nihuINT: 多様な人文科学研究資源を統合するプラットフォーム

東京大学史料編纂所　山田太造

１．はじめに

人間文化研究機構（以下，機構）では，人文学に関わる多種多様な研究資源の有機的な利活用による研究および教育を促進するため，2008 年 4 月から nihuINT （nihu INTegrated retrieval system；研究資源共有化統合検索システム）を公開している．nihuINT が検索対象とする DB は，2013 年 10 月時点で 138 である．その内訳は，機構を構成する各機関（国立歴史民俗博物館，国文学研究資料館，国立国語研究所，国際日本文化研究センター，総合地球環境学研究所，国立民族学博物館）から 122，機構が運用する地域研究拠点のデータベース nDP（nihu Data Provider）から 4，2010 年 7 月 14 日より連携を開始した国立国会図書館から 12 である．nihuINT のアクセス数は月あたり約 350 万件に達している．

nihuINT は，各 DB の所在や操作方法を意識することなく共通の操作で横断した検索・検索結果表示が可能であり，DB を横断した検索結果一覧表示以外にも，時空間情報を用いた検索の機能などを有する．これまで以上に研究資源を探しやすい環境を目指すための新機能を追加するなどシステム更新を行ない，人文科学データベースを統合し，一元的・網羅的検索を確保しつつも，多角的に検索結果を分析することが可能なプラットフォームとして進化させ，2012 年 5 月 7 日公開した．

本稿では，nihuINT のシステム概要，検索インターフェースおよび nihuINT における主な機能について述べる．さらに，人名・地名などの知識をベースにさらなる統合を推進しつつある．この構想についても触れる．また今後の展望について述べる．

２．nihuINT 概要

2.1. システム構成

nihuINT のシステム構成を図 1 に示す．

検索クライアントとして機能する GWS（Gateway System），共有化検索サーバとして機能する FES（Front End System）および，図 1 にはないが，nihuINT 全体の運用管理を行う MGR（Manager）より構成される．

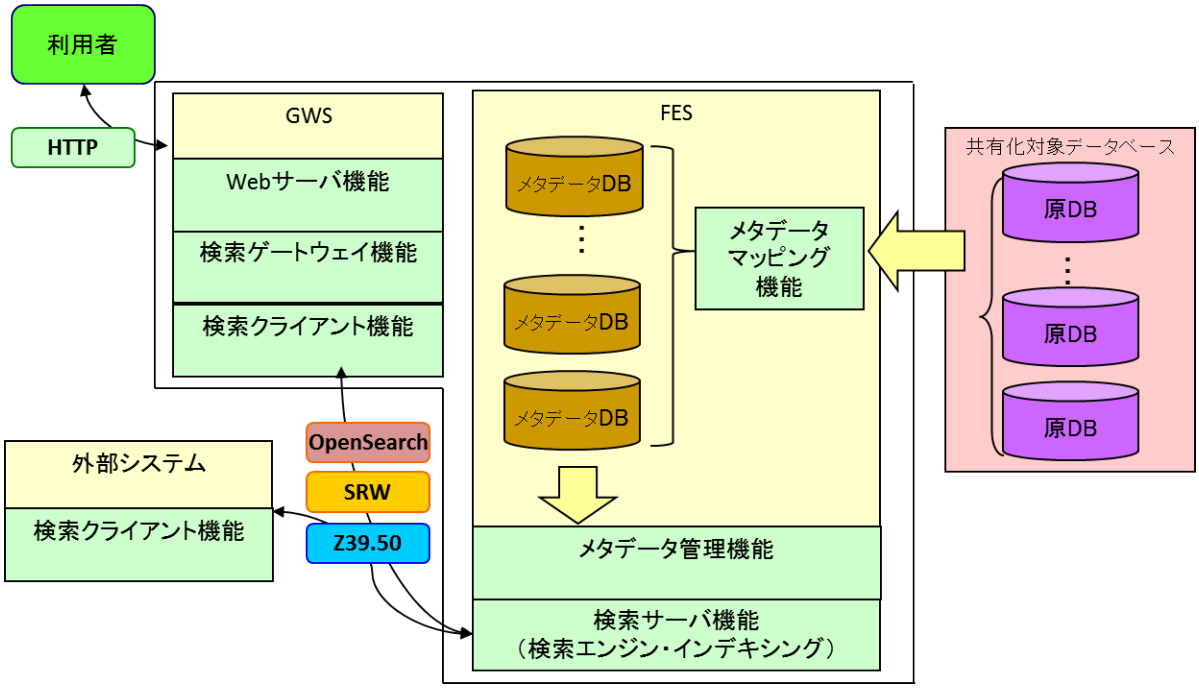

図 1　nihuINT の各機能

GWS は，利用者と対話するためのユーザインターフェースを有し，利用者からの検索リクエストを FES へ伝える．FES はそれ自身に格納してあるデータベースを対象に，GWS からの検索リクエストについての検索処理を実行し，検索結果をレスポンスとして GWS に渡す．GWS は，FES から得られた検索結果を利用者に提示する．

nihuINT に参加している各機関所有のデータベース について，メタデータは FES に登録される．FES は機関ごとに用意しており，基本的には各機関のデータベースはその機関の FES に登録される．FES はそれ自身に格納してあるデータベースを対象に検索処理を実行する．MGR（Manager）は，nihuINT 全体の運用管理を行う．メタデータの更新状態・データベースの公開状態などを把握するなど運用管理を行う．

FES は，共有化対象データベースから抽出したデータからメタデータを生成するメタデータマッピング機能，メタデータを格納管理するメタデータ管理機能，メタデータに対する検索サーバ機能およびメタデータ・データベースを有する．

メタデータマッピング機能は，データベースから抽出したデータの対応する項目を，基本共通メタデータおよび他のメタデータにマッピングし，メタデータ・データベースに格納する．

メタデータ管理機能は，メタデータマッピングの結果として出力されるメタデータと共有化対象データベースの情報を登録・管理する機能である．メタデータは XML 形式で登録するため，XML データベースとして機能する．管理するデータベース情報としては，データベース名，機関名，データベース概要，キーワード，種類，時空間情報の有無，公開設定（公開・試験公開・非公開の区分），データ更新終了時刻などがある．

検索サーバ機能は，GWS や外部システムなどの検索クライアントからの検索リクエストに対し，検索エンジンを用いて検索を行い，検索結果を検索レスポンスとして返戻するサ

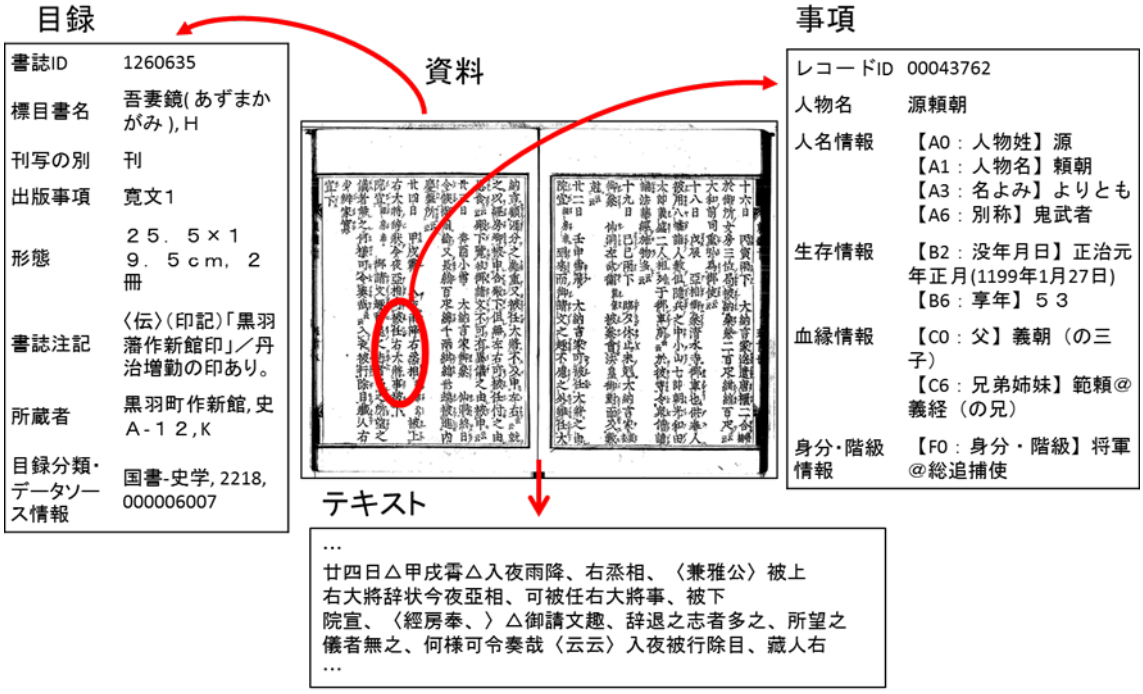

図 2　多様な表現

種類	DB数
所蔵資料目録	39
本文・テキスト	11
画像・映像・音響	9
所蔵図書・雑誌目録	11
研究文献目録	19
事項・ファクト	49

図 3　データベースグループ（種類別）

基本共通メタデータ	
identifier	識別子
title	名称・題名
keywords	種別・主題
who	人物・組織
when	時期・日付
where	地域・場所

図 4　基本共通メタデータ

ービスを提供する．

2.2. 共有化対象データベースとメタデータ

共有化対象データベースの数は 138 にも及ぶ．機関別の内訳を図 3 に示す．この内，地域研究拠点のデータベース nDP は，機構のもとに設置された「地域研究推進センター」における研究成果として作成されている．

ある資料をデータベース化するにあたり，研究内容や目的によって異なる種類のデータベースが作成されることがある．図 2 は nihuINT 内における『吾妻鏡』データの多様性を

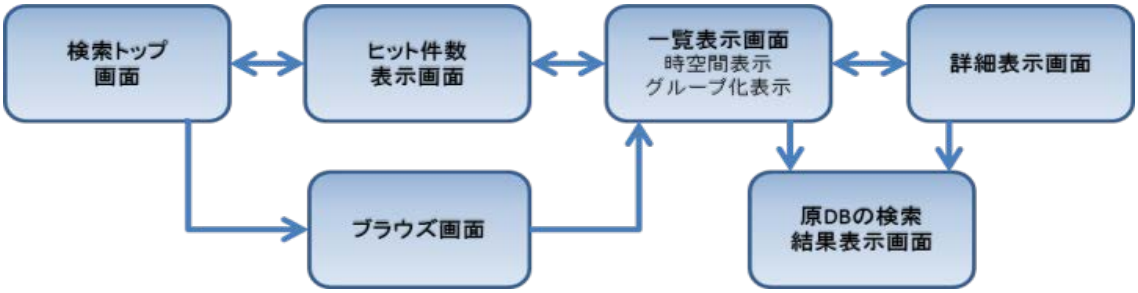

図 5　検索フロー

示す.『日本古典籍総合目録』データベースでは『吾妻鏡』の目録情報,『吾妻鏡』データベースには『吾妻鏡』のテキスト,『地下家伝・芳賀人名辞典』データベースには人名（事項）に関する情報が収載されている．すなわち，同じ資料から目的により，目録化，テキスト化，事項抽出され，異なるデータベースが作成された.

また，各機関が所有するデータベースには，歴史学，日本文学，日本語学，環境学，民族学などのような人文科学での様々な研究分野，それらに付与されているメタデータにおける情報の粒度のばらつきなど，さまざまな点で不均質性がある.

統合検索においては，このような多様性や不均質性のあるデータを一元的に検索し提示しうる環境が必要である．そこで，必要最低限の共通要素を定義したメタデータである基本共通メタデータを用意した．基本共通メタデータの要素は，図 4 に示すとおり，識別子，名称・題名，種別・主題，人物・組織，時期・日付，地域・場所の 6 項目のみであり，多くの人文科学研究資源に共通し，一覧する場合において重要であると思われる要素を端的に表現したものである.

基本共通メタデータ以外にも，共通して保持するメタデータ（共通メタデータ）として，DCMES[1]の基本記述要素で構成するメタデータ（Simple DC），および Simple DC を拡張した汎用メタデータ（NIHU メタデータ[2]）を用意した．Simple DC は Web 上で流通するデータに対するメタデータ記述として標準であるため用意しており，主にシステム外部からの検索リクエストに対応するときに利用する．NIHU メタデータは，Simple DC（ただし，空間的・地理的な記述項目である Coverage.spatial および時間的な記述項目である Coverage.temporal の 2 つの拡張項目を含む），Who，When，Where 等の 5W1H メタデータ，および，時空間情報メタデータを有する構成であり，特に時空間情報について詳細な記述を施すために拡充したものである.

共有化対象データベースには博物館資料に関するデータベースがあるが，Simple DC や NIHU メタデータでは情報の記述が不十分だと考えた．そこで，博物館資料に特化した博物館コアメタデータ[3]も用意した.

2.3. 外部システムとの連携

国立国会図書館は PORTA[4]やそれ以外の国立国会図書館における目録システムを統合し，新たに NDL Search（国立国会図書館サーチ）[5]を 2012 年 1 月 6 日に一般公開した．2010 年 7 月 14 日より nihuINT と PORTA が，相互にデータベースを検索できるよう，シ

ステム連携を開始した．システム連携に関する詳細は[6]に記載されている．NDL Search 公開に伴い，新たに nihuINT と NDL Search との連携を開始した．2011 年 10 月 26 日に

図 6　一覧表示（画面左側は検索結果分類におけるファセット）

NDL Search から nihuINT への検索を，2012 年 1 月 27 日より nihuINT から NDL Search への検索を開始した．

3．利用者インターフェース

3.1. 検索フロー

nihuINT の検索フローを図 5 に示す．検索トップ画面で検索語入力し，データベースごとのヒット件数を表示し，検索結果の一覧を表示し，選択した検索結果の詳細を表示するフローである．

検索語入力では，簡易検索と詳細検索の 2 つを用意した．簡易検索は基本共通メタデータおよび他のメタデータの全項目に対する検索である．詳細検索は基本共通メタデータの項目を指定した検索である．

一覧表示画面では検索結果を表形式（図 6），もしくはスニペット形式で一覧表示する．表形式による一覧では，基本共通メタデータの“名称・題名”，“識別子”に加え，データベース名，機関名が表示される．スニペット形式による一覧では，ヒットしたメタデータの項目名・ヒットした箇所とその前後のテキスト，機関名，およびデータベース名が表示される．詳細表示画面では一覧表示画面で選択した検索結果に対して，設定した共通メタデータの各項目名とその値を表示する．また、両画面とも原データベースの詳細表示画面

（図 7）へのリンクが表示される．

▶ 検索TOP ▶ ヒット 件数 ▶ 一覧表示 ▶ 詳細表示 ▶ オプション ▶ ヘルプ

1件目(検索結果の合計：169件)

次のデータ >>

館蔵錦絵 / 国立歴史民俗博物館 / 博物館コア形式

[原DBレコードを表示]

項目名	内容
名称	東海道五十三次の内 京 二 真柴久吉 Actors at the Fifty-three Stations of the Tokaido: Mashiba Hisayoshi at Kyo2
識別番号	H-22-1-1-39
種別	錦絵コレクション Kabuki, Landscapes
主題	[主題分類]役者絵 名所絵 [内容分類]見立役者絵 東海道 [人名]尾上菊五郎3(真柴久吉＝羽柴秀吉) [その他固有件名]京都東山 鴨川 三条大橋 泥棒
人物	歌川豊国3(歌川国貞1)/豊国画(年玉枠) Utagawa Toyokuni3(Utagawa Kunisada1) [彫師]彫竹 辻岡屋文助 Tsujiokaya Bunsuke
時間	嘉永5年8月 1852/0
空間	[国名]山城 [地名]京都
物理属性	[判型]大判 [法量]36.7×24.7 [形態]縦 [員数]1枚
関連	錦絵コレクション
その他	[画工署名]豊国画(年玉枠) [彩色]錦絵 [改印]子八 衣笠 村田 [備考]背景は歌川広重1の保永堂版「東海道五拾三次之内 京師」の図様を借用。「京 石川五右衛門」(H-22-1-1-66)と続絵をなす。

図 7　詳細表示

3.2. 検索結果分類機能

nihuINT に限らず，多くの検索システムにおいて，利用者は検索する際に検索したい語が思いつかない状況から試行錯誤して検索することが少なくないと考えられる．特に多様性・不均質性の高い nihuINT のデータベースを検索する場合は，明確な検索条件を設定することが困難な場合もある．そこで nihuINT では，検索条件の設定を“ゆるやか”にする代わりに，利用者が必要な検索結果を探すのを補助する機能を提供する．この 1 つが検索結果分類機能である．この検索結果分類はいわゆるファセット型ナビゲーション（Faceted Navigation）[7] として機能する．nihuINT では、基本共通メタデータの各項目をファセット，ファセット値を項目の値としている．図 6 のようにキーワードを“東海道”とした検索を行うと 17,879 件と非常に多くの検索結果を得ることになり，個々を確認していくのは困難である．検索結果のファセット分類により，利用者は必要な検索結果を発見する切り口を得ることができる．

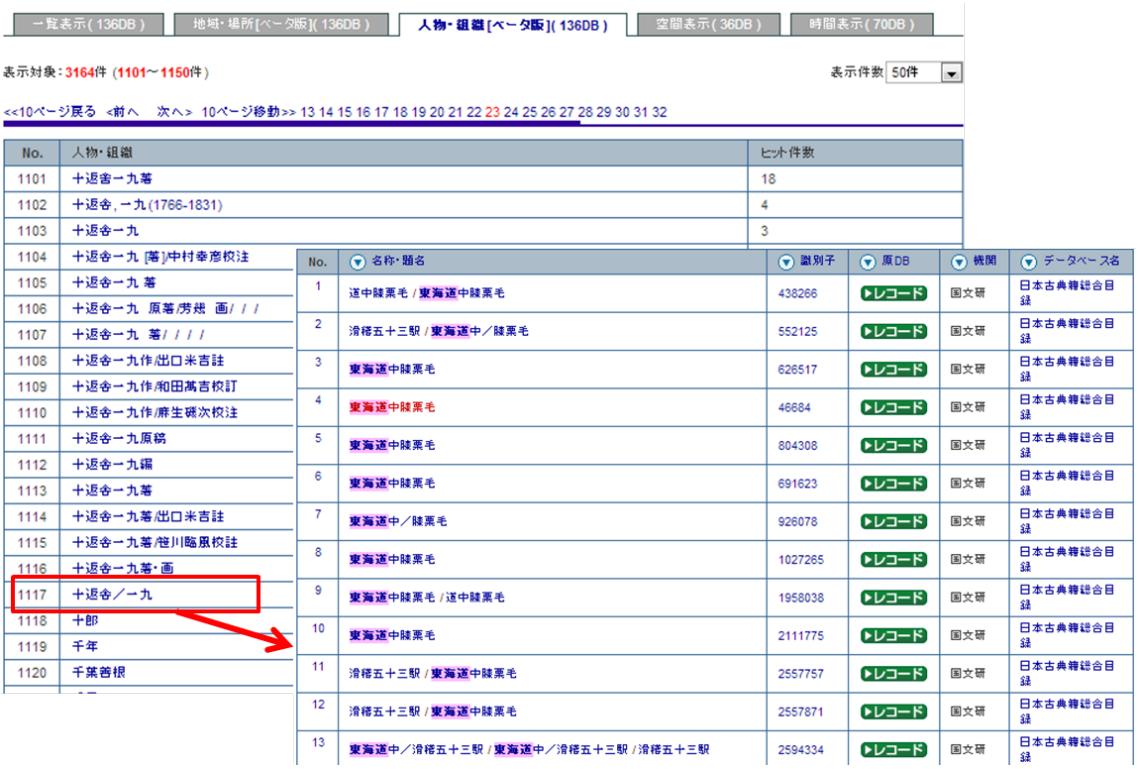

図 8　グループ化表示（人物・組織）

3.3. グループ化表示

この機能は検索結果を“人物・組織”もしくは“地域・場所”を単位にグループ化して表示する．図 8 はキーワードを“東海道”とした検索を行い，“人物・組織”でグループ化した検索結果を示す．この機能はグループ化した人物・組織名とその件数を表示する．あるグループを選択すれば ，それに関連する検索結果が表示される．

3.4.ブラウジング

詳細検索や検索結果分類からの関連情報の検索では，何かしらの検索条件を設定して検索を進める．しかしながら，どのようなデータが存在しているか分からないため，検索条件を設定することが困難な場合がある．nihuINT は，データベース単位にどのようなデータが存在しているかを一覧することができるブラウジング機能を提供する．検索トップ画面から各データベースをブラウズすることができる．ブラウジング機能では識別子以外の基本共通メタデータの項目ごとにどのような項目値があるか，その項目値であるデータは何件あるかを表示している．これにより，どのようなデータがあるかを把握することができ，その後の検索において，利用者による検索条件設定のヒントとなり得ると考えている．

3.5. サジェッション

検索語の入力の支援を行うため，入力途中でも後続する文字列を補完して，利用者が入

力したいキーワードの候補を提示するサジェッション機能を提供する．サジェッションによって入力誤りを防ぐ効果も期待できる．サジェッションで提示するキーワードは過去の検索でヒットしたキーワードであり，キーワードごとにスコアリングし，その値に応じてランキングした結果の上位 10 件を提示する．このスコアは，影響力の逓減モデルにおける忘却係数[8,9]を用いて算出しており，重みを 6 ヶ月で半減するように設定している．

3.6. 異体字同定検索

異体字を吸収した検索結果を提示する検索の方が望ましいことが多々ある．そこで，国立国語研究所が開発した異体字漢字対応テーブルを用いて異体字を同定した検索を実現している．このテーブルは，単漢字のみであるが，``国"と``國"のような新字体・旧字体だけではなく，繁体字・簡体字にも対応している．

3.7. 時間検索・時間表示

検索条件の設定において，時間範囲を指定することができる．直接入力して指定することもできるが，時間範囲指定はタイムラインシステム（TimeLine[10]を利用）を用いて範囲を指定するができる．時間情報をもつ検索結果に対して，タイムラインシステムによる時間表示機能を用いて検索結果を一覧することができる．

データベース中に記述されている時間情報の表記は多様である．その時間情報に対し，その開始と終了を太陽暦による表現で正規化している[11]．特に和暦に対しては，国文学研究資料館相田満氏作成による『日本暦基本対照表』を用いることで正規化している．

４．考察

4.1. 基本共通メタデータと他のメタデータの関係，クレンジング

基本共通メタデータの各項目はデータに“関連する”内容を記述しているに過ぎない．しかしながら，基本共通メタデータの各項目は人文科学における典拠データとして利用できる可能性がある．例えば，グループ化表示機能において，“人物・組織”でグループ化したとき，検索結果として表示されるデータは，“人物・組織”に関する典拠データであり，各“人物・組織”に関連するデータを得ることができる，という見方もできる．どのような関連であるかはデータベースごとに設定したメタデータによって表現されているため，これを述語として用いれば，ある人物の著作物や血縁者などのデータを得ることが可能だと考えられる．これを更に発展させることができれば，人文科学データベースにおけるレファレンスサービスとして展開することも可能だと考えられる．

典拠データベースやレファレンスサービスを実現するためには次の問題がある．検索結果分類機能のファセット，グループ化表示機能の検索結果，およびブラウジング機能では，単に文字列一致するかどうかで各データを区別している．そのため，ごみデータが存在する，同姓同名の人物が区別できない，表記が異なるため別の場所として判断してしまう，

などの問題がある．この解決のためには，メタデータの記述を洗練する，各データを識別する，シソーラス等を用いて名寄せを実施するなどのタスクに取り組む必要がある．また，典拠データ等の表現方法も重要である．Linked Data[12]のようなセマンティック Web を含め，既存の手法が適用できるかどうか，我々のフレームワークの利点を損なうことなく表現できるかどうかについて注意深く検討する必要がある．

4.2. データベースの拡大

人間文化研究機構各機関のデータベースは 200 を越えるがその半数程度しか nihuINT の検索対象となっていない．機構各機関のデータベースをより多く検索対象としていく必要がある．また機構のデータベースだけは人文科学に関する研究資源の一部分しかカバーできない．現在，nihuINT は NDL Search とシステム連携しているが，さらに他機関の外部システムとの連携を進めて，より充実した人文科学データベースの検索ができるようにすることも重要な課題である．

５．おわりに

本論文では，人文科学データベースを一元的かつ網羅的に検索するために導入した基本共通メタデータについて述べた．さらに，統合検索システム nihuINT において基本共通メタデータを用いた検索支援機能を示した．nihuINT が人文科学研究における情報提供の中心的なデータベースシステムとしての役割を担えるよう今後も積極的に発展させていく予定である．

参考文献

[1] Dublin Core Metadata Initiative: Dublin Core Metadata Element Set, Version 1.1 (2010). http://dublincore.org/documents/dces/.

[2] 人間文化研究機構：NIHU メタデータマッピング規則 ver.2.00 (2007). http://www.nihu.jp/sougou/kyoyuka/pdf/reference/03.pdf.

[3] 山本泰則，安達文夫：博物館資料情報統合検索のためのコアメタデータ，情報処理学会シンポジウム論文集，Vol.2009，No.16，pp.287-294 (2009).

[4] 柴田昌樹：PORTA によるデジタルアーカイブの連携について，人間文化研究情報資源共有化研究会報告集 1，pp.123-131 (2010).

[5] 国立国会図書館：国立国会図書館サーチ (2012). http://iss.ndl.go.jp/.

[6] 山本泰則：国立国会図書館 PORTA と人間文化研究機構 統合検索システムとの連携について，人間文化研究情報資源共有化研究会報告集 2，pp.53-68 (2011).

[7] Morville P. and Callender J.: 検索と発見のためのデザイン―エクスペリエンスの未来へ，オライリージャパン，Nov 2010.

[8] Khy, S., Ishikawa, Y. and Kitagawa, H.: Novelty-based Incremental Document

Clustering for On-line Documents, Proceedings of 22nd International Conference on Data Engineering Workshops (ICDEW06), p.40 (2006).

[9] 石川佳治，北川博之：忘却の概念に基づくインクリメンタルな文書クラスタリング手法，電子情報通信学会技術研究報告． DE, データ工学， Vol.101, No.192, pp.145-152 (2001).

[10] Massachusetts Institute of Technology: SIMILE Widgets | Timeline. http://www.simile-widgets.org/timeline/.

[11] 安達文夫：統合検索システムの概要と今後の展開，人間文化研究情報資源共有化研究会報告集 1，pp.33-43 (2010).

[12] Tim Berners-Lee: Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData, 2006.