目的は，多様なディジタルリソースを長期にわたって保存するために，適切な方法を選択するためのガイドラインを構成することである。現在まで，方法選択のための基本的な要因に関する検討を進めてきた。ディジタルリソースの保存方法を判断し，保存作業を行う過程では，保存のためのメタデータが必要とされる。これまで保存されたリソースの内容やその構造，保存履歴等を記述するためのメタデータが作られてきた。ここでは，既存の保存のためのメタデータを基礎として検討し，保存方法を与えるガイドラインを作るための視点を明確にしていきたいと考えている。

なお，本研究は，一部科学研究費補助金（基盤(B)19300081）による。

参考文献

[1] National Library of Australia: "PADI: Preserving Access to Digital Information , Emulation" (online) , http://www.nla.gov.au/padi/topics/19.html (2007 年 10 月 5 日参照).

[2] National Library of Australia: "PADI: Preserving Access to Digital Information , Migration" (online) , http://www.nla.gov.au/padi/topics/21.html (2007 年 10 月 5 日参照).

[3] 内閣府: "中間段階における集中管理及び電子媒体による管理・移管・保存に関する報告書" (online) , http://www8.cao.go.jp/chosei/koubun/kondankai14/houkoku.pdf (2007年10月5日参照).

[4] ICA; International Council on Archives: "ISAD(G): General International Standard Archival Description, Second edition" (online) , http://www.ica.org/en/node/30000 (2007 年 10 月 5 日参照).

[5] The Library of Congress: "EAD: Encoded Archival Description" (online), http://www.loc.gov/ead (2007 年 10 月 5 日参照).

[6] The Librarian of Congress: "METS: Metadata Encoding and Transmission Standard" (online), http://www.loc.gov/standards/mets (2007 年 10 月 5 日参照).

[7] OCLC; Online Computer Library Center: "Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group" (online) , http://www.oclc.org/research/projects/pmwg (2007 年 10 月 5 日参照).

[8] 難波忠清ほか:" 核融合アーカイブズデータベースの共有化" (online), http://www.nifs.ac.jp/archives/01aA31P.pdf#search='ead の特徴' (2007 年 10 月 5 日参照).

[9] Consultative Committee for Space Data Systems: "Reference Model for an Open Archival Information System (OAIS)" (online) , CCSDS Blue Book 650.0-B-1, 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf (2007 年 10 月 5 日参照).