

利用履歴に基づくデジタルコンテンツへの価値情報付与方式の提案

大森 伸宏 齊藤 典明

NTT 東日本 研究開発センタ

〒180-8585 東京都武蔵野市緑町 3-9-11 NTT 武蔵野研究開発センタ本館 4F

{n.oomori,saito.n}@rdc.east.ntt.co.jp

あらまし：

電子図書館では、収集・選書機関としての立場により様々な情報を分類整理し、より質の高い情報を提供する役割が期待されている。しかしながら、従来の人手による情報の分類整理には熟練を要するため、情報の大量・多様化に伴い支障が出始めている。本稿では、デジタルコンテンツに対する分類整理の判断基準となる価値を表現する指標の導入について述べる。提案する手法により、大量のコンテンツ中から自動的に等価値と判断される集合が抽出できる。よって、この集合を扱うことにより雑多なコンテンツ中からの良質なコンテンツが容易に抽出可能となり、コンテンツの利用が促進される。

Digital Contents Categorizing Method Based on Access Histories

Nobuhiro OMORI Noriaki SAITO

NTT EAST Research and Development Center

3-9-11 Midori-Cho Musashino-Shi Tokyo 180-8585 Japan

{n.oomori,saito.n}@rdc.east.ntt.co.jp

Abstract:

Digital Libraries are required to supply high quality digital contents like traditional libraries. But recently a Librarian is unable to classify contents, because contents on digital libraries have increased enormously in quantity and variety. In this paper, for the solution to this problem, we propose the value information of digital contents based on access histories. Since with this information we can search a group of contents of the same value automatically, high quality contents from a variety of contents can be used easily.

1 はじめに

電子図書館では、テキスト、音声、動画、音楽といった様々なデジタルコンテンツを蓄積、提供している。ここで、電子図書館の特徴として、各図書館が独自の価値判断にて情報を収集している点や、情報に付随する二次情報（＝メタデータ）を有効活用している点が挙げられる。

しかし近年の情報の量や種類の増大により図書館側からの人手による分類整理が難しくなり、利用者が望む情報を探し出すことが困難となってきている現状がある。そこで、コンテンツの持つメタデータに着目し、新たに価値情報を加えることにより、コンテンツのグループ化を行う手法を提案する。

以下、2章では電子図書館に関する話題を基に問題点を抽出し、3章でその解決策として提案する手法を述べる。さらに4章にてシステム化を目指した実現方法を提案し、5章でその効果について論じる。

2 背景

2.1 電子図書館のコンテンツ

電子図書館で扱う情報資源は、コンテンツそのものである一次情報、およびコンテンツの属性情報である二次情報（＝メタデータ）に分けられる。この一次情報、および二次情報をコンピュータによって検索し、参照することによって大量のコンテンツの中から目的のコンテンツに容易にアクセスすることが可能となる。一方で、大量のコンテンツの中から効率的に目的のコンテンツを得るためにには、単純に情報を電子化し、検索するだけではなく電子図書館固有の情報発見手法の実現が必要であると考えられている。このような目的に対し、従来メタデータを用いて自動的なコンテンツ間の関連付けを行うことによりコンテンツの発見を容易にする研究[1]や、蓄積や取り出しの単位の細分

化という手法[2]などがあった。特に、電子図書館においては、章や節ごとに情報を取り出す、文章中の図のみを取り出すといった、一冊の書物の任意の部分（以下、これを取り扱い単位とよぶ）を取り出すことが要望されている。

2.2 ゲートウェイ図書館

従来の図書館では、情報の「媒体」を独自の判断基準で収集、分類整理、蓄積、提供してきた。しかし、電子図書館においては全ての情報はコンピュータ上のデータとしてネットワークを介して相互に利用可能なことから、全ての電子図書館が、情報の実体を持つ必要がなくなった。このことから、ネットワーク上の情報資源に対する一元的なアクセスの提供が要望されており、この機能を持つ電子図書館をゲートウェイ図書館と呼ぶ。

特に、ネットワーク上の全ての情報資源へのアクセスを提供するのではなく、対象をある分野に特定し、専門スタッフにより厳選されたメタデータをデータベースに持つサブジェクト・ゲートウェイと呼ばれるサービスあるいはプロジェクトもまたいくつか存在している[3]。

2.3 次世代電子図書館システム NetLibra

著者らは上記の要求条件を満たすために、コンテンツへのメタデータの付与、書籍単位・記事単位での検索、ネットワークを介した他電子図書館への横断検索、を実現した次世代電子図書館システム NetLibra を開発し、運用実験を行ってきた[4]。

2.4 問題点

NetLibra 運用実験より、現在の電子図書館が抱えている情報資源へのアクセスに対する問題点を以下に整理する。

- 1) コンテンツの量の増大
- 2) コンテンツの種類の増大
- 3) コンテンツ選定作業の増大

1) コンテンツの細分化および、他図書館へのアクセスが可能であることによるコンテンツの量が増大し、欲しい情報があるのにその情報に辿り着けない問題が起きる。2) 各図書館のそれぞれの判断

基準で集められた情報が混在する事によりコンテンツの種類が増大し、ある情報の存在に気づいてもその情報の持つ価値に気が付かない、という問題が起こる。3) ゲートウェイ図書館を実現するためには、人手によるコンテンツの選定作業の増大が発生し、システムによる自動化が望まれている。

2.5 コンテンツの価値

従来、図書館の蔵書（コンテンツ）は、各図書館の司書による判断基準に基づいて、価値あるコンテンツが選書され、図書館のコレクション（ある目的を持って収集された蔵書の集まり）として収集されてきた。その価値判断の基準はそれぞれの図書館により異なり、例えば大学図書館であれば学術資料を中心に収集し、公共図書館であれば地域に密着した古文書や実用書を中心に収集している。

このように、コンテンツの価値は判断をする人、場所、状況により変化するものであり、数値等で直接的に表現出来ないものである。一方、前述の通り、図書館においてはある個人文庫に関するコレクション、地域に関連する地誌等のコレクション、また歴史資料集といったコレクションのように様々なコレクションが形成されてきている。そして、このコレクションを形成することこそが、利用者に対して各コンテンツの価値付け（意味付け）をする役割を担っていると言える。つまり、図書館ではいくつかのコンテンツを集める事によって、それらのコンテンツの持つ価値を際立たせる役割を持っている。

これより、コンテンツを複数個集める事により、これらのコンテンツの持つ価値を表現することが可能になると考えられる。

3 提案する手法

3.1 コンテンツのグループ化

コンテンツの量と種類の増大という 2.4 節の問題点 1)、2)を解決するために、著者らはコンテン

ツの取り扱い単位に着目した。細分化され多様な価値を持つコンテンツは、著者や編集者が作成したリンクや、前述の様なシステムが自動的に作成したリンク[1]により関連付けられている。このリンクを用いる事により、コンテンツはひとつの集合として扱うことが出来るが、そのリンクは複雑に張り巡らされ、コンテンツの集合は無限に大きくなる可能性がある。そのため、情報を効率的に発見するために行われたコンテンツの細分化が、逆に雑多なコンテンツからなる巨大なコンテンツの固まりを形成する事になり、利用者が求めるコンテンツを発見することを難しくしている。

そこで、これらのコンテンツの固まりの中から利用者の求めるコンテンツを提供するためには、取り扱い単位を最適化する必要がある。この取り扱い単位は、従来の図書館が持つコレクションの単位に相当するものであり、利用者がこの取扱単位を基にコンテンツの価値を判断できるものでなければならぬ。

つまり、従来の図書館におけるコレクションのように共通の価値を持ったコンテンツをグループ化することとし、これをもってコンテンツ集合体（Digital Contents Set）と呼ぶ事とする（図 1）。

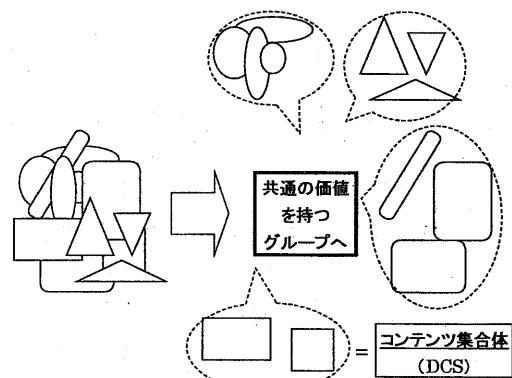


図1 コンテンツのグループ化

3.2 グループ化の指標

共通の価値を持ったコンテンツをグループに分けるには、各コンテンツに価値を表わす指標を導入する必要がある。しかし前述の通り、価値とは

各個人・場所等によって異なるものであり、コンテンツの価値の直接的な記述は非常に難しい。そこで、本研究では「連続して利用されたコンテンツは同じ価値を持っている」と仮定した。つまり、リンクにより関連付けられているコンテンツの集まりを考え、利用者がこのリンクを辿り、その中のいくつかのコンテンツを利用した時、これらのコンテンツはある価値判断の基で利用されたと捉える事とする。

以上のことから、価値を表わす指標としてコンテンツ間の利用履歴を利用し、これをもってコンテンツの価値情報（Digital Contents Value）とする（図2）。さらに、この価値情報を用いてコンテンツ間の価値の合致の度合いを判定し、コンテンツ集合体を形成する。

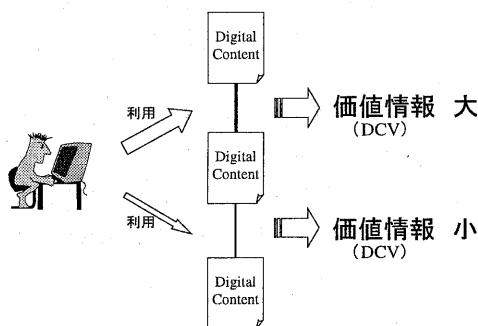


図2 価値情報としての利用履歴

この利用履歴は利用者の参照動作を自動的に記録することにより得られるため、この利用履歴を利用した動的なグループ化を行う事が可能となる。このため、2.4 節の問題点 3)に挙げた人手による作業の増大を防ぐ事が可能となる。同時に、場所、時間によって利用履歴は変化するため、価値の動的な変化に対応できる。

3.3 メタデータツリーの導入

価値情報としてのコンテンツ間リンクに対する利用履歴を管理する方法として、本稿では従来手法であるログファイルによる一括管理ではなく、

各コンテンツが持つメタデータに着目し、メタデータを用いた個別管理を提案する。メタデータの一つの項目として価値情報を管理する事で、

- 1) 他のメタデータと同様に、コンテンツの効率的利用のため、コンテンツ自身とは独立して検索等に利用される。
- 2) 作者等の他のメタデータと連携した価値判断が可能である。
- 3) コンテンツを図書館間で移動する際にも、他のメタデータと共に移動できる。

等の利点がある。この 1)の利点により、数多くのコンテンツ同士が互いにリンクでつながっているコンテンツ空間において（図3-①）も、コンテンツ間の利用履歴を価値情報とし（図3-②）メタデータにて管理する事により、メタデータのみを利用したコンテンツのグループ化（図3-③）が可能となる。

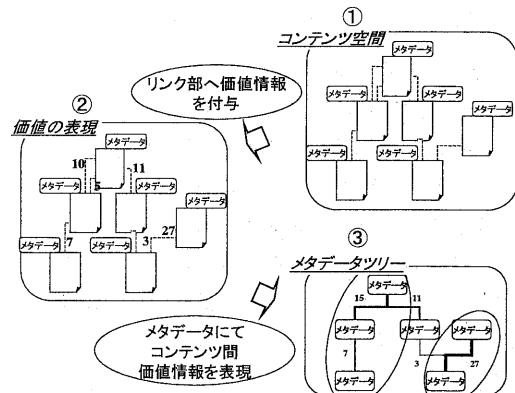


図3 メタデータツリーの生成

つまり、価値情報（利用履歴）を含んだメタデータだけを管理すれば良いことになる。これらのデータはメタデータをノードとするツリー構造を取るため、これをメタデータツリー（MetaData Tree）と呼ぶこととする。また、このようにこれらの価値情報をメタデータにより管理する事により、メタデータのコンテンツ本体（一次情報）からの分離が可能であり、コンテンツを持たずに、複数図書館へのアクセスの入り口となるゲートウェイ図書館に適用可能である。

以上の機能を実現するためには、メタデータツリーを蓄積するデータベース（メタデータツリーDB）、および利用履歴を収集しこのデータベースに追加する利用者履歴参照機能、またデータベースの情報を基にコンテンツのグループ化を行うグルーピング機能が必要となる。これらの実現方法について以下で検討する。

4 実現方式の提案

4.1 現在の NetLibra

本研究でベースとして用いた NetLibra のシステム概要を図 4 に示す。図の様に 3箇所に分散配置された 5つの電子図書館における運用実験を行って来た。各電子図書館は電子図書館本体 (DL)、その入り口となり他図書館へのアクセスも可能とするゲートウェイ (GW)、ネットワーク上のサービスを案内するトレーダ、から構成されており、各図書館間は ATM ネットワークにより接続されている。また、利用者はこのネットワーク上のどこからでもすべての電子図書館にアクセスできる。この分散環境下でのシステムを実現するために

NetLibra のプラットフォームには CORBA (Common Object Request Broker Architecture) [5] を用いており、この CORBA 上に検索サービス、一覧サービス、参照サービスが実装されている。ここでは、各サービスから、コンテンツ情報の一次情報、およびメタデータが蓄積されたコンテンツ DB にアクセスし、その結果を利用者クライアント上のブラウザに表示させることができる。

コンテンツの一次情報として、静止画 (JPEG、GIF 等)、テキスト (TEXT)、PDF 等各種文書処理アプリケーションに対応する形式、および動画像 (AVI、MPEG 等) などがある。また、メタデータとしては、書誌情報の標準化動向を考慮し、Dublin Core Metadata (RFC2413) にて規定されている 15 項目 (表 1) [6][7] の他に、サブタイトル、著作権処理時の半開示の有無、電子透かしに関する各種情報および著作権クリアランス費用などを追加し、全部で 40 以上の項目によりコンテンツを管理している。

4.2 実現方式

本研究を実現するにあたって、この NetLibra 上に、グルーピングサービス、および利用者履歴収

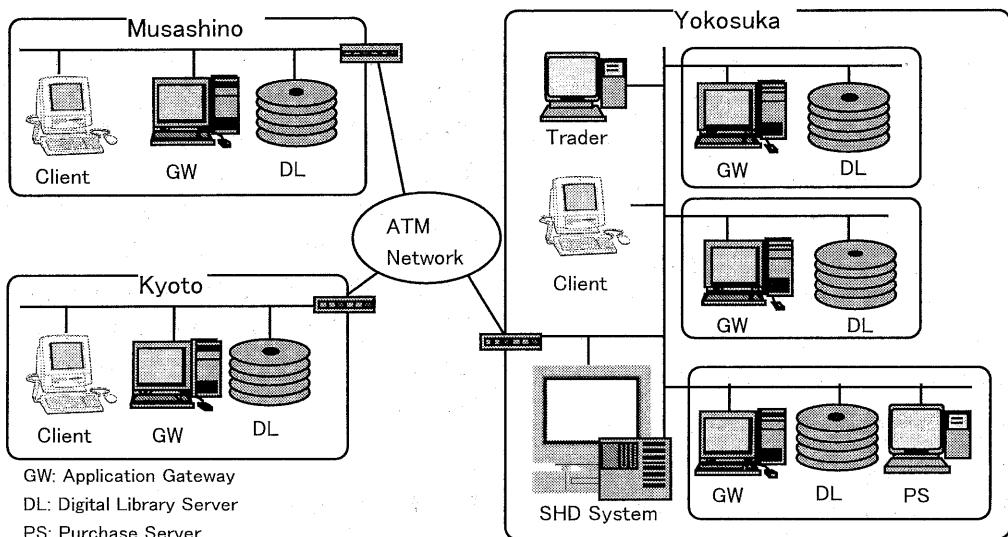


図4 NetLibra 実験システム構成図

集サービス、ならびにこれら両サービスからアクセス可能であるメタデータツリーDBの3つの機能を実装する(図5)。ここで、グルーピングサービスはNetLibraにおける検索サービスの上位に位置し、また利用者履歴収集サービスも同じくNetLibraの一覧サービスおよび参照サービスの上位に位置することで、従来のNetLibraサービスへのアクセスと同様のアクセス手段での利用が可能となっている。

それぞれについて、以下で詳しく説明する。

表1 Dublin Core Metadata(RFC2413)

Element	Label
1. Title	Title
2. Author or Creator	Creator
3. Subject and Keywords	Subject
4. Description	Description
5. Publisher	Publisher
6. Other Contributor	Contributor
7. Date	Date
8. Resource Type	Type
9. Format	Format
10. Resource Identifier	Identifier
11. Source	Source
12. Language	Language
13. Relation	Relation
14. Coverage	Coverage
15. Rights Management	Rights

4.2.1 グルーピングサービス

グルーピングサービスは、1)検索要求に応じメタデータツリーDBにアクセスし、検索キーに合致するメタデータツリーを切りだす機能、2)切り出されたメタデータツリーを基にコンテンツのグルーピングを行う機能、の2つの機能から成る。また同時に通常の検索サービスでのコンテンツDBへの検索を行い、メタデータツリーDB内に存在しないコンテンツも提示することが可能である。

4.2.2 利用者履歴収集サービス

また、利用者履歴収集サービスは、利用者のコンテンツ参照要求の際呼び出され、メタデータツリーDBへ利用情報を登録すると同時に、通常の参照サービスを呼び出し、その結果を利用者に提示する。

4.2.3 メタデータツリーデータベース

メタデータツリーDBには、コンテンツの各種メタデータを格納する。メタデータには、関連付けられているコンテンツへのリンク情報および、そのリンク間にに対する参照履歴が含まれており、これを価値情報としてコンテンツのグルーピングに用いる。また、他のメタデータはNetLibraが持つDublin Core Metadataを基にした書名、作者等の書誌情報であるが、今後データ項目が変更されたり、他アーカイブシステムとの連携をする場

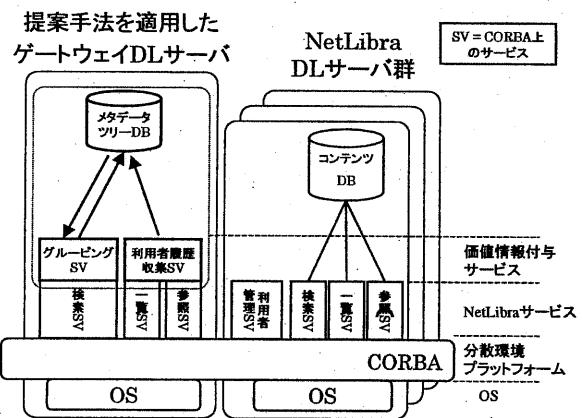


図5 NetLibraサービス構成図

合には、各コンテンツのメタデータ項目数が異なる可能性がある。さらに、複数人からなる団体が著者となるように、各メタデータは階層構造をとる場合がある。

以上の理由により、メタデータツリーDBの実装において、柔軟なデータ項目の設定および、データの階層化表現が可能なXML技術を適応する事とする。さらにXMLをDBにおいて管理するために、より適しているオブジェクト指向DBを用いる事とする。

5 本手法の効果

コンテンツの検索時に本提案手法を適用すると、図6の様に従来は検索結果の一覧表示であったも

のが、等価値の集合体として検索結果が表示される。ここでは、その効果について述べる。

5.1 コンテンツ集合体の発見

コンテンツが集合体として検出されることで、利用者にとってそれぞれ独立した存在として既知であるコンテンツが、ひとつの意味を成す集まりである事が発見されうる。例えば今、「ドコモ」、「携帯」、「OCN」、「116」、「ポケベル」、「ISDN」、「iナンバー」、「シティホン」、「インターネット」、等の情報が蓄積されていたとすると、このときNTTに関する様々な話題についてそれぞれ既知であった場合にも、その中の「ドコモ」、「携帯」、「ポケベル」、「シティホン」が一つの価値を持つ集まりであることが新たに認識される。このため、コンテンツを既知の価値観とは異なった価値観で利用することが可能となり、一つのコンテンツの利用範囲が広がりコンテンツの有効利用が促進される。

5.2 集合体内でのコンテンツの価値発見

コンテンツが集合体として検出されることで、

その集合体内に限定してコンテンツを利用する事が可能である。よって、既知のコンテンツ集合体内でのいくつかの未知のコンテンツに対し、他のコンテンツのもつ価値を利用し、容易にその価値を見出すことが可能となる。例えば、「ISDN」、「116」、「インターネット」、「OCN」については一つの価値を持つ集まりであると認識している時に、その集合に未知の情報「i ナンバー」がある場合にも、その価値を容易に想像することが可能になる。

5.3 検索の効率化

また、集合化されることにより検索結果の項目数も集合体の数にまで減少し、集合体の選択後に目的のコンテンツへ段階的に辿って行く事が可能であり、コンテンツにたどり着きやすくなるという利点もある。例えば、図6では従来は13個のコンテンツが羅列され表示されており、欲しいコンテンツを探す手間があったが、本手法によるとまずは、3つの入り口から入り、目的のコンテンツを発見しやすくなっている。

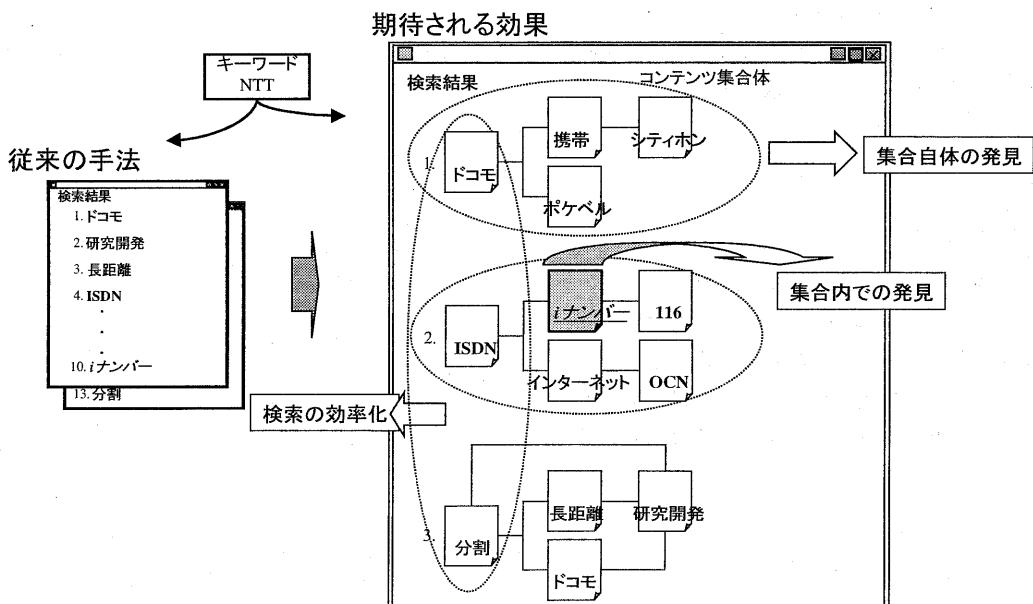


図6 期待される効果

このように、コンテンツ集合を利用単位としてることで、その価値の認識を容易にし、良質のコンテンツの利用促進が期待できる。

6 まとめ

電子図書館におけるコンテンツの量および種類の増大により、求める情報に辿り着けない、コンテンツの持つ意味に気がつかない、という問題に着目した。その解決手法として、コンテンツは集合化されることによってその価値をより際立たせられることを利用した。本稿では特に、価値を表現する指標としてコンテンツ間の利用履歴を利用し、その指標をメタデータの一つとしてコンテンツに付与する事を提案した。そして、与えられたメタデータを用いてコンテンツを集合体として扱うことにより問題の解決を試みた。

今後の課題としては、「連続して利用されたコンテンツは同じ価値を持っている」と仮定した価値が妥当であるかどうかの検証、および、集合体の大きさに対する評価がある。また、コンテンツ集合体の階層化を行い、多段階絞り込み検索の実現についても検討する予定である。さらに、3.3節の2)および3)の利点を利用したコンテンツのグループ化についても考察の余地がある。

参考文献

- [1] 日高、斎藤、阿部、関：情報組織化によるコンテンツ流通システム：Net-X、情報処理学会GW研究会、32-2、1999.5
- [2] 長尾 真：電子図書館の正しい概念を持とう、IPSJ Magazine Vol.40 No.3 Mar.1999
- [3] 尾城 孝一：理工学系ネットワーク情報資源へのゲートウェイ、
<http://www.ne.jp/asahi/coffee/house/ARG/compass-013.html>
- [4] 萩野 忠、西野 正和、関 良明、爰川 知宏：分散ネットワーキング電子図書館 NetLibra の提案、第57回情報処理学会全国大会、3-i、1998.10
- [5] Object Management Group: "The Common Object Request Broker, Architecture and Specification", CORBA Version 2.0, 1995
- [6] <http://purl.org/dc>
- [7] S.Weibel,J.Kunze,C.Langoze,M.Wolf : Dublin Core Metadata for Resource Discovery,RFC2413,1998

謝辞

本検討を進めるにあたって、有意義なコメントを頂いた、東日本電信電話株式会社の長島雅夫氏、日本電信電話株式会社の藤木直人氏、その他ディスカッションに参加して下さった東日本電信電話株式会社研究開発センタの多くの皆様に感謝致します。