図書館情報大学ディジタル図書館のメタデータ作成

平岡　博
図書館情報大学
〒305 茨城県つくば市春日 1-2
Tel: 0298-59-1220, Fax: 0298-59-1212, E-Mail: liru@ulis.ac.jp

概要

図書館情報大学ディジタル図書館 (ULIS-DL) の中心となるサービスは、図書館情報学に関するネットワーク上の情報資源のメタデータを作成し、この分野のサブジェクトゲートウェイ機能を提供することである。本稿では、実際にメタデータ作成に関わった立場から現状と問題点について述べる。なお、ULIS-DL は"http://lib.ulis.ac.jp/"でアクセスできる。

キーワード

ディジタル図書館、電子図書館、図書館情報大学、ダブリン・コア・メタデータ・エレメント・セット、サブジェクト・ゲートウェイ

Metadata Creation at The Digital Library of the University of Library and Information Science

[Abstract]

The major functions of the Digital Library of the University of Library and Information Science (ULIS-DL) are to build and provide a collection of metadata of network resources for libraries and library and information science, and to work as a subject gateway to those resources. This paper describes the present state and lessons learned through our services for creating, maintaining and retrieving the metadata. ULIS-DL is accessible at "http://lib.ulis.ac.jp/".

[Keywords]

Digital Library, University of Library and Information Science, Dublin Core Metadata Element Set, Subject Gateway

1. はじめに

1999年2月に運用を開始した図書館情報大学のディジタル図書館システム(ULIS-DL)は図書館情報学に関連した分野のサブジェクトゲートウェイの機能を提供することが中心的なサービスとなっている。すなわち、図書館情報学や情報メディア研究及び図書館、情報センターなどに関するネットワーク上の情報資源を収集し、そのメタデータを作成して提供している。

収集する情報資源の範囲は次のとおりである。

(1) 図書館情報学および情報メディア研究関係の研究者が発信している研究情報

(2) 国公私立大学図書館、国公私立図書館等が提供している情報

(3) 各種情報センター等が提供している情報

(4) 図書館情報学および情報メディア研究関連学会および団体が提供している情報

(5) 図書館情報学および情報メディア研究関連企業が提供している情報

(6) その他図書館・情報センター等に関連した情報

また、ネットワーク上の情報資源以外にも、一部の雑誌記事や本システム上に登録される一次資料などもメタデータの作成対象としている。

2. メタデータの構成

作成するメタデータはDublin Core(ver1.0)を基本としている。Dublin Coreの15エレメントに出版国、文字コード及びメタデータID(管理用のユニークな番号)の3エレメントを追加した以下の18エレメントで構成されている。

(1) タイトル（Title）

情報資源に与えられた名前：html文書では<TITLE>タグの内容を入力する。<TITLE>タグが無い、空値、内容が不適切などの場合は適当な値を補記する。

(2) 著者あるいは作者（Creator）

情報資源の創造に責任を持つ人あるいは機関。

(3) 主題およびキーワード（Subject）

情報資源の主題や内容を説明するキーワード：主題を表す言葉のほか、重要な語句や頻繁に出てきた単語を入力する。

(4) 内容記述（Description）

情報資源の内容に関する説明記述、抄録：トップページについてはそのページだけでなく、サイト全体の概要を記述する。

(5) 公開者（出版者）（Publisher）

情報資源を現在の形態で利用可能にしたことに責任を持つもの：「著者あるいは作者」と異なる場合に入力する。

(6) 寄与者（他の関与者）（Contributor）

「著者あるいは作者」以外で、当該情報資源をの創造に知的に重要な寄与をしたもの（編集者、翻訳者、イラストレータなど)。

(7) 日付（Date）

情報資源が作成された、あるいは有効になった日付：Web 文書の場合は最終更新日を入力する。

(8) 資源タイプ（Type）

情報資源の種類：Dublin Core の Working Draft を参考に、「text」、「image」、「sound」などを入力。（参考 http://purl.org/dc/documents/wd-typelist.htm）

(9) 形式（フォーマット）（Format）

情報資源のデータフォーマット：http://www.isi.edu/in-notes/iana/assignments/media-types/media-types を参考に、「text/html」のように記述する。

(10) 資源識別子（Identifier）

当該情報資源を一意に識別するための文字列もしくは番号（URI、ISBN など）：文書の URL を入力する。

(11) 情報源（出処）（Source）

当該情報資源を作り出す元になった別の情報資源に関する情報：当面、Relation エレメントを使用し、Source エレメントは使用していない。

(12) 言語（Language）

情報資源の知的内容を記述するために用いられている言語。

(13) 関係（Relation）

別の情報資源の識別子および当該情報資源とその情報資源との間の関係：関係のタイプ（IsPartOf など）と関連する情報資源をスペースで区切って記述する。（参考 http://purl.org/dc/documents/working_drafts/wd-guide-current.htm）

(14) 対象範囲（空間的・時間的）（Coverage）

当該情報資源の知的内容に関する空間的 (地理的) あるいは時間的特性：内容が特定の地域、時代などを扱っている場合に記述する。

(15) 権利管理（Rights）

権利管理に関する声明文、権利管理に関する声明文へのリンクを表す識別子、あるいは当該情報資源の権利管理に関する情報を提供するサービスへのリンクを表す識別子：主に著作権に関する記述があれば入力する。

(16) 文字コード（Charcode）

情報資源の内容記述に用いられている文字セット。

(17) 出版国（Country）

情報資源の出版国。

(18) メタデータ ID（metaid）

メタデータに与えられるユニークな番号：システムで自動的に付与される。

(1) から (15) までが Dublin Core で規定されているエレメントで、(16) から (18) が本システムで規定したものである。

3. メタデータ作成における現状と問題

3.1 情報資源の収集とメタファイルの作成

ネットワーク上の情報資源は自動収集ソフトウェア（ロボット）によって収集される。あらかじめ用意したリストに含まれるURLを起点として、そこからリンクをたどって同一ドメインに含まれる文書を収集してくる。現在用意しているリストは図書館を中心に国内600、海外3000のサイトを含んでいる。

収集した文書からタイトル、日付、文字コードなどの情報や文書の内容を解析してキーワードや内容記述を自動的に抽出して、文書毎の自動抽出メタデータを作成する。このように自動的に作成されたメタデータをメタファイルと呼んでいる。

たとえば、TitleエレメントはHTMLの<TITLE>タグから抽出される。Dateは収集文書のサーバーから取得される文書作成・更新日を、IdentifierはURLをそのまま取り込む。TypeとFormatは文書の種類を識別し、Languageも記述言語を識別することにより値が与えられる。また、著者自身によりHTMLの<META>タグでメタデータが付与されている場合には、そこから抽出した値がそれぞれのエレメントに設定される。

著者によりメタデータが付与されている場合には、SubjectやDescriptionも含めて、ほとんどのエレメントがかなり正確に自動抽出されることになる。<META>タグを含むHTML文書も増えており、中にはDublin Coreのメタデータを与えられているものもあるが、全体としては、まだそれ程多くない。一方、文書の本文から自動生成されたSubjectやDescriptionは不完全なものが多く、人間による修正が必要であり、大幅な省力化や完全な自動化はしばらく先のことになりそうである。

3.2 メタデータ作成作業

次に、自動的に作成されたメタファイルを呼び出して、メタデータ作成者がメタデータを作成する。この時点でメタファイルは仮登録中という状態のメタデータとなる。自動抽出されたエレメントの値を確認し、必要に応じて修正や値の追加を行う。作業は収集文書とメタデータをそれぞれ画面上に表示しながら入力を行う。作成し終わったメタデータは登録申請を行うことにより、登録申請中という状態になる。管理者は登録申請がなされたメタデータを確認し、問題がなければ登録承認を行う。承認されたメタデータは検索システムへ送られて、検索が可能になる。

メタデータの作成単位は個々のWebページであるため、メタファイルのデータはページ単位にリストされている。これを順次呼び出して編集するのであるが、あるWebページをいきなり開いてもそれだけでは対象文書の前後関係（リンク関係）がわかりづらい。そのページの内容だけでなく、そのサイトで提供されているWebページ全体の中でどのような位置にあるのかがわからなければ、メタデータを作成できない場合も多い。このため、文書の関係を調べるのにある程度の時間を費やすことになる。現在は担当者ごとにサイトを分担して同じ担当者が同一サイトのメタデータを作成するなどして効率化を図っているが、収集先のサイトマップを自動作成するような仕組みができないものかと考えているところである。

メタデータの作成作業を担当しているのは、図書館情報大学や筑波大学の学生、卒業生などのパートタイマーが中心である。全員が図書館情報学を専攻しているわけではなく、専攻していても目録作成の経験者ではない。主として、Creator、Subject、Descriptionなど自動生成が困難なエレメントを中心に入力することになる。Subjectはフリーキーワードで文書中の重要と思われる単語や、対象分野、図書館のホームページの場合には図書館の種類などを入力してもらっている。図書館情報学に馴染みの無い担当者は用語の使い方などが難しいようである。Descriptionは文書内容の簡単な説明を付けてもらっている。単純なWebページであればほとんど問題はないが、記事や論文のような場合や、外国語で書かれているものには

苦心しているようである。ただし、外国語の場合には、韓国や中国からの留学生も参加しているため、それぞれの母国語を担当してもらうことにより助かっている面もある。

3.3 メタデータ作成基準

メタデータ作成のための基準としては、Dublin Core の Working Draft などが参考になるが、個々のエレメントの記述のための情報源をどこに求めるかといった問題や具体的な記述方法などは、作成対象となる情報資源の性質によっても異なると考えられる。従来、図書館で作成してきた図書や雑誌などの目録の場合には既に目録規則が存在しており、これに従って作成すればよい。メタデータの場合には作成基準の整備から始めなければならないが、実際にデータを作成しながら入力基準を考えているといった状態である。

とはいえ、図書館で作成している蔵書目録もメタデータの一種であり、そのノウハウを活用することである程度は解決できる部分もある。メタデータの各エレメントは目録データの項目と対応づけることが可能であるので、これまで図書館で扱ってきた資料を対象にする限りはそれほど大きな問題は発生しないだろうと考えられる。

一方、ディジタル資料は近年になって急速に発展し、新たな媒体が次々と出現している。目録規則もこれに対応できるように改定が続けられているが、実際の目録作成において十分な経験を積むまでに至っていないのではないだろうか。特に、メタデータの作成対象の大きな部分を占めるネットワーク上の情報資源は、メタタグによりキーワードや抄録まで与えられているものから、HTML の<TITLE>タグさえ付いていないものまで、その差が大きく、エレメント記述のための安定した情報が得にくいことが問題であるように思われる。

3.4 その他

メタデータの作成単位は個々の Web ページやファイルなどとしているが、雑誌などをイメージデータに変換した場合などは、一つの記事が複数の画像ファイルにまたがったり、一つの画像ファイルに複数の記事が含まれるなど、必ずしも１：１の関係にならないことがある。このような場合のメタデータをどのように作成するか検討の必要がある。

また、Web ページのメタデータについては、トップページも最下層のページも区別なく作成している。このため、サイトごとなどの大きな単位で検索したい場合などには不都合である。Web ページの種類のような情報を与える方法を工夫する必要がある。

この他にも、細かい事柄を挙げればきりがないが、主に経験不足から様々な問題が発生していると思われる。しかし、これらのほとんどは今後、徐々に解決していくことができると考えている。

4. おわりに

図書館情報大学のディジタル図書館は、メタデータの作成と提供という、国内ではあまり例のないサービスを先駆けて開始した。そのため、様々な面で戸惑うことも多い。しかし、メタデータやゲートウェイに対する関心は高まっており、図書館情報学を専門とする大学にふさわしいテーマであると考えられる。ここでの経験が多くの方々の参考になれば幸いである。

参考文献

[1] 杉本重雄, 平岡博, 阪口哲男, 田畑孝一 "図書館情報大学におけるディジタル図書館システム". ディジタル図書館. No.15 pp.17-28.

[2] 平岡博, 真中孝行, 横山敏秋, 阪口哲男, 杉本重雄, 田畑孝一 "図書館情報大学ディジタル図書館システム". 情報管理. Vol.42, No.6, pp.471-479.

[3] Dublin Core 参照記述. http://www.DL.ulis.ac.jp/DC/.

[4] Dublin Core Metadata Initiative. http://purl.org/DC/.