資料１－２

ジャパンサーチ（仮称）の連携フォーマット（案）

目次

1　基本的な考え方

1.1　連携フォーマット

- ジャパンサーチ（仮称）（以下「ジャパンサーチ」とする）は、連携のための単一のメタデータモデルを定義せず、データを提供する連携機関のメタデータモデルをそのままの形（オリジナルのモデル）で受け入れる。
- 受け取ったメタデータは、ジャパンサーチが実装する管理ツールを用いて処理することで、全項目を対象とする単純な検索ができるようにする。
- 異なるデータベースのメタデータの一覧表示や横断検索を可能にするため、「共通項目ラベルの付与」を行う。作業負荷を軽減するため、自動で候補を推定するとともに、必要最低限の項目に絞るものとする。

1.2　連携仕様

- 連携を容易にするため、原則、ファイルベースでの連携を行う（OAI-PMH は必須としない）。
- 多様なファイル形式に対応する。

1.3　利活用フォーマット[1]

- メタデータの利活用促進のため、オリジナルのメタデータモデルとは別に、分野を横断する標準的なメタデータモデル（＝利活用フォーマット）を定義する。
- メタデータは、オリジナルモデルと利活用フォーマットの両方で保持する。
- 利活用メタデータモデルの生成のためのマッピング作業は、連携後に、国立国会図書館が行う。必要に応じて連携機関の確認を得る。

2　連携仕様

2.1　メタデータ項目

連携機関のメタデータモデルをそのままの形で受け入れる。メタデータモデルは原則として自由であるが、最低限の必須項目（①オリジナル（ソース）データの一意の ID、②名称/タイトルの 2 項目）がある等の制約を設ける。

必須項目の詳細については、3.3.1 参照。

2.1.1　各メタデータ項目の値

各メタデータ項目の値は、下表にあるデータ種別に対応する。基本は全て文字列として扱われるが、ラベル定義（3.3 参照）の際にデータ種別を指定することも可能とする。

[1] 利活用フォーマットについては、資料１－３及び資料１－４を参照。

表１　メタデータ項目として対応可能な値

データ種別	内容
文字列	文字列としての検索が可能。改行については、「¥n」にエスケープされ、UI に表示される際には BR タグに変換される。また、URL はリンクとして表示される。
HTML	文字列のヴァリアントで、UI で表示される時に HTML として表示される。XSS を防ぐため以下のタグ・属性しか利用できない（データ登録時に消去される）。 H、P、BR、DIV、SPAN、TABLE、TR、TH、TD、attr:style 項目名の最後に「_h」が追加される。
日付	日付として変換される（UnixTime として保持する）。変換に失敗した場合、null になる。変換対応フォーマットは別途用意する資料（技術情報を集約した詳細版を作成予定）に示す。 項目名の最後に、「_d」が追加される。
真偽値	真偽値。JSON の true/false に対応。「true」「TRUE」「1」が true として、それ以外の全ての値は false として解釈される。 項目名の最後に、「_b」が追加される。
数字	数字（小数点含む）。JSON の number に対応。数字に変換できないエラーデータが入っている場合、0 が入る。 項目名の最後に、「_n」が追加される。

2.2　連携方式

ファイルによる連携を基本とし、表２のとおりの連携方式の選択を可能とする。ファイル連携の場合は、連携・更新の都度、全データが格納されたファイルを用いて連携を行う。

表２　ファイルの連携方法

①ファイルアップロード	対応可能なファイルフォーマット（2.3 参照）のファイル（20MB まで）を、管理画面から手動でアップロードする方式。
②ファイルを Web に掲載し、ジャパンサーチがファイル取得する	管理画面でファイルの URL を入力し、ジャパンサーチから取得しに行く方式。ベーシック認証、ダイジェスト認証にも対応可能。応答の形式が対応ファイルフォーマット（2.3 参照）であれば、API の URL でもよい。
③ファイルを Web に掲載し、ジャパンサーチがファイルを定期取得する	ファイル取得と同様だが、ジャパンサーチ側で定期自動実行する。実行時間の指定が可能。
④ハーベスト用 API（OAI-PMH）	データが大容量、かつ高頻度で更新を行う場合の連携方法を想定。※ただし、プロトタイプ版での実装は想定しない。

2.3　ファイルフォーマット

TSV、CSV、XLSX の表形式ファイルフォーマット[2]、及び JSON、XML の構造化ファイルフォーマットに対応する。

また、複数のファイルを zip 形式で圧縮したファイルにも対応する。（なお、指定されたファイルフォーマット以外のファイルが含まれている場合には、エラーとなる）。

表３　対応ファイルフォーマットの種類と制約

ファイル形式	制約
TSV・CSV	UTF-8 でエンコーディングされたファイルのみ受け付ける。 CSV の場合、RFC4180 に準拠していれば問題はなく、ダブルコーテーション（”）で囲まれた範囲であれば改行を含んでもよい。 TSV についても、エスケープ等は CSV に準拠する。
XLSX	1 ファイルにつき 1 シートのみ（複数シートには非対応）。セルの結合等にも対応できない。XLS 非対応のため、古いファイル等に注意が必要。
JSON	1 行 1 レコードとする JSON Lines[3]形式を推奨する。 **（例）JSON Lines** **{"id":"0001", "title":"タイトル 1", ... }** **{"id":"0002", "title":"タイトル 2", ... }** 但し、ルートを array とし、1 レコード 1 オブジェクトとする形式、ルートを object とし、key:Value の Value を 1 レコードとする形式にも対応可能とする。 なお、JSON ファイルは原則としてそのままの形で問題ないが、以下の制約がある。 ・pair の key に使えるのは、アルファベット小文字、数字、アンダーバー（_）のみ。それ以外の文字が含まれる場合、削除される。 ・1 つの array の中に、異なる種類の Value が存在してはならない。
XML	1 行 1 レコードとする形式を推奨する。 **（例）1 行 1 レコードの XML** **<root><id>001</id><title>タイトル 1</title>...</root>** **<root><id>002</id><title>タイトル 2</title>...</root>** 但し、XPath 等で、ルートの下に複数のレコードを続けていく方式にも対応可能とする。

表形式ファイルフォーマット（TSV・CSV、XLSX）では、列を項目、行をメタデータの単位として解釈する。1 行目はヘッダとして指定することを可能とする（アルファベット小文字・数字のみ使用可、重複不可）。これらの項目名は、ジャパンサーチのシステム内部で

[2] 表形式のデータを作成する際のポイントについては、『デジタルアーカイブの構築・共有・活用ガイドライン』（平成 29 年 4 月）p.39 を参照。
https://www.kantei.go.jp/jp/singi/titeki2/digitalarchive_kyougikai/guideline.pdf

[3] http://jsonlines.org/

のみ利用され、検索結果詳細画面等に表示される項目の名称は、ラベル定義（2.3 参照）で別途定義する。

構造化ファイルフォーマット（JSON、XML）では、その構造を原則としてそのまま取り込むこととする。入れ子構造にも対応するが、検索結果詳細画面等で、その入れ子の単位で表示することはできず、入れ子を展開したフラットな形式（「key：値」）での表示とする。外部 API で取得した場合は、入れ子構造での表示も可能とする。

3　連携に係る作業の流れ

ジャパンサーチとの連携作業は、大まかに、(1) データベース登録・基本情報の提供→ (2) データ登録→ (3) ラベル定義（共通項目ラベル付与、個別項目ラベル定義）→ (4) 公開の手順に沿って行われる。作業分担は以下の表のとおり。

表 4　登録から公開までの流れ

作業内容	連携機関	ジャパンサーチ
(1)データベース登録	・管理画面から、データベースの基本情報を入力	（情報をもらってジャパンサーチ側で登録も可能）
(2)データ登録	・管理画面からメタデータファイルをアップロード又はファイルを Web に掲載	
(3)ラベル定義	・共通項目ラベル候補の確認・修正 ・個別項目ラベル定義（メタデータの各項目の名前、データ形式、定義の確認のみ）	・（メタデータアナライザー[4]による）データの自動解析により、共通項目ラベル候補を提示
(4)公開	・テスト環境での確認・修正	・公開（公開後、利活用フォーマットへの変換作業）

3.1　データベース基本情報の提供

連携機関には機関 ID・パスワードが発行され、管理画面へのアクセスが可能になる。連携機関は、この管理画面から、ジャパンサーチと連携するデータベースの基本情報について、データベース単位で定義を行う。ジャパンサーチ側で情報をもらって代わりに入力することも可能とする。

設定が必要な項目は表 5 の通り。

[4] ジャパンサーチ側のシステムにおいて、データ種別や項目の充足率等から共通項目ラベル（名称/タイトルや ID 等）の候補を推定し、提示する仕組み。

表５　データベース定義項目一覧

項目名	種別	意味	制約
ID	必須	データベースの ID	アルファベット小文字、数字で 4 文字（ジャパンサーチ側で付与）
名称	必須	データベースの名称	
名称（英語）	必須	データベースの名称（英語）	
説明		データベースの説明	100 字まで（100 字以上は折り畳み表示）
説明（英語）		データベースの説明（英語）	
タイプ/カテゴリ ※選択の区分は要検討	必須	データベースが扱うコンテンツの種別。選択式。	データベースあたり一つが推奨だが、複数選択も可
メタデータの権利表示	必須	メタデータの権利情報	クリエイティブ・コモンズライセンス、政府標準利用規約等。権利情報について記述されている外部資源へのリンク（URL）も可
コンテンツの権利表示	必須	対象のデジタルデータがある場合の権利情報。例外がある場合は（メタデータの）共通項目で定義する。	
メタデータの権利区分	必須	メタデータの権利情報だが、検索・絞込み用に選択式になっている。	選択項目としては、クリエイティブ・コモンズライセンスのバリエーションその他が考えられるが、要検討。
コンテンツの権利区分	必須	コンテンツの権利情報だが、検索・絞込み用に選択式になっている。	
URL		データベースの URL	
組織名	必須	データベースの所有者の名称	
組織名（英語）	必須	データベースの所有者の名称（英語）	
組織 URL		データベースの所有者の URL	

3.2　データ登録

連携機関は、任意の連携方式（2.2 参照）によってメタデータファイルを登録する[5]。

3.3　ラベル定義

連携機関は、管理画面を通じて、共通項目ラベル・個別項目ラベルの定義を行う。この定義作業を行うことで、ジャパンサーチ上での一覧表示や横断検索が可能になる。

3.3.1　共通項目ラベル付与

3.2 で登録されたメタデータは、ジャパンサーチのシステム（メタデータアナライザー）により自動的に分析され、「共通項目ラベル」候補が自動的に提示される。連携機関は、提示された候補が適切かどうか確認し、必要に応じて管理画面で修正を行う。

共通項目ラベルとは、ジャパンサーチと連携している全てのデータベースに共通するメタデータ項目（名称/タイトル、提供者等）に付与するラベルである。ラベルを付与することで、検索結果表示が分かりやすくなる、検索の絞込みができるようになるなどのメリットがある。

共通項目ラベルには、「必須項目」（ラベルの付与が必須の項目）、「あれば必須項目」（オリジナルのメタデータに対応項目がある場合、ラベルの付与が必須の項目）、「任意項目」（ラベルの付与が任意の項目）の 3 種類がある。具体的な項目は表 6 のとおり。

表 6　共通項目ラベル一覧

項目名	種別	意味	制約
ID	必須	オリジナル（ソース）データの一意の ID。レコードの URI に使われる。	オリジナル（ソース）データ内で一意であること。 アルファベット大文字小文字、数字のみで構成されていること。
名称/タイトル	必須	レコードの名称。検索結果の表示に使われる。	
名称/タイトルヨミ	あれば必須	レコードの名称の読み。	
名称/タイトル英語	あれば必須	レコードの英語名称又はローマ字	
最終更新日	あれば必須	データの最終更新日	日付型であること。
URL	あれば必須	レコードのリンク先の URL	

[5] 登録されたデータは、ジャパンサーチのシステム側で JSON Lines 形式に変換される。

サムネイル画像 URL	あれば必須	サムネイル画像の URL	
提供者	あれば必須	オリジナルのコンテンツの提供者	当面は、ID 等では無く、文字列とする。
コンテンツの権利表示	あれば必須	対象のデジタルデータがある場合、その権利情報が、データベース定義の情報と異なる場合のみ	クリエイティブ・コモンズライセンス、政府標準利用規約等。権利情報について記述されている外部資源へのリンク（URL）も可
寄与者	任意	対象の作成に関わった人（作者、発行者、出演者等）	複数可
時間	任意	対象に関連する時間（制作年、対象時期等）	複数可
場所	任意	対象に関連する場所（発行地、制作地等）	複数可

3.3.2　個別項目ラベル定義

連携機関は、メタデータ項目全体について、各項目の内容をシステムが正しく認識できるよう、各メタデータ項目に関する名称や説明、インデックス方法、データ種別等の情報を付与する作業を行う。これを、個別項目ラベル定義という。

※個別項目ラベル定義に当たっては、入力画面で Excel 等のインプットファイルを受け付けて画面で入力する必要をなくす、元データからコピー&ペーストできるようにする等の運用を検討中。連携機関は、各メタデータ項目の名称、データ形式、定義の確認を行うのみでよいようにする。

表 7　個別項目ラベルの定義項目

項目名	種別	意味	制約
名称（日本語）	必須	詳細画面等に表示される、その項目の名称	
名称（英語）		英語画面に切り替えた時の名称。未定義の場合、検索結果は日本語が表示される。	
説明（日本語）		その項目が何を意味するかの説明。定義があれば、詳細画面でユーザが見ることが可能。	
説明（英語）		同、英語。	
格納種別/イン	必須	以下から選択（デフォルトは通常）	

デックス方法		通常 : 項目が格納され、検索できるようになる。 検索しない : 項目は格納され、詳細表示や API で取得できるが、検索はされない。(項目名の最後に、「_s」が追加される。) 除外:項目はジャパンサーチに登録されない。	
データ種別	必須	2.1.1 のデータ種別から選択（デフォルトは文字列)	

3.4　データ登録、公開

連携機関は、ラベル定義の作業が終了すれば、テスト環境で実際のメタデータの提供状況を確認することができる。必要に応じて、連携機関による管理画面での修正又はジャパンサーチ側での修正後、一般公開の運びとなる。