タイ語と日本語による Dublin Core: 世界共通のメタデータセマンティクスを管理する (Dublin Core in Thai and Japanese: Managing Universal Metadata Semantics)

Thomas Baker
Asian Institute of Technology
Bangkok, Thailand
Stuart Weibel
Online Computer Library Center
Dublin, Ohio
(訳： 杉本重雄, 図書館情報大学)

概要

図書館や博物館、企業、あるいは学術分野といった情報資源を扱うコミュニティは、これまで個別に情報を蓄積し利用してきた。ところが、こうしたコミュニティがインターネット上の情報を共有するようになるに従い、既存のあるいは今後作り出される数多くの様々な情報資源に関する記述能力を持つ単一かつ共通のメタデータの記述方式が必要とされるようになってきた。World Wide Web コンソーシアムが開発を進めている Resource Description Framework はこうした要求を満たそうとするものである。Dublin Core はインターネット上に提供されるいろいろな分野の情報資源のためのメタデータを記述するための共通の記述モデルとして最も開発の進んだものである。Dublin Core のセマンティクスは世界共通であるので、英語であれ、タイ語であれ、あるいは日本語であれメタデータの表現は可能である。一方、数多くの言語毎の記述形式を互いに矛盾しないように維持管理していくことは困難であると考えられる。これまで Dublin Core は、いわば混成語 (Pidgin Language) －異なる母言語を持つ人たちによって自然に作り出された単純化された共通言語－として成長してきた。しかしながら、Dublin Core の将来を考えると、進化と成長のための柔軟性を失うことなく、かつ利用分野の広がりによる多様性をうまくコントロールしていくことが必要であると思われる。そのため、メタデータを表すためのことばの意味を取り決めていくための場として Interlingua（言語とは独立に表した概念のスーパーセット）となる分散レジストリ (distributed registory) が必要である。

キーワード

Dublin Core, メタデータ, World Wide Web, Resource Description Framework (RDF), 多言語, オントロジ, 人工言語, 混成語と混合語（Pidgin と Creole）

Keywords

Dublin Core, metadata, World Wide Web, Resource Description Framework (RDF), multilinguality, ontologies, artificial languages, pidgins and creoles

1. 情報資源の記述コミュニティとインターネット上の情報資源の共有地 (Resource Description Communities and the Internet Commons)

インターネットが発展する以前は、情報資源の記述のために組織毎に異なるメタデータ基準が採用されていても大きな問題にはならなかった。たとえば、図書館で本を探したり、地図室で地図を探したり、博物館で文化財を探したりというように、資料や情報を探す人は一時にはひとつの場所（建物）にしか行くことができないので、行った場所にカードカタログや索引があれば、仮にそれらが利用者に馴染みのない方法で組織化されていたとしても、図書館員、地図専門家、学芸委員といった人に助けを求めることができる。

現在、インターネット上に提供されるコレクションやカタログが世界的に共有される情報空間を形作り、いわばインターネットを情報資源の共有地 (Internet Commons) と呼んでもよい状況が生まれている。しかし、民族的に複雑に入りくんだ土地を訪れた旅行者がいくつもの見知らぬことばに戸惑うのと同じように、この共有地を訪れた人は自然言語や計算機用言語によって表された情報資源や検索方法等に関するとても理解しづらい記述に出くわすことになる。情報資源の記述を行ってきた伝統的なコミュニティと並んで、首尾一貫した記述形式を持たないアドホックなコミュニティ、たとえば電子商取引の業者のようなコミュニティが多く生まれ、そこでは要求に適した書き方が手探りで作り出されている。

さらに好ましくないことには、インターネット上で情報を探しながらいろいろな情報資源にアクセスする利用者、いわばインターネット上を旅する旅行者（ヴァーチャルツーリスト）は必要な情報を探す手伝いをしてくれる通訳やガイドを簡単に探すことができない。こうした情報資源のためのいろいろな記述の間での自動翻訳の開発を進める努力がなされるとしても、記述モデルの種類がどんどん増えるので翻訳はますます難しくなる。これまで経験したことのないような大きな情報空間の広がりの中で、首尾一貫した情報資源記述を行い、そうして作り出された記述を交換することができるようにするには、世界共通のセマンティクスとシンタックスを持つ情報資源の記述方法を確立することが必要である。こうした背景の下、メタデーターすなわち、データに関する記述のための構造化されたデーターがWebの基盤開発における主要な問題と認められるようになった。

情報の共有地としてのインターネットにおいて、様々な情報資源の記述モデルがあることは分野間の境を越えて情報を探す上での障害となる。一方、記述モデルがたくさんあるということは分野やコミュニティによって記述に関する要求が多様であることを意味している。詳細なレベルでの記述の場合、記述対象が異なるのでこうしたメタデータの記述モデルは異なっている。たとえば、文書の記述者にとって「雲で覆われた範囲」を書く必要などほとんどないが、農地の衛星写真の場合には非常に重要な記述項目となり得る。多様な情報資源を見渡すと、より一般化されたレベルの記述の場合、ほとんどの情報資源に共通な、すなわち分野に関わらない基本的な属性の記述からなるメタデータの基本要素（コアセット、Core Set) を見出すことができる。このように異なった記述モデルから共通要素として取り出せるものは、見方を変えると、別々の記述モデルとして発展してきたがために意味は同じで単に名前だけが異なっていると考えることができる。そのため、分野に関わらずに共通の属性としてとらえることのできる基本的要素をまとめることによって、分野にまたがった情報資源の検索をより効率よく行うことができるようになると考えられる。たとえば、「著者 (author)」と「作者 (creator)」は情報資源の発見という目的では同一の属性の記述とすることができる。

Dublin Coreは、インターネットのような巨大な情報空間から分野によらずに情報資源を見つけ出すという要求にこたえるために開発されてきた。Dublin Coreは電子的情報資源の記述のために、作者 (Creator)、タイトル (Title)、出版者 (Publisher) 、主題 (Subject)、内容に関する記述 (Description)、他の関与者 (Other Contributors) 、日付 (Date)、情報資源の型 (Resource Type)、形式 (Format)、情報資源の識別子 (Resource Identifier)、情報資源のソース (Source)、情報資源の記述言語 (Language) 、他の関連する情報資源との関係 (Relation)、地理的・空間的範囲 (Coverage)、権利管理 (Rights Management) の15項目を定めている。

こうした要素のほとんどは、図書館の目録カードに書いてあるもののように、一般的な共通理解を得られるものである。Dublin Core は、いわば情報の共有地としてのインターネット上で不慣れな分野の情報を得ようとするヴァーチャルツーリストのために用意された簡単な会話のための慣用表現集 (Phrase Book) のようなものであるといってよいであろう。マシンレベルのプロトコルが、異なったハードウェア間での相互運用性 (interoperability) を保証するために必要とされるのと同じように、情報処理における最も重要なプラットフォーム、すなわち情報を理解し扱う利用者の間で情報の意味的な相互利用を促進するために、データ内容に関するいくつもの標準の意味的な定義の間での共通性を見つけ出すことが有益であると考えられる。

1.1 多言語による Dublin Core

Dublin Core はもともと英語で定義され、開発が進められてきた。しかし原理的には、その 15 項目のカテゴリーは、現代の言語であればどのような言語でも表現することは可能である。もし技術的用語に適切な訳が存在しなければ句として説明することも、あるいは新しい語を作ることも可能である。Dublin Core はこれまでにドイツ語 [14]、タイ語 (図 1 参照)[15]、フィンランド語、スウェーデン語、ノルウェー語、デンマーク語、ハンガリー語、フランス語、ポルトガル語、それに日本語に訳されている。

TITLE ชื่อเรื่อง

ชื่อของทรัพยากรสารสนเทศที่กำหนดโดยเจ้าของผลงาน หรือ สำนักพิมพ์

AUTHOR OR CREATOR ผู้แต่ง หรือ เจ้าของผลงาน

บุคคล หรือหน่วยงานที่รับผิดชอบเนื้อหาเชิงปัญญาของสารสนเทศ

SUBJECT OR KEYWORDS หัวเรื่อง หรือ คำสำคัญ

หัวข้อ คำสำคัญ วลี รหัสวิชา เลขหมู่ ที่อธิบายเรื่องและเนื้อหา

DESCRIPTION ลักษณะ

รายละเอียดของสารสนเทศ เช่น บทคัดย่อ (กรณีที่เป็นเอกสาร)

หรือ บรรยายรูปร่าง ลักษณะการใช้งาน (กรณีที่เป็นวัตถุ)

PUBLISHER สำนักพิมพ์

หน่วยงานที่ผลิตสารสนเทศชิ้นที่เผยแพร่ในรูปแบบปัจจุบัน(อิเล็กทรอนิกส์) เช่น สำนักพิมพ์ มหาวิทยาลัย

บริษัท เป็นต้น

OTHER CONTRIBUTORS ผู้ร่วมงาน

บุคคล หรือ หน่วยงานอื่นนอกจากผู้แต่งหรือเจ้าของผลงานที่มีชื่อปรากฏในข้อผู้แต่ง

หมายถึงบุคคลหรือหน่วยงานที่มีส่วนร่วมสร้างผลงานเชิงปัญญาในระดับรองจากผู้แต่ง

DATE ปี

ปีที่ผลิตผลงานในรูปแบบปัจจุบัน(อิเล็กทรอนิกส์)

RESOURCE TYPE ประเภท

ประเภทของสารสนเทศ เช่น home page นวนิยาย คำประพันธ์ ร่างบทความ บทความ

รายงานวิชาการ เรียงความ พจนานุกรม (จะมีรายการให้เลือก)

FORMAT รูปแบบ

รูปแบบที่บันทึกสารสนเทศ เช่น text/html, ASCII, Postscript file,

โปรแกรมที่นำไปใช้งานได้ , JPRG image (จะมีรายการให้เลือก เช่น registered

Internet Media Types (MIME types)

RESOURCE IDENTIFIER รหัส

สัญญลักษณ์ หรือเลข ที่ระบุเฉพาะว่าหมายถึงสารสนเทศอิเล็กทรอนิกส์รายการนั้น ๆ เช่น URLและ URN

SOURCE ต้นฉบับ

ผลงานที่เป็นที่มาของสารสนเทศ ไม่ว่าจะเป็นเอกสารหรืออยู่ในรูปอิเล็กทรอนิกส์

LANGUAGE ภาษา

ภาษาที่ใช้ในการเรียบเรียงสารสนเทศ

RELATION เรื่องที่เกี่ยวข้อง

สารสนเทศเรื่องอื่น ๆที่เกี่ยวข้อง (ยังไม่มีข้อยุติว่าจะกำหนดคำจำกัดความว่าอย่างไร)

COVERAGE สถานที่และเวลา

(ยังไม่มีข้อยุติว่าจะกำหนดคำจำกัดความว่าอย่างไร)

RIGHTS MANAGEMENT สิทธิ

ประกาศระเบียบปฏิบัติเรื่องลิขสิทธิ หรือ อาจให้Serverจัดการเพื่อให้ผู้ใช้สารสนเทศรับทราบและยอมรับข้อปฏิบัติเรื่องลิขสิทธิทันทีที่สารสนเทศเรื่องนั้น ๆปรากฏบนจอภาพ

図 1: DC-Simple, defined in Thai

こうしたいくつもの言語への翻訳は英語で表された標準形の単なる訳であると思われるかもしれない。実際、多くの図書館における標準はそのように決められてきた。たとえば、Universal Standard Bibliographic Description (USBD) は多くの言語に翻訳されている。また、言語には依存しない数値表現によって多くの言語に対する普遍性を備えることを目的として作られている Universal Decimal Classification (UDC) や Dewey Decimal Classification (DDC) もまた数多くの言語に翻訳されている。たとえば、DDC は 30 の言語に翻訳され、135 カ国で利用されている。しかしながら、こうしたシステムは、新しい知識が生み出さ

れるのに合わせて更新され続けなければならない。そして、現実問題として、多くの場合英語で表された標準を変更し、各国語への翻訳はそこから時間的に遅れて作り出されることになる。

本論文では、Dublin Core を多言語に適用するのにこれまでのようなモデルが不要であることを議論する。いろいろな言語によって具現化をされた Dublin Core をある地域で実現された単なる標準版からの単なる翻訳（すなわち、サブタイトル付きの Dublin Core）ととらえるのではなく、取り決めと改訂作業からなる標準の策定プロセスにおける (英語版のものと) 同等の参加者としてとらえる。以下の節では、はじめに背景を示す。

2. Web のための単純なメタデータ

Dublin Core は、目録作りの方法を学んだことのない著者やウェッブ管理者 (Webmaster) が自分のドキュメントに自身でメタデータを付加することができるようにすることを目的としたものであり、また作成されたメタデータは Web Harvester と呼ばれるデータ収集ソフトウェアや検索システムが利用することを目的としたものである。そのため、非専門家であっても十分に単純であるように意図されている。Dublin Core はより多くの情報を記述することを目的とする既存のメタデータのモデルに取って代わることを意図して作られたものではなく、目録専門家であれ素人であれいずれもが単純な情報資源記述のために利用できるメタデータ記述要素の基本セット（コアセット）として定義されたものである。

しかしながら、目録のエキスパートにとっては、より詳細な記述構造を付加することやより詳細なレベルの意味的記述をすることが可能なように Dublin Core が十分な柔軟性を持つことも重要である。図書館の予算縮小、通貨の変動、熟練した目録専門家の不足、そして世界規模での情報量の増大に際し、図書館で広く使われ、かつより洗練された目録基準である AACR2 や MARC に対しても、Dublin Core はメタデータとして経済的な選択肢のひとつである。実際、Dublin Core は WWW 文書のためのメタデータとして将来にわたって利用される適切な技術であるとの見方もある。

初期の頃から Dublin Core の開発に参加している人たちのコミュニティは大きく分けて二つのグループに別れる。ひとつは Minimalist と呼ばれる人たちのグループである。このグループの立場は、Dublin Core メタデータは単純であることが望ましく、かつ記述条件をできるだけ与えないというものである。したがって、15 項目の要素の記述は基本的に構造を持たないテキストとし、外部で定義された記述に依存することやより詳細な構造を要素内部に持ち込まないことを主張するものである。もう一方のグループは Structuralist と呼ばれる人たちである。このグループは基本要素に付加的な情報や構造を与えることが適切かつ有用であるという立場である。また、場合によっては Dublin Core を特定の分野に応用するにはそうした付加情報が必要であるとするものである。たとえば、作者 (Creator) の要素に与えられた名前が著者ではなく作曲家であるということを限定したり、主題 (Subject) の要素が Library of Congress の Subject Heading に基づいて記述されていることを指定したりすべきであると主張する立場である。実際には多くの人たちがこの両者の中間に位置しており、簡明さの重要性を認めており、複雑な構造を持ち込むことによって得られる利益が明らかな場合にのみ簡明さを犠牲にすることができると考えている。

2.1 実際的な利用の方法

Dublin Core をウェッブ上で利用する最も簡単な方法は HTML の META タグを用いて記述することである。バージョン 2.0 以降の HTML を用いる場合には慣習的に簡単な方法が用いられている。バージョン 4.0 ではより詳細な構造を表すための属性を指示をする qualifier を記述する属性を META タグの中に指定することができる。HTML 文書に埋め込まれたメタデータは文書の一部であり、Web 上の文書を収集し、索引付けをするソフトウェアによって直接収集される。しかしながら、情報資源が分散しているのでメタ

データの更新と維持を非常に難しいものにしている。たとえば、ある文書上での変更や修正がそのコピーに伝わらず、矛盾を生じることにつながることがある。

また、埋め込み型とは別の方法として、Dublin Core で記述したメタデータのレコードを、記述対象の文書とは別に蓄積、維持することも可能である。これは図書館や博物館で目録や索引を作るのと同じである。たとえば、利用者の年令に対して適切な資料であるかどうかや利用に関する適合条件の評価付けといったサービスを行う組織によって提供される新しい種類のメタデータの場合はこうしたメタデータだけで蓄積されることになるであろう。

第 3 の方法は Dublin Core をデータベースの不均一なコレクションへの窓として利用することである。Dublin Core 以外のいくつかのメタデータ基準で構成される不均一なコレクションの場合、そうした他のメタデータ基準から Dublin Core への写像を実現することで、不均一なコレクションを統合的に検索することが可能である。こうした写像を実現するため、「Crosswalk」と呼ぶ Dublin Core と他のメタデータ基準との対応関係定義がなされてきている。たとえば、(図書館向けの)MARC[2]、(政府情報向けの) GILS との間の Crosswalk が作られている。また、Z39.50 の profile との間の Crosswalk により、Dublin Core を用いて Z39.50 サーバに対して検索質問を発することができるようになる [11]。こうした写像のコレクションは英国の Michael Day によって維持管理されている [3]。

2.2 モジュール性のためのシンタックス

ウェッブ上のメタデータに対する要求は非常に多様である。そのため、別個に開発され維持管理されているひとまとまりのメタデータ (メタデータパッケージ) が共存できる環境が必要とされている。Dublin Core は情報資源の発見 (Resource Discovery) のために設計されたものであるが、他の目的を指向した別の機能を持つメタデータパッケージが用いられる場合も多いであろう。たとえば、利用条件 (Terms and Conditions) パッケージは情報資源の権利保持者の識別、価格表示、再利用や出版に関する制約条件の指定等のために利用されると考えられる。こうした異なる基準に基づいて作られるメタデータをひとまとまりのものとして扱うこと、すなわちメタデータにおけるモジュール性の必要性が Warwick で開催されたワークショップで認識され、Warwick Framework として形式化されたことによって、Dublin Core は大きく進化した。

Warwick Framework として形成された概念は WWW のために進められていたメタデータ開発にも大きな影響を及ぼした。WWW の標準化を進める組織である World Wide Web コンソーシアム (W3C) の下で進められているメタデータの基本概念が Resource Description Framework (RDF) として確立された [18]。このメタデータの構成方式によると、たとえば図書館の目録、第 3 者機関による内容評価 (Rating)、電子商取引等、多様な種類の構造化された情報を表現することができる。これが実現されると、いろいろな別個の専門分野の組織によって独立的に作られたいろいろなメタデータの共存が可能になる。より重要なことは、RDF の実現によっていろいろなメタデータ基準で書かれたメタデータのための Plug-and-Play(導入するだけですぐに利用できる) 環境が提供されるようになり、利用者の必要性に適合した記述的メタデータ (descriptive metadata) を利用することが容易になることであると考えられる。

2.3 標準化への過程

Dublin Core の開発はメタデータを構成する基本要素を決めることであると言える。1995 年 3 月以来開催されてきた一連の Dublin Core ワークショップには図書館、コンピュータネットワークとディジタル図書館、さらにいろいろなコンテンツ専門家が集まり、単純な記述レコードとして実現されるメタデータの基本要素に与えられる意味に関する国際的なコンセンサスを得るための議論を重ねてきた。1997 年 10 月に開催されたヘルシンキでのワークショップにおいて、正式にこの目標が達成されたことが宣言された (the

Finnish Finish)。ここでは15項目の基本エレメントからなる定義(DC-simpleと呼ばれる)は、より詳細な定義の記述と修正の必要性が認められたものの、Dublin Coreの基本としては健全であり、自信を持って実際の利用に結び付けることができることが確認された。現在、Internet Engineering Task Force (IETF)における標準化に向けた準備のための文書として、実利用に向けた記述形式(syntax)と意味(semantics)がいくつかのRFC(Request for Comments)としてまとめられつつある。また、こうした作業を進める上でより詳細な議論と文書の作成を行うためのワーキンググループが作られている。ワーキンググループを作って進められている仕事には、基本エレメントに関する詳細な定義を明確にすること、データのモデル化のための問題点を明らかにすること、Dublin Coreを非電子化資料に適用するための拡張に関すること、情報資源の記述やWWWの標準化に関する委員会等との協力を進めること、Dublin Coreとして認められ暗黙のうちに利用できるサブエレメント(Sub-elements)を明らかにすることなどがある。

メタデータの相互利用性を実現するために記述形式(syntax)、意味(semantics)、および構造(structure)という3本の柱がある。DC-Simpleの定義が安定するに従い基本的な情報資源記述の意味の基礎が固められる。記述形式は記述のための文法規則を形式的に与えるものであり、単純なメタデータに関しては定義済みである(HTML 4.0のMETAタグ)。また、任意の複雑な構造を持つものに関しても検討が進められている(Resource Description Frameworkにおけるassertion block)。第3の柱である構造にはこれから注目が集まることになるであろう。たとえば、エレメント内に書かれた複数の値を区切る記号を何にするか、名字と名前を書く順序をいかにするかといった問題のように、情報資源の記述を行ってきたコミュニティの文化的な問題ともいえる。数多くの問題に関して完全な合意を得ることは不可能であるかもしれない。しかしながら、全体としての意味的な枠組みを与えることで、少なくとも検索者が目標に意味的な近傍にまでたどり着けることを支援することができるようになると期待している。

3. 世界共通の意味を管理する

Dublin Coreは言語や学術分野の違いを超えてメタデータの意味的な相互利用性を提供するものとして有望視されている。ところが、地域や分野毎にサブエレメントを増やしすぎるとDublin Coreが進めてきたメタデータの意味の定義を不明確なものにしてしまう危険がある。この点に関して、世界的な共通性と単純化という性質という観点からエスペラント(Esperantos)、混成語(Pidgin)、混合語(Creole)という3つの言語的な現象を考察することで多くを学ぶことができる。

3.1 エスペラント: 人為的に組み立てられた統合

エスペラントは1875年頃以降の数十年の間に主に欧米で発明された何十もの人工言語のうちで最も有名なものである。既存の自然言語を分析し合成することで作られたので、これらはひとまとめにa posteriori languageと呼ばれる。それらのほとんどは単純な形式と文法を持ち、西ヨーロッパの言語の単語を使用していたり、そうした言語を基礎としていたりする。こうした言語を作ろうとする動きはヨーロッパ各国が海外の植民地を拡張するのと並行して広まった。当時、言語的な多様性は国際間の摩擦の原因とみなされていたが、かといってラテン語に戻ることは考えられず、また英語とフランス語の間の合意がまとまることは政治的に実現可能であるとは思えなかった。科学や社会の進歩、国際間での平和的な共存を進めるには、国際的に利用できる補助言語を利用することが最善であると信じられていた。しかしながら、エスペラントをも含めてどれひとつとしてこうした言語は成功しなかった。

a posteriori languageの多くは単独で定義を進める一人の作者によって作り出され、そしてそれに従う小さなグループの中でのみ使われていた。運動が盛んになるにつれ、言語が作られたはじめの頃からの使用者の間では、新しい単語や構文を採用すべきであるかに関する議論がなされた。しかしながら、そう

したものの多くは意見の一致を見ることのできないものであった。そこでの議論ではその言語を日常に利用しようとする人と言語の専門家との間の要求の違いによるものであった。そうした言語のひとつである Volapük ("World Speak") での議論は、いわば Minimalist と Structuralist との間の衝突によるものであった。Volapük の発明者は自然言語にあるような十分な意味的表現能力を入れようとしたが、一部の使用者は国際的な補助言語として利用が広がる可能性を高めるために簡明さを残すように希望した。エスペラントの活動においても、たとえば曲折アクセント (circumflex) の使用に関する問題などでの議論があり、いろいろな版のものを推進する派閥に別れた。Umberto Eco は「こうした問題は人工言語にとっては避けられない問題である。すなわち、単語というものは利用が広がらない間においてのみ意味的な純粋性が保たれる。ところが、いったん利用が広がると変節者のコミュニティの所有物となる。そして（最善を求めることは良いものを求めることの敵であるので）その結果は Babelization（多くのことばができて互いにコミュニケーションできなくなるという事態）を招くことになる」と結論づけている [6]。

国連がエスペラントの採用を検討したことはあるものの、これまでに人工言語が政府からの援助の獲得に成功したことはない。エスペラントの頑固な支持者は、エスペラントの利用が広がるのはエスペラントが補助言語として利用される場合のみであると考え、マスメディアでの利用を推進すること、標準の維持と新しい提案の吟味、そして言語の発展を制御していくための国際的な管理組織を作ることを進めている。Eco は、これまでの過去の失敗によって将来における補助言語に対する政治的な合意を得るための試みが行われないことを意味するのではないと指摘している。また、Eco は、日常的に生まれる新しい概念をも表現する能力までも持つほど厳密にはする必要はないが、補助言語が成功するには言語の定義を上から (トップダウン的に) 与えることが必要であると考察している。

上からのトップダウン的なコントロールと同じように下からのボトムアップ的な自然な変化によって解決することも必要であろう。二人の言語工学の研究者 Donald Laycock と Peter Mühlhäusler が解答への道筋を示唆している。彼等は、自然言語はどんなものにでも対応し、かつ開放的であるが、その一方人工的に作られた言語は閉じたものであり、規則によって厳密に縛られ、言語学的な自然さがなく、変化には向いていない、と指摘している。人工言語が成功するには、言語の利用者である人間が規則を変更したり、あるいは作ったりすること、システムを状況に合わせたり、意味に関する新たな取り決めをしたりすることがあることに言語の設計者が対処しなければならないと彼等は論じている。そして、こうした方向で人工言語が進歩していく上で、言語工学者は言語の利用者コミュニティがいかに自然発生的に混成語 (Pidgin) を作り出していくかということを十分に検討しなければならないと結論づけている。

3.2 混成語 (Pidgin): 自然発生的ハイブリッド

混成語（Pidgin）は、異なる母言語を持つもの同士が一緒に働いたり、あるいは取り引きをしたりする際に生じるまにあわせのハイブリッド言語である。そこには（社会的に有力なグループの言語から持ち込まれたものがほんとんどで）あまり語彙はなく、語形変化は少なく、語の順も一定しない。強調したい内容は反復やジェスチャーで伝えられる。文法的な正確さにかけるので回りくどい表現に頼らざるをえないことがあり、話し手の間でことばづかいが一致しないこともある。観光地、港町、あるいは移住者の入植地で混成語化 (Pidginization) が続いている。歴史的には、混成語は民族的に混ざり合った植民地の農園の雇い人や奴隷から生じてきたものである。

商売を進める際に混成語が用いられるようになるなどして、使用者にとって言語の利用価値が上がるに従い、ことばとしては落ち着き、語彙は広がり、話し手の母言語として十分な柔軟性を持つようになる。そうなるには、全ての話し手の言語的な要求にこたえられるものでなくてはならない。そうして、前置詞が入り、単語が増え、そして文脈に依存しない構文が用いられるようになる。子供が青年期前の重要な時期に混成語を母言語として用いながら成長すると、本能的な言語能力で文法的に複雑な表現を加え、両親の混

成語から文法的に豊かで表現力に富んだ混合語 (Creole) へと変化してゆくとの研究がある。混合語は正真正銘の言語であり微妙な表現の構文要素や一貫した語順を持つ。しかしながら、この複雑化の過程は世代が進むことによるものだけではない。混合語化 (Creolization) までには至らなかったが言語としては落ち着いたものになりかつ拡張もなされた言語の例として十分に研究されてきた混成語に Tok Pisin と呼ばれるものがある。また、これは前世紀においてはそれは 150 万もの人たちにとっての共通語 (lingua franca) のひとつであり、パプアニューギニア議会での主要な言語であった。

こうした過程はインターネットを情報資源の共有地とする情報資源記述コミュニティにおいて起こってきたことと同じであろう。Dublin Core は 1994 年の第 2 回 World Wide Web の国際会議での立ち話から始まった。この年は一般の人たちがインターネットのことを知るようになった年でもある。いろいろな情報資源記述コミュニティが作り上げてきた記述方法を単純化しハイブリッド化する Dublin Core における努力から、値の属性を指定しない Minimalist と呼ばれる基本エレメント集合（あるいは Minimalist と呼ばれる人達によって指示される基本エレメント集合）が作り出された。前に示したように旅行者は一般的に片言でしか話せないため、いわば混成語で話しているようなものである。そのため、ヴァーチャルツーリストのメタファはこの過程をうまく言い表わしていると言えるであろう。

利用者はセマンティクスとシンタックスにより多くのニュアンスを込めようとするので、minimalist によって示される自然な Pidginization に続いて現れるものは、再び複雑な内容を表そうとする動きであり、Creolization である。その結果混成語より表現能力に富む混合語化したメタデータが現れることになる。Dublin Core における structuralist の展望は複雑な情報を表現しようという意図に基づくものである。すなわち、より詳しい意味を表す新たなサブエレメントを作り出すことによって特定分野のコミュニティにとってより都合の良いメタデータの記述が得られることになる。このふたつのグループ、すなわちより高い相互利用性を目指す minimalist と、より高い意味表現を目指す structuralist の間のトレードオフは、自然言語での混成語と混合語のトレードオフにあたるものであると言えよう。

Dublin Core の発展過程はエスペラントのような人工言語の発展とは大きく異なっている。一連のワークショップの主導した人たちは新しい技術の発明者としてではなく、情報資源記述という分野で働いてきた人たちの持つ知恵を引き出し、経験を集約することで Dublin Core の定義プロセスを進める役割を担ってきた。このプロセスは前例がないほど電子メールやウェッブ上においたドラフト、メーリングリストを活用することによって効率よく進められ、また安価な航空券によって数多くの実践家と Dublin Core に関心を持つ人たちが集まって標準を決定するプロセスに参加することができた。

しかしながら一方、現在のコミュニケーションの道具だけでは何百、何千もの参加者による合意を得ることは難しい。現在でも、Dublin Core の活発なメンバーでさえもメーリングリストといろいろなワーキンググループでの議論に並行して参加することは難しいと考えている。グループの詳細に関する決定が多数の電子メールの山の中に埋もれてしまっているようなことさえもある。

混成語 (Pidgin) も含めて、自然言語は使用されている間に変化していく。これと同じ事がメタデータについても言える。また、自然言語の利用が広がり標準的な言語になっていくのは、日常的な使用、継続的な技術革新、マスメディアや教室、辞書における認知の広がりが互いに影響を及ぼしあうことによる。もし Pidgin メタデータが自然に進化していくことができないほど厳しい制約が与えられるのであれば、このメタデータを技術革新にさらしたり、意味に関する合意を得たり、適切な実践例を正式に認知したりするための公開討論の場を設けることが必要になるであろう。この場は次に述べる Interlingua のようなものでなければならない。

3.3 Interlingua: 言語に対して中立なスーパーセット

ここで述べるInterlinguaは言語に対しては中立に構成されたものでいくつかの言語で表された概念間の意味的関係を表すために用いられるものである。Interlinguaのひとつとして、EuroWordNetプロジェクトにおける言語間(Cross-Language)検索のために開発されたものである[17]。EuroWordNetシステムは単一の言語(monolingual)のためのスペイン語、英語、オランダ語、ドイツ語の全ての基本語を含むwordnet(オントロジ)を持つ。wordnetの内部は、"same as"、"kind of"、"part of"といった意味的な関係が与えられた単語のクラスタからできている。(関係の中には、"near-synonym"、"sub-element"、"role"といったより意味的に深いものも含まれている。)

wordnetを一つのシステムとして統合するため、EuroWordNetでは言語の対ごとにクラスタを結ぶことにしている。しかしながら、この構造は言語の数が増えると結び付けられる言語対の数が多くなるので新たに言語を加えることは非常に難しくなり、それを維持管理していくことは悪夢のように思えることさえある。このプロジェクトでは英語のwordnetに他の個々の言語のwordnet(すなわちmonolingual)を結び付けることも考慮した。しかしながら、monolingualなwordnetを他の言語のwordnetに写像すると、辞句の構成上の違いや言語依存の意味が失われてしまうことがあることがわかった。たとえば、イタリア語のditoは指(finger)であり、またつまさき(toe)でもある。言語に依存した微妙な意味を他の言語で表すことは非常に難しい。

以上のようなことから、EuroWordNetプロジェクトでは、monolingualなwordnetを言語非依存のInterlingua ーすなわち、全ての言語に共通な概念のフラットで構造を持たないスーパーセットーに結び付けることにした。単語はInterlinguaの中に含まれる最も近い意味を持つ要素に対して、等価関係(equivalence)あるいは近等価関係(near-equivalence)を用いて結び付けられる。図2はライオン(lion)が哺乳動物であり、つめのある足(paw)を持ち、たてがみ(mane)を持つことを表す。ライオン、足、たてがみに対するオランダ語、スペイン語、英語およびフランス語の単語が、Interlinguaに定義された対応する概念への並行するリンクによって同義語であることがわかる。

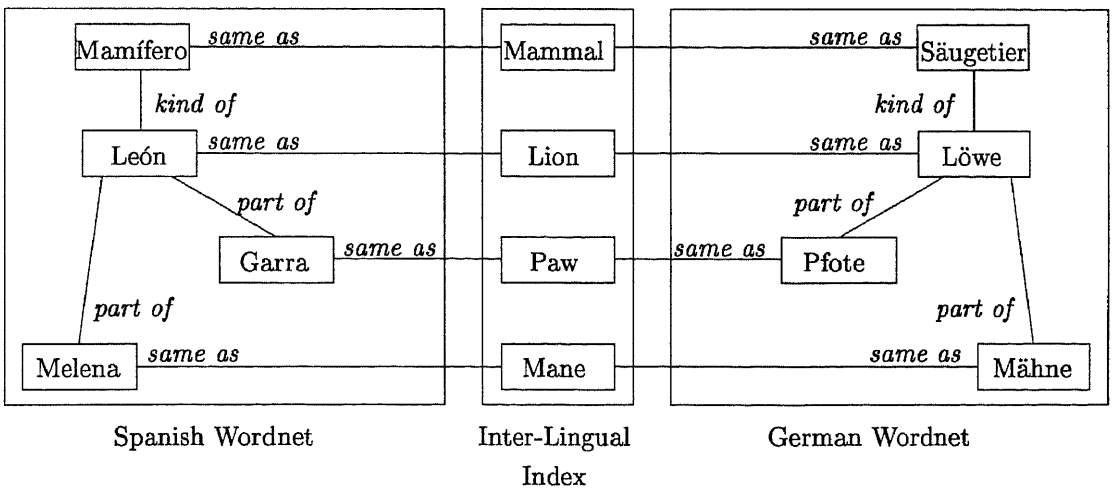

図2: A conceptual interlingua between wordnets

言語に依存する辞句に基づいて概念間の関係を定めようとしても言語によって概念の位置付けが異なることがある。そのため、Interlinguaの中では概念間は意味的なリンクで結ばれていない。言い換えると、Interlingua内で概念間のリンク付けを行ったとしてもそうしたリンクが全ての言語に対して合理的な意味

を持つとはいえない。このように設計することで各言語の wordnet によって多言語による豊かさや広がりを持ち続け、かつ一方で、たとえば dito が fingers、toes、fingers-and-toes に結び付けられるというように言語間での単語間の意味の関係の曖昧さを表すこともできる。

4 レジストリとしての Interlingua

Dublin Core をある種の Interlingua として概念的に理解することもできよう (図 3) 。wordnet 間を結ぶ架け橋として役に立つのみならず、より詳しいメタデータ記述、たとえば GILS や MARC 等への橋渡し、また他の言語や目的や組織に応じてカスタマイズされた Dublin Core への橋渡しのために役立つと考えられる。また、EuroWordNet の場合、既存の wordnet をボトムアップ的に Interlingua に結び付けようとしたのに対し、メタデータ間の Interlingua としての Dublin Core の場合、目的が異なるためトップダウン的である。すなわち、必要に応じてサブエレメントが階層的に深まり、メタデータの内部構造がより詳細になる。こうしたサブエレメントのうちのいくつかは Dublin Core のために作られるものであり、また、それ以外のものは crosswalk を介して取り込まれるものである。

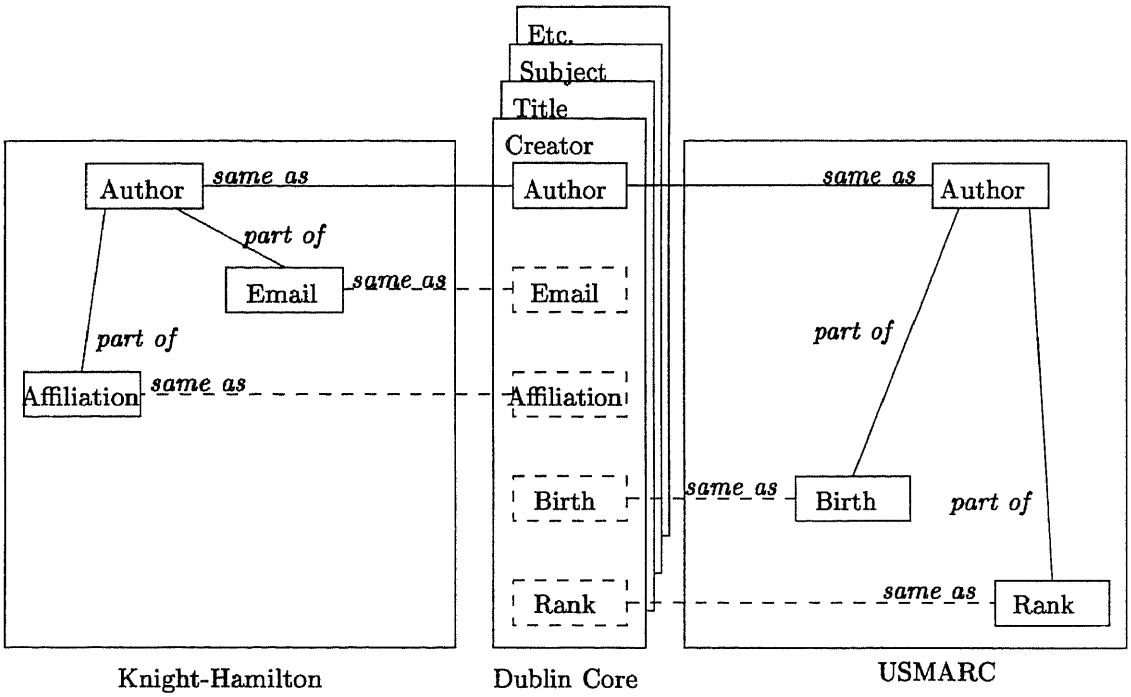

図 3: Dublin Core as an interlingua between description models[7]

4.1 複雑化を管理する

EuroWordNet プロジェクトでは、Interlingua への新しい概念の追加に対して厳密な更新手続きを用意している。あるひとつの単語に対して等価な意味を持つ概念を見つけることができないサイトがあれば、そのサイトはそれを英語で明確に定義し、新たに加えられることになる項目として付け加えなければならない。画面分割型のナビゲーションツールを用いて各サイトからの新たな概念の追加の可能性の報告を定期的にチェックし、既存の概念と重ならないことを確かめた上で Interlingua の更新のための勧告を出す。

Interlingua が更新されると全サイトはそれぞれが持つ wordnet と新たに追加された概念の間のリンクを確かめる。Interlingua を中心に置くことで数多くの wordnet の間での更新管理を多対多ではなく 1 対多にしている。

世界のどこかで"Dublin Core"というラベルを使って用いられたメタデータのサブエレメントとを記述的に結びつけることで、Dublin Core は値には依存しない Interlingua とすることができよう。こうした結びつきによって作られる空間を (そこに参加する人によって秩序が作られる)Dublin Core Market と呼ぶことができよう。しかし、Interlingua としてのエレメントを Dublin Core の規範 (Dublin Core Canon) として認定することを望む人もいるであろう。John Kunze は将来の基準 Core について、地域や組織を限定した拡張や試験的な拡張を公表するメカニズムとそれらを正当と認め基準に組み込むための審査と認定のプロセスに関して論じている [8] 。Core の維持管理には提案された追加が既存のサブエレメントと重複したり、衝突していないかをチェックする必要がある。EuroWordNet の場合と同じように、大きく異なる分野の関連語は Interlingua の中にならべて登録することができるであろう。自然言語の辞書によくあるように、そうした定義はそのエレメントに与えられた代替の意味を表すと考えることができる。Interlingua は認定されたエレメントからなる安定した基本要素部分 (コア) と、コアを取り巻く正式には用いられていない発展段階にある要素によって構成されることになるであろう。

また、Dublin Core Market 自体は実践のための究極的な調停役であると考えることができる。この場合、Interlingua は単に実践の例を示すための目録として働くのみであろう。すなわち、自動的にデータを読み、機関の重複等を考慮しながら利用状況を勘定するメカニズムとして働くのみであろう。特定のサブエレメントがよく用いられること、すなわち相互利用性におけるそうしたサブエレメントの価値は、それらがいかに広く利用されるかによって上がることもあれば下がることもある。

しかし一方、それは標準的な形を持たない言語のようなものであろう。たとえば American Heritage という辞書の利用性評価に関する委員会 (Usage Panel) のようなものを Dublin Core は必要とするであろう。この委員会は、この辞書の編集者が実際の利用例の記述と望ましい形式に関する規定との間のバランスをとることを評価するもので、辞書の 173 人の著者に加えて批評家や学者の助けを得て、順序、明確さ、正確さといった基本的な言語学的美点に対して意味のある項目を評価する役割を持っている。

Market と Canon の両方をカバーするシステムであれば、どのような形式であれ、あるいは言語であれ新しいサブエレメントを提案し公表することで Dublin Core の発展に寄与することができる。Dublin Core のコミュニティが、DDC のように言語に依存しない何らかの方法でエレメント名を定義しようとしない限り、EuroWordNet で採用されたように英語で書かれた概念の辞書的な定義のようなスタイルを踏襲するのが好都合であるように思える。(たとえば、脊椎の指状の部分、新陳代謝される物質。)

4.2 多言語分散レジストリ

世界中のいろいろな意味の情報を集めた定義の中に英語以外の情報を加えるために、非英語の Dublin Core に関する説明を英語で付加したとしても、非英語言語の利用者に対しては彼等自身の言語だけで Dublin Core を利用できるようにしなければならない。それでは、異なる言語で表されているサブエレメントが意味的に等価であると索引付けるというように、異なる言語で記述された Dublin Core 同士をどのようにして相互的に関係づけることができるのであろうか。

RDF の開発者と協力し、Dublin Core コミュニティのメンバーはこうした機能を実現するために必要な簡単なレジストリの設計に参加している。Renato Iannella と Eric Miller によると RDF のレジストリは図 4 に示す機構を提供するであろうとのことである。この図で、ドイツ語の Dublin Core は基準 Dublin Core から DC.title という機械可読形式の名前を継承するが、そのラベルや人間が読むためのエレメント記述にドイツ語のテキストを重ね合わせる。

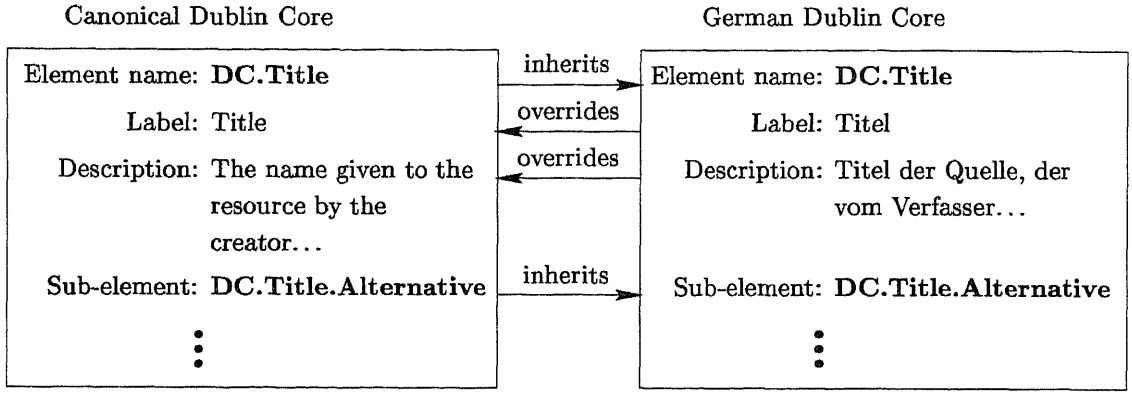

図 4: A registry model in RDF

この概念はシンプルではあるが能力は高い。それはメタデータのエレメントから、定義と正当性の保証を与える参照モデルへのリンクを RDF が与えることになっているためである。Dublin Core に基づくいろいろな実現例におけるサブエレメントが Interlingua として働く基準 Dublin Core へリンクを持つのみならず、基準自身から機械可読形式を得ることもでき、それによってエラーが起きる危険性を減らすことができる。この分散レジストリを言語間参照のために利用する多言語検索ツールの開発にはいろいろな研究が必要となるであろう。

これを実際に利用するには実際的な観点からの問題が多く残されている。1997 年 12 月時点では Unicode はまだ世界中で利用可能であるとはいえず、ごく近い将来にどこででも表示可能になるといえない。フォントの問題や競合関係にある標準間の問題があり、特に日本語に関してはこうした問題は大きい。たとえばブラウザ上での表示のために DC-simple をビットイメージとして送るよりは、日本の図書館情報大学で開発された MHTML を使うことで解決可能な問題であると思われる。これはタイ語、日本語他の言語に関してヘッダ、フォントをテキストと一緒に送り、Java の利用できるブラウザ上で多言語のテキストを読めるようにするものである [16]。

4.3 将来に向けてのより大きな構図

Dublin Core コミュニティと World Wide Web コンソーシアム (W3C) のメタデータ構造定義に関わる人たちが両グループのアイデアと努力を互いに交換し理解を広めあうことで WWW で利用可能なメタデータ構造に関する短期的な展望が大きく開けた。Dublin Core では、電子的情報資源の単純化された記述のために必要な意味基盤を明らかにすることに焦点を当てた開発努力がなされてきた。それによって分野の違いを越えた非常に大きな広がりを持つ情報資源へのアクセスのための障壁間に橋渡しができること、また、いろいろな分野でウェッブ上の情報資源が増やされるに従って別のより豊かな情報を持つ情報資源記述へ橋渡しができることが期待できる。RDF によって、Dublin Core と他のメタデータ記述を相互に利用可能な形で配布するための技術的基盤が作り出される。図書館、美術館、商用のソフトウェア販売業者、大学等、いろいろなコミュニティの専門家の協力によって全てのウェッブ利用者にとって役に立つ解決策に向かってまとまりつつある。情報資源記述の質と網羅性が高まるにつれ、様々な分野を含む全世界で利用可能な情報資源としてのウェッブの価値が恒常的に高まると考えられる。

参考文献

[1] Thomas Baker. Dublin Core in Multiple Languages: Esperanto, Interlingua, or Pidgin? International Symposium on Digital Libraries. Tsukuba, Japan, November 1997.

[2] Priscilla L. Caplan and Rebecca S. Guenther. Metadata for Internet Resources: the Dublin Core metadata elements set and its mapping to USMARC. Cataloging and Classification Quarterly 22(3/4): 43-58, 1996.

[3] Michael Day. Metadata: Mapping between Formats. UKOLN (confirmed 30 September 1997). http://www.ukoln.ac.uk/metadata/interoperability.

[4] Lorcan Dempsey and Stuart L. Weibel. The Warwick Metadata Workshop: a framework for the deployment of resource description. July/August 1996, D-Lib Magazine, July/August 1996. http://www.dlib.org/dlib/july96/07weibel.html.

[5] Dublin Core Homepage. http://purl.org/metadata/dublin_core.

[6] Umberto Eco. The Search for the Perfect Language. Oxford: Blackwell, 1995, pp. 319, 346.

[7] Jon Knight and Martin Hamilton. Dublin Core Qualifiers, ROADS Project, Department of Computer Studies, Loughborough University, http://www.roads.lut.ac.uk/Metadata/DC-Qualifiers.html, 1997.

[8] John Kunze. A Unified Element Vocabulary for Metadata. http://www.ckm.ucsf.edu/personnel/jak/dist.html, 1996.

[9] Carl Lagoze, Clifford A. Lynch, Ron Daniel Jr. The Warwick Framework: a Container Architecture for Aggregating Sets of Metadata. TR96-1593, 21 June 1996. Acrobat version: http://www.nlc-bnc.ca/ifla/documents/libraries/cataloging/metadata/tr961593.pdf.

[10] Donald C. Laycock and Peter Mühlhäusler. Language Engineering: Special Languages. In: An Encyclopaedia of Language. London: Routledge, pp. 843-875, 1994, p. 871.

[11] Ralph LeVan. Dublin Core and Z39.50: Personal Reflections. http://cypress.dev.oclc.org:12345/~rrl/docs/dublincoreandz3950.html.

[12] André Martinet, 1991. Cited in Eco, p. 332.

[13] Geoffrey Nunberg. Usage in the American Heritage Dictionary: the Place of Criticism. In: The American Heritage Dictionary of the English Language, Third Edition. Boston: Houghton Mifflin Company, Pp. xxvi-xxx, 1992.

[14] Diann Rusch-Feja. Dublin Core Version 1.0 in German. http://www.mpib-berlin.mpg.de/DOK/metatagd.htm, 1996.

[15] Praditta Siripan. Dublin Core in Thai. National Science and Technology Development Agency, Bangkok, Thailand, 1997.

[16] Tetsuo Sakaguchi, Akira Maeda, Takehisa Fujita, Shigeo Sugimoto, Koichi Tabata. A browsing tool for multi-lingual documents for users without multi-lingual fonts. Proceedings of the 1st ACM International Conference on Digital Libraries (March 20-23, 1996), pp. 63-71.

[17] Piek Vossen, Pedro Diez-Orzas, Wim Peters. Multilingual design of EuroWordNet. http://www.let.uva.nl/~ewn/Vossen.ps, 1997.

[18] W3C Web Site: Resource Description Framework. http://www.w3.org/Metadata/RDF.