nihuINT: A Platform for Integrating a Variety of Humanities' Research Resources

Taizo Yamada[1]

1 Historiographical Institute, The University of Tokyo

1. Introduction

The National Institutes for the Humanities ("NIHU") launched nihuINT (Nihu INTegrated retrieval system; an integrated search engine for shared research resources) in April 2008 with the goal of promoting research and education through organic beneficial use of various humanities-related research resources. As of October 2013, nihuINT can search 138 databases. These include 122 DB from the organizations that comprise NIHU (The National Museum of Japanese History, The National Institute of Japanese Literature, The National Institute for Japanese Language and Linguistics, The International Research Center for Japanese Studies, The Research Institute for Humanity and Nature and The National Museum of Ethnology), four from the nDP (nihu Data Provider) DB at the area studies centers operated by NIHU, and 12 from the National Diet Library, which has worked in cooperation with NIHU since July 14, 2010. The number of searches using nihuINT has climbed to approximately 3.5 million a month.

nihuINT enables cross-DB searches and the display of retrieved results using common operations without the need to consider the location or operating procedures of each DB, and in addition to displaying a list of the cross-DB retrieval results, includes functions such as searches utilizing spatiotemporal information. Upgrades to the system have been carried out, including the addition of new features aimed at creating an environment in which research resources can be searched more easily than before, and a platform that integrates the humanities databases and enables retrieval results to be analyzed multidirectionally, while continuing to ensure uniform, comprehensive retrieval, was unveiled on May 7, 2012.

This paper provides an overview of the nihuINT system, and describes the search interface and nihuINT's main functions. In addition, further integration is being pursued based on knowledge such as personal names and place names. This concept is touched upon as well. Finally, a view of the future is described.

2. Overview of nihuINT

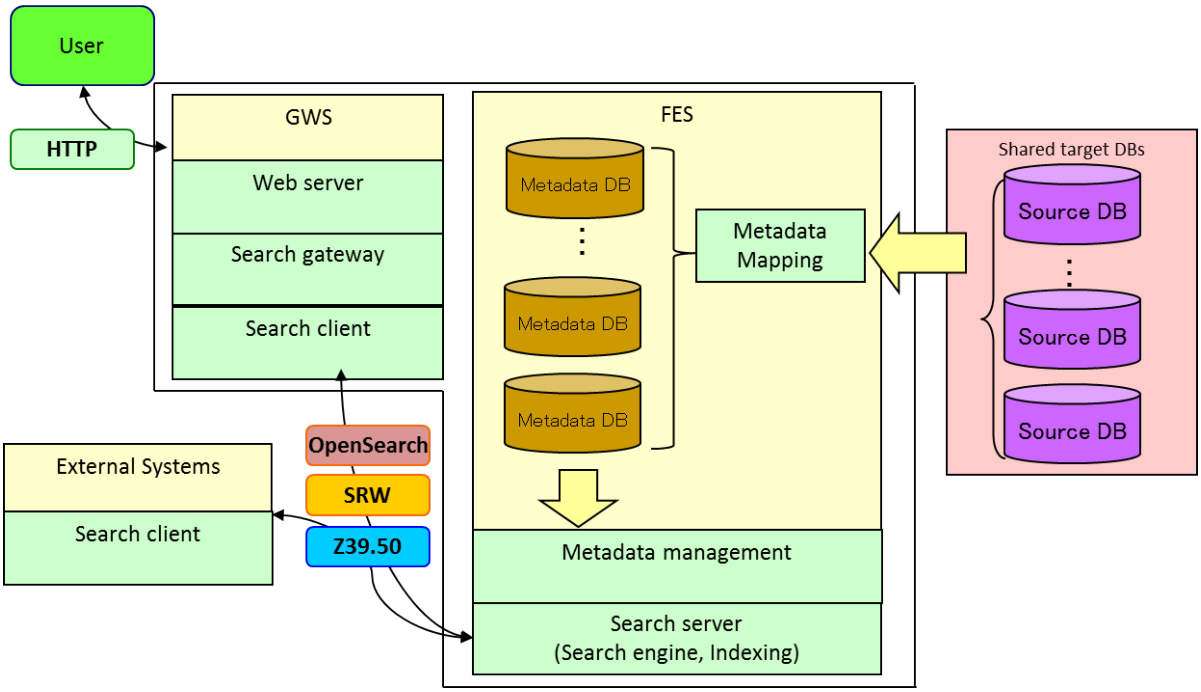

Figure 1 The various nihuINT functions

2.1. System configuration

The system configuration of nihuINT is shown in Figure 1. nihuINT is comprised of a GWS (Gateway System) that functions as a retrieval client, an FES (Front End System) that functions as a shared retrieval server and a MGR (Manager), which is not shown in Figure 1, that handles the operations and management for the entire nihuINT system.

The GWS possesses a user interface for communicating with users, which transmits the retrieval requests from users to the FES. The FES executes retrieval processing of retrieval requests from the GWS by searching the databases stored within it, and returns the retrieval results to the GWS as responses. The GWS displays to users the retrieval results obtained from the FES.

Metadata for the database owned by each institution participating in nihuINT is registered in the FES. The FES is prepared by each organization, and basically the database of each organization is registered in that organization's FES. The FES executes retrieval processing for the databases stored in it. The MGR (Manager) handles the operating management for the entire nihuINT. Operating management performed by the MGR includes understanding various conditions, such as the status of metadata updates and the open status of the databases.

The FES is comprised of a metadata mapping function for forming metadata from data extracted from the databases available for sharing, a metadata management function for storing and controlling the metadata, a retrieval server function for

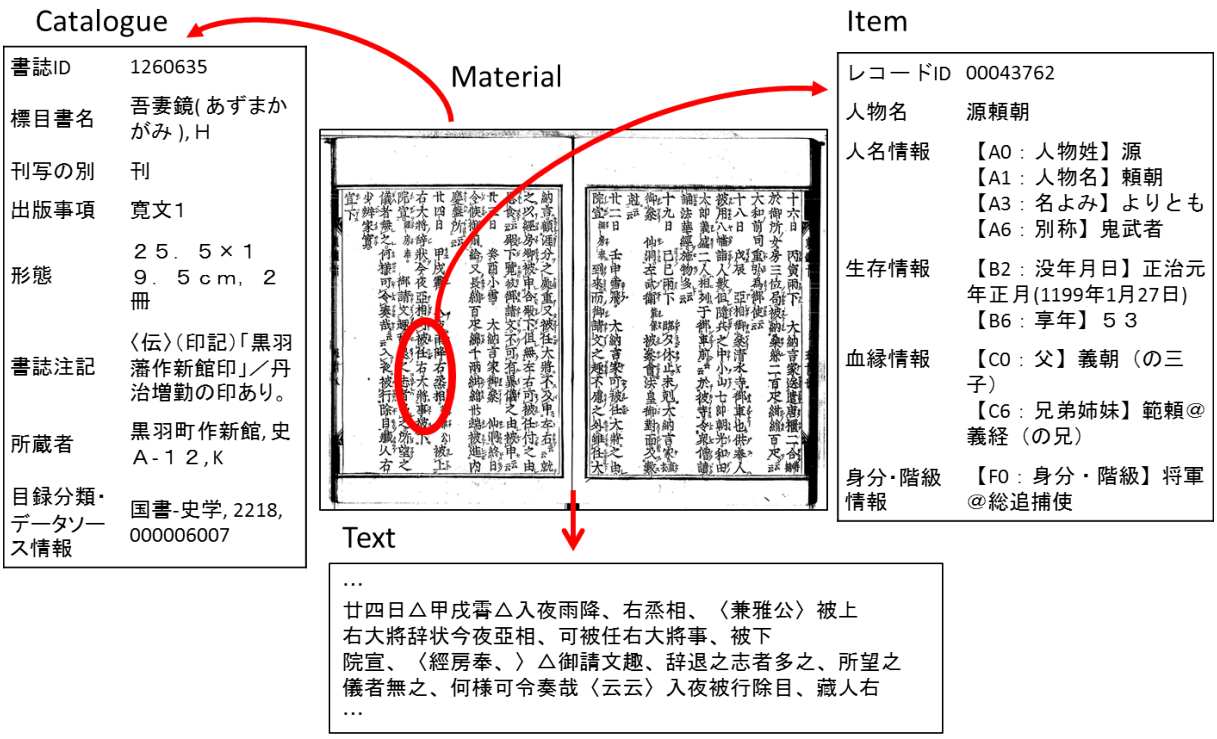

Figure 2　Various renderings

Type	Number of DB
Materials	39
Full Text	11
Images / Audio	9
Libraries	11
Bibliography	19
Facts	49

Figure 3　Database groups (by type)

Basic Common Metadata
identifier
title
keywords
who
when
where

Figure 4　Basic shared metadata

searching the metadata, and the metadata databases.

The metadata mapping function maps the items that correspond to data extracted from the databases into the Basic Common metadata and other metadata, then stores the results in the metadata database.

The metadata management function is a function to record and manage the metadata that is output as a result of the metadata mapping, and the information in the shared databases. Because the metadata is recorded in XML format, it functions as an XML database. The metadata stored as management database information includes the

database name, the organization name, the database summary, key words, type, whether there is spatiotemporal information, the disclosure settings (classification as public, trial publication or not public), and data update finish time.

The retrieval server function provides a service to use the search engine to perform searches in response to retrieval requests from search clients, such as the GWS or an external system, and return the search results as retrieval responses.

2.2. Shared databases and metadata

There are 138 shared databases; a breakdown by organization is shown in Figure 3. Among these, the nDP database of the area studies centers has been created as study results at the Center for Area Studies that was established at NIHU.

In some cases, a different type of database might be created when converting certain resources into a database, depending on the research content or objectives. The diversity of data on the *Azuma Kagami* in nihuINT is shown in Figure 2. Catalogue information for the *Azuma Kagami* is listed in the database for the *Union Catalogue of Early Japanese Books*, the text of the *Azuma Kagami* is listed in the *Azuma Kagami* database, and information concerning person's names (categories) is listed in the *Jige Kaden HagaJiinmei Jiten* database. That is, the same resource has been converted into a catalogue and text, depending on the purpose, and then the particulars were extracted and different databases were created.

Moreover, the databases owned by each organization exhibit a lack of homogeneity in many respects, including both the various fields of research in the humanities, such as history, Japanese literature, Japanese linguistics, environmental studies and ethnology, and disparities in the granularity of information in the metadata contributed by each organization.

Integrated searches require an environment that will enable such diverse and non-homogeneous data to be uniformly searched and displayed. We therefore prepared the Basic Common metadata, which is metadata that defined the necessary minimum common parameters. As shown in Figure 4, the Basic Common metadata parameters are only six items – an identifier, title and keywords and who, when and where – that succinctly expressed those parameters thought to be important when sharing and viewing numerous humanities research resources.

In addition to the Basic Common metadata, metadata comprised of the basic descriptive parameters of DCMES [1] (Simple DC), and general-purpose metadata that expanded the Simple DC (NIHU metadata [2]), were prepared as metadata that is shared and maintained (Common metadata). The Simple DC has been prepared to have a standard as the metadata description for data that circulates on the Internet, and is used

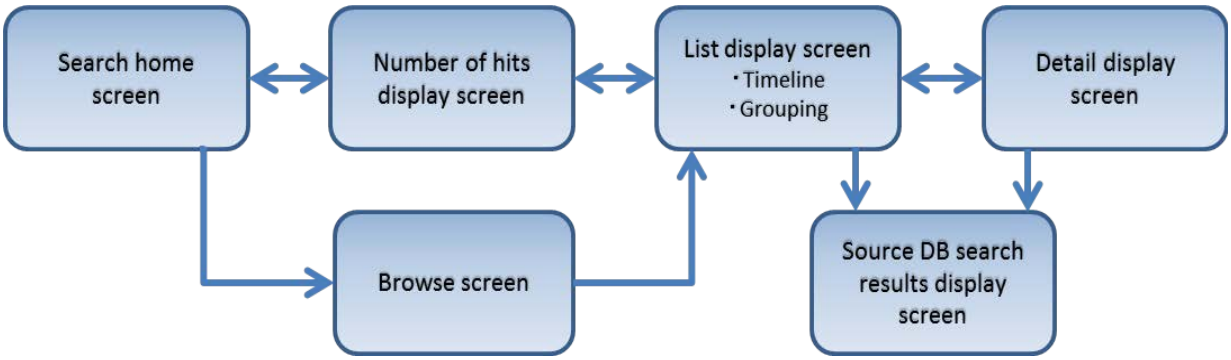

Figure 5　Search flow

mainly when responding to search requests from outside the system. The NIHU metadata is configured to have the Simple DC (but including two enhancing items: Coverage.spatial, the spatial and geographic description items, and Coverage.temporal, the temporal description items) and 5W1H metadata such as who, when and where, and the spatiotemporal information metadata, with the spatiotemporal information in particular expanded to give a detailed description.

The shared databases are databases related to museum resources, but the information descriptions for the Simple DC and NIHU metadata were thought to be inadequate. Consequently Museum Core metadata [3] that are specialized in museum resources also were prepared.

2.3. Linkage with external systems

The National Diet Library integrated PORTA [4] and the other catalogue systems at the National Diet Library, and NDL Search [5] was newly opened for public searches on January 6, 2012. System linkage to enable nihuINT and PORTA to mutually search each other's databases was begun on July 14, 2010. Details concerning the system linkage are described in [6]. In conjunction with opening NDL Search for public use, a new link between nihuINT and NDL Search was launched. Searches of nihuINT from NDL Search were begun on October 26, 2011, and searches of NDL Search from nihuINT were begun on January 27, 2012.

3.　User Interface

3.1. Search flow

The nihuINT search flow is shown in Figure 5. The workflow order begins with input of a search word on the search screen, after which the number of hits for each database is displayed, a list of the search results is displayed, and details of the search results selected are displayed.

Figure 6　List display (left side of screen shows facets for search results Figure 5)

Two search word inputs, enabling a simple retrieval or a detailed retrieval, have been prepared. A simple search examines all of the items in the Basic Common metadata and other metadata. A detailed search performs a search of the specified items in the Basic Common metadata.

On the list display screen, the search results are displayed as a list in either a table format (Figure 6) or in snippet format. For lists using the table format, the Basic Common metadata title and identifier are displayed, together with the database name and name of the organization. For lists using the snippet format, the item name of the metadata that was selected, the location of the hit, and the text before and after that are displayed, together with the organization name and database name. On the detailed display screen, the name of each item of shared metadata that was set and its value are displayed for the retrieval results selected on the list display screen. In addition, the links to both screens and the original database detailed display screen (Figure 7) are displayed.

3.2. Search result classification

It's thought that with many search engines, not only nihuINT, users frequently will search using trial and error when performing a search because it's difficult to recall the word or words that they want to search. In some cases, setting clear search conditions

▶ 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。

Figure 7 Detailed display

might be especially difficult when searching the databases in nihuINT, which are diverse and have a high degree of heterogeneous. Therefore in place of setting the search conditions to "loose," functions are provided with nihuINT to help users look for the required search results. One of these is the search results classification function. This search result classification as so-called Faceted Navigation [7]. In nihuINT, each Basic Common metadata is set as a facet, and the facet value is set as the item's value. As shown in Figure 6, when a search using "Tokaido" ("東海道") as the key word is performed, the search results obtained – 17,879 – is extremely large, making it difficult to confirm each individual result. Using facet classification of search results, users can obtain means to dig into the search results and locate the search results they need.

3.3. Grouping

This function displays the search results by grouping them into units according to "who" or "where." Figure 8 shows the search results when a search was performed using "Tokaido" as the key word and the results grouped by "who." This function displays the

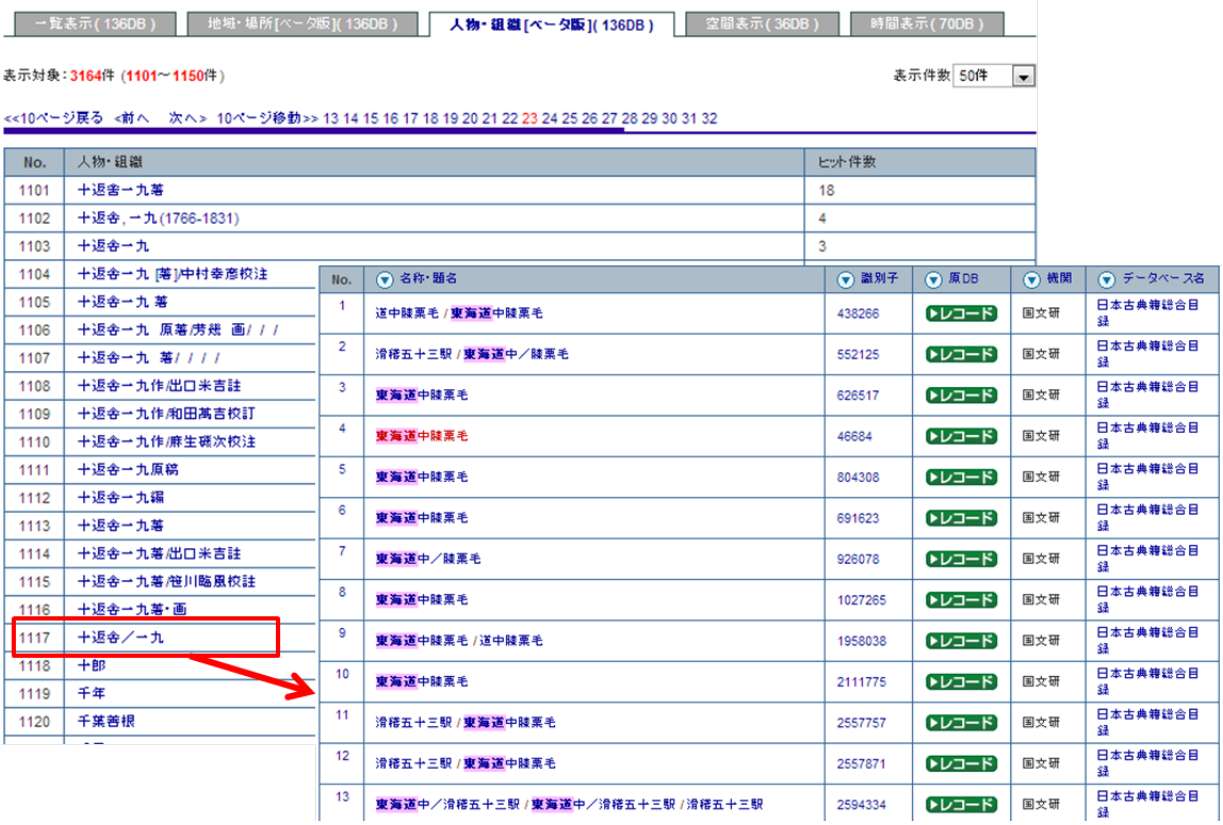

Figure 8　Grouping

“who” names clustered into a group and the number of hits returned. If a certain group is selected, the search results related to that group are displayed.

3.4. Browsing

For a detailed search or search for related information based on search results classifications, a search is performed after some search conditions have been set. Because the type of data that exists is unknown, however, establishing the search conditions is sometimes difficult. NihuINT provides a browsing function that enables users to look at what data are available in each database. Each database can be browsed from the search home page. The browsing function displays what item values are provided for each item in the Basic Common metadata other than the identifier, or how many data have that item value. As a result, users can understand what data are available, which should provide hints for the search conditions set by users for their subsequent searches.

3.5. Suggestions

To support search word input, a suggestion function is provided to supplement a follow-on character string during key word input and display candidates for key words a user might want to enter. This can also be expected to have the effect of preventing input

errors, based on the suggestions. Key words displayed by the suggestions are key words that received hits in past searches; each key word is scored, and the ten highest scoring words as a result of the ranking are displayed according to their values. These scores are calculated using forgetting factors [8, 9] in a successively diminishing influence model, and are set to reduce the weighting by half every six months.

3.6. Variant character identification search

In many cases a search that displays search results that subsume variant characters is desirable. Consequently the table of kanji variants developed by the National Institute for Japanese Language and Linguistics is used to realize searches that have identified variant characters. This table, which contains only individual characters, supports not only a new character forms and former forms for characters, such as "国" and "國," but also traditional Chinese characters and simplified Chinese characters.

3.7. Timeline

Users can specify a time range when setting their search conditions. While users can specify a time by inputting it directly, the time range specification enables them to flag a range using the timeline system (Timeline [10]). For search results with temporal information, users can analyze the search results by using the timeline system.

There is a variety of marks for temporal information described in the database. For that temporal information, the beginning and end are normalized by expression based on the solar calendar [11]. In particular, the Japanese calendar has been normalized by using the "Japanese History Basic Comparison Table" prepared by Mitsuru Aida of the National Institute of Japanese Literature.

4. Discussion

4.1. Relationship between Basic Common metadata and other metadata, and cleansing

Each Basic Common metadata item describes only the contents "related" to the data. However, there is a possibility that each Basic Common metadata item can be used as authority data in the humanities. For example, with the grouped display function, when grouped according to "who" the data displayed as the search result is the authority data concerning "who," and can also be viewed in a way that makes it possible to obtain the data related to each "who." Because the type of relationship is expressed by the metadata established for each database, it is thought that data such as someone's copyrighted works or relatives can be obtained if this is used as a predicate. It's believed that if this can be developed further, it will also be possible to develop this as a reference service in the humanities databases.

In realizing the authority database and reference services the following problems are encountered. For the facets of the search result classification, the grouped search result display function and the browsing function, each data is distinguished merely by whether or not the character strings correspond. As a result there are problems such as the existence of bad data, the inability to distinguish persons who share the same family name or same first name, and locations being judged as different locations because data are worded differently. To resolve these problems, it will be necessary to address tasks such as refining the metadata descriptions, distinguishing each data and implementing name identification using a thesaurus or other tools. Moreover, the mode of expression of the authority data etc. is also critical. This will require carefully studying whether existing methodologies can be applied, including the semantic web techniques such as Linked Data [12], and whether results can be expressed without losing the advantage of our framework.

4.2. Database expansion

Although there are over 200 databases for the NIHU's organizations, only about half of these are searchable using nihuINT. It will be necessary to make more of the databases at each of NIHU's organization available as targets. Moreover, NIHU's databases can cover only part of the humanities-related research resources. While nihuINT currently has system linkage to NDL Search, establishing further links to the external systems of other organizations, and making it possible to search a more enhanced range of humanities databases, are critical issues.

5. Conclusion

This manuscript describes the Basic Common metadata introduced for uniform, comprehensive searches of humanities databases. It also highlights the search support function that utilizes the Basic Common metadata in the integrated search engine nihuINT. We plan to actively develop nihuINT in the future so that it can fulfill its role as a core database system for the provision of information in the humanities.

References

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

[2] National Institutes for the Humanities: NIHU metade-ta mapping kisoku [NIHU Metadata Mapping Rules] (in Japanese), ver. 2.00 (2007). http://www.nihu.jp/sougou/kyoyuka/pdf/reference/03.pdf.

[3] Yamamoto Y. and Adachi F.: Core Metadata to Search for Museum Object Information

across Databases (in Japanese), *Proceedings of Computers and the Humanities Symposium "JinMonCom 2009",* Vol. 2009, No. 16, pp. 287-294 (2009).
[4] Shibata Masaki.: PORTA ni yoru dejitaru a-kaibu no renkei ni tsuite [Digital Archive Linkage using PORTA] (in Japanese), *Proceedings of the Study on Information Resources of the Human Science 1*, pp. 123-131 (2010).
[5] The National Diet Library: NDL Search (2012). http://iss.ndl.go.jp/.
[6] Yamamoto Y.: Kokuritsu kokkai toshokan PORTA to jinken bunka kenkyuu kikou tougou kensaku shisutemu to no renkei ni tsuite [Linkage between The National Diet Library PORTA and the Integrated Search Engine of the National Institutes for the Humanities] (in Japanese), *Proceedings of the Study on Information Resources of the Human Science 2*, pp. 53-68 (2011).
[7] Morville P. and Callender J.: Search Patterns (Design for Discovery), O'Reilly Media, Inc., Jan 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] Ishikawa Y. and Kitagawa H.: Incremental Document Clustering Based on Forgetting Factors (in Japanese), IEICE Technical Report. DE, Data Engineering, Vol. 101, No. 192, pp. 145-152 (2001).
[10] Massachusetts Institute of Technology: SIMILE Widgets | Timeline. http://www.simile-widgets.org/timeline/.
[11] Adachi F.: Tougou kensaku shisutemu no gainen to kongo no tenkai [Integrated Search Engine Overview and Future Development] (in Japanese), *Proceedings of the Study on Information Resources of the Human Science 1*, pp. 33-43 (2010).
[12] Tim Berners-Lee: Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData, 2006.