第13回 韓・日国立図書館業務交流 Ⅱ

韓・日自動翻訳についての現況と課題

―MDR基盤メタデータ相互運用性確保―

パク・ジンホ / 国立中央図書館デジタル企画課

1. 序論

知識情報社会において情報を入手、管理、保存、サービスしなければならない諸機関の共通の関心事でありトレンドは、セマンティックウェブ(Semantic Web)である。セマンティックウェブは、W3C の Tim Berners-Lee が使った新造語で、機械が理解できる(Machine Understandable)ウェブを作ろうという新しい環境を意味する。

図書館が Semantic Web に関心を持ち肯定的な態度を持つようになったのは、機械可読型(Machine Readable)メタデータ(図書館の MARC)を活用するところにあるのであり、他の知識情報機関よりも有利な位置に立つことができるのは、これまで管理し続けてきた主題名標目表、著者典拠などの統制語彙(Controlled Vocabulary)が存在するためである。

Semantic Web についての議論は、結局、利用者がより意味のある情報探索をできるようにするためのものであり、図書館分野では国際図書館連盟で発表された FRBR(Functional Requirement for Bibliographic Records)が代表的な例として挙げられる。1997 年に承認され 1998 年に公表された FRBR は、1,2,3 Group を持つフレームワークで、現在は“FRBR Family”形態で以下のような構造を有する。

- FRBR:Functional Requirement for Bibliographic Records (1 Group 中心)
- FRAD:Functional Requirement for Authority Data(2 Group 中心)
- FRSAD:Functional Requirement for Subject Authority Data(3 Group 中心)

第 13 回韓日業務交流主題発表のテーマは、「韓・日自動翻訳についての現況と課題」であり、日本の国立国会図書館はこれまで、自然言語処理(Natural Language

Processing)の応用分野である機械翻訳(Machine Translation)について多くの研究開発を試みてきた。韓国語と日本語は類似語族であり、機械翻訳による効果は他の言語よりはるかに良い結果を示す。これまで、民間の検索サービス業者や図書館などでは、この長所を活かして多様な試みをしてきており、初期は形態素の置換からはじまり、形態素、品詞の分析をおこなって文法に適うよう再配列する直接方式を経て、変換規則を基盤に言語を分析した後に文章を生成する変換方式まで、絶え間ない発展をしてきた。自然語処理の様々な分野の中で、機械翻訳は文書を自動で目標とする言語に変換してくれるという点で、情報に対するアクセスポイントではなく、目標とする情報への内容的なアクセスが可能である。その点で、学術研究者だけでなく該当言語の文化に対する好奇心と探索欲求を持った利用者にとって重要な意味を持つと考えられる。

図 1　韓国 NHN 提供日本語ウェブ翻訳機活用 NDL ホームページ翻訳画面

これまで列挙した Semantic Web、FRBR そして機械翻訳はすべて、利用者がより容易かつ便利に意味ある情報を活用できるようにするという共通点を持っている。韓日自動翻訳(機械翻訳)についての現況は、日本の事例を通じてより具体的な例を調べることができるため、本稿では韓中日デジタル図書館協力においても核心的な推進課題であるメタデータの相互運用性確保のための方策として、メタデータレジストリ基盤の相互運用性確保について述べることにする。

2. メタデータと相互運用性

2.1 メタデータ

一般的にメタデータは、データについてのデータと解釈され、デジタル知識情報資源を記述するための用語として使われていたが、1995 年、Dublin Core Metadata Element の開始とともに図書館分野でも本格的に使われ始めた。現在では、文書の構造と表現に対する分離の要求とともに、情報源の特徴を記述するための構造化されたデータとして、より具体化された意味を持って活用されている。伝統的に図書館では目録と標目という用語が長く一般的に使われていたが、現在はメタデータという用語がより一般的に使われている実情であり、機械可読型目録である MARC は、図書館分野で世界的に標準化された(もちろん国ごと、地域ごとに活用には差がある)編目法だ。

MARC は、これまで図書館が伝統的に使ってきた編目規則をそのまま適用することができ、標準の語彙と構造を用いて共同編目が可能、利用する図書館に合うように部分的な修正が可能、共同目録作業による質的上昇効果、時間的経済性などの長所にもかかわらず、書誌情報が持つ構造的な難しさ、新しい要素の追加が困難なこと、複雑な記述規則による使用の難しさなどの短所が長らく指摘されてきた。

反面、ダブリンコアは、メタデータの多様性を認めてこれを受容できる包括的な構造(Lagoze,1996)であり、ウェブ上での電子的、物理的資源に対する探索が容易なように、15 個の基本要素と限定語から出発し、現在では最も一般的かつ標準的なメタデータとして活用されている。

2.2 相互運用性

相互運用性は、異種システム間において問題なく自動化された情報交換が可能になることであり、またメタデータとの関連においては、互いに異なるメタデータの交換において、その意味するところをそれぞれのシステムが理解できるようにすることであり、

主にそれぞれのメタデータスキーマが有する語彙のマッピングにより相互運用性を確保する。ダブリンコアは、相互運用性の確保のために寄与したところが大きいと言える。

これまでの韓日業務交流を通じても確認してきたが、両国立図書館は MARC という同一の記述規則を有しているが、その活用には差があり、データ交換のためには別途のマッピング規則を必要とした。図書館の相互運用性以外にも、Dublin Core、ONIX、CERIF、INDEX、GDAS などの多様なメタデータの存在は、相互運用性を阻害する要因として作用したが、同様に主にマッピングを通じて問題を解決してきた。

しかしマッピングは、メタデータ間の完璧な 1:1 関係の設定が不可能な場合が発生し、構築後に変更事項が発生すると再作業が必要になるなど、また別の問題を生じている。

互いに異なるメタデータの必要性と存在を認めるなかで、より効果的な管理をするためには、共同で標準的な組織と手続きにより、メタデータの意味を管理して活用することが最も良い代案になり得るが、代表的な事例として ISO11179 Metadata Registry が挙げられる。

3. ISO 11179 MDR

3.1 MDR 概要

ISO/IEC11179 MDR は、ISO/JTC1 SC32 WG2(メタデータ、Metadata)において研究開発され制定された標準であり、データを記述するために必要なメタデータの品質と種類(kind)を説明し、メタデータレジストリでメタデータの管理と運営を説明する。11179 はデータの表現(representations)、概念(concepts)、意味(meanings)を形成する場合に適用されるが、このような要素間の関係が、データを生産する組織に関係なく人と機械の間で共有できるようにすることにその目的がある。

このようなアクセスが可能なのは、メタデータはやはりデータであるがためにデータベースに保存することができるというところにある。このようなメタデータの登録機能を持っているデータベースを、メタデータレジストリという。

11179 は 6 個のパートで構成され、次の通りである。

表 1　11179 の構成と主要内容

名称	主要内容
Part 1:Framework	全体のフレームワーク
Part 2: Classification	分類
Part 3: Registry metamodel and basic attributes	データを記述するための MDR の概念的ドメイン
Part 4: Formulation of data definitions	望ましい定義を与えて、命名規則を作れるようにする規則と指針
Part 5: Naming and identification principles	
Part 6: Registration	11179 の要求と手続きによるメタデータ登録に関する側面

3.2 11179 Part 1:Framework

Part1 は、ISO/IEC 11179 の各部の連結関係と理解のための方法を提供していて、メタデータとメタデータレジストリの概念的な理解のための基礎情報を盛り込んでいる。11179 標準で使う用語の概念を明確にし、11179 の各パート間の関係を記述している部分であり、特に MDR において最も重要な概念的モデル(conceptual model)の要素である、データ要素概念(data element concepts)、データ要素(data elements)、値領域(value domains)、概念的領域(conceptual domains)間の関係を説明している。

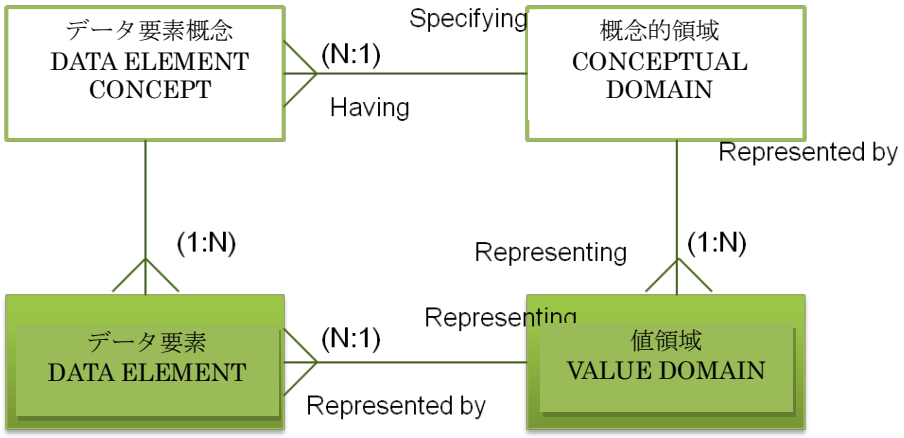

図 2　MDR 概括モデル

概括モデルのデータ要素概念と概念的領域は、概念層(Conceptual Level)に該当するが、抽象的な概念でまだ具体化されていないものを意味する。データ要素と値領域は、表現層(Representational Level)として具体化され、目で確認することができ、明白に表現（例Country_Name_Code/ISO3166 2 Alpha Code）できるものだ[1]。

3.3 11179 Part 2:Classification

Part2 では、データ要素概念とデータ要素を分類方法に基づいて分類する手続きについて記述していて、レジストリを管理する登録機関(Registry Authority)がデータ要素の登録時に考慮または使用できる分類システムについて叙述している。

3.4 11179 Part 3:Registry Metadata and basic attributes

Part3 では、MDR を人が情報を理解する方式と最も近い形式の概念的なデータモデルであるメタモデルとして記述している。メタデータレジストリの構造を概念的なメタモデルの形態で提示して、標準化されたメタデータを通じたデータ要素の正確な意味の把握と体系的な管理について説明する。

メタモデルは管理と記述、命名規則、分類、管理アイテムに関するもので、次のように図を用いて表現することが可能だ。

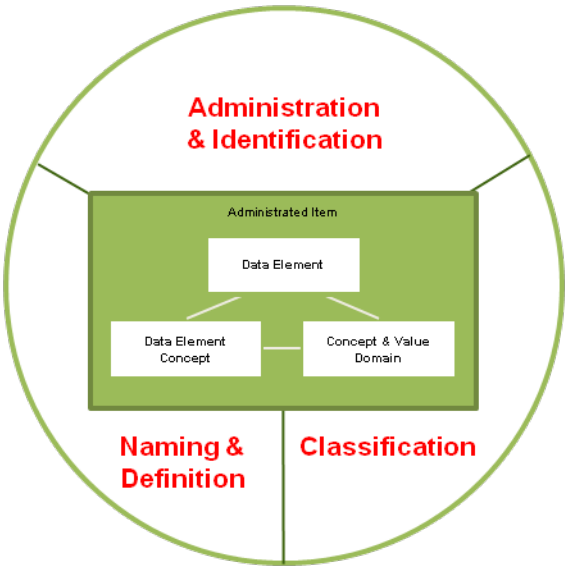

図 3 MDR メタモデル

1 概括モデルについての具体的な例と説明は業務交流当日の発表時に行い、本稿では省略する。

管理と識別は、データ要素の管理のためのメタデータ項目で、管理項目の登録者情報、識別情報、提案者情報、担当者情報、参照情報、登録/生成日、発効日、注記事項などを扱う。

命名および定義は、管理項目が活用される主題分野(Context)、使用言語の明示、同意語など多様な表現方法の支援、言語ごとの多様な意味の使用支援を扱う。

分類は、分類スキームと関連分類スキーム内に存在するリポジトリ構成要素を管理するためのもので、使用される分類システムの類型定義(分類、キーワードリスト、用語辞典、ダイヤグラムなど)、分類システムを構成する項目(分類群名、用語、関係など)、項目の関係定義を扱う。

管理項目は、データ要素、データ要素概念、概念領域、値領域に関するもので、上記ですでに説明したところである。

3.5 11179 Part 4:Formulation of Data Definitions

Part4 では、データの定義と構成について説明しているが、データ定義の規則は、代表的に次のように要約することができる。

-単数項目を定義(「紙（複数)」に対する「紙（単数)」についての定義)
-肯定文を使用
-同意語の使用を避け、概念を最もよく説明する節、句を使用し定義
-できる限り略語は避け、略語を使う場合はよく知られた普遍的な略語のみ使用
-異なる定義を混ぜて定義することは禁止

また、データ定義指針の内容は次のように要約することができる。

-本質的概念のみ記述、簡潔かつ明確な記述、曖昧だったり包括的な表現は避ける
-理解のための付加的な情報が必要ない語彙の使用
-該当分野の背景知識や脈絡を理解していなくても理解可能なように記述
-文章よりも関係型名詞節を使用(The name of the country where mail is delivered)
-表現クラスを名詞にした関係節(the name of,the code that represents,the text that describes,the measure of the..)

3.6 11179 Part 5:Naming and Identification

Part 5 では、データ要素の概念、概念的領域、データ要素、値の領域などの管理項目に関する命名と識別に関する指針を提供する。登録データ(管理項目)の識別のためには各登録項目を識別できる識別子が必要だが、この標準では IRDI(International

registration data identifier)を使用し説明している。IRDI は RAI(registration authority identifier)、DI(data identifier)、VI(version identifier)で構成される。

体系的な命名規則のためには、Scope(範囲)、Authority(命名者情報)、Semantic rules(意味付与規則)、Syntactic rules(意味の文字配列規則)、Lexical rules(同意語、語の長さ、綴りなど)、Uniqueness rules(同音異義語などの規則)について説明している。

3.7 11179 Part 6:Registration

Part6 は、実際に登録者(Registrar)を通じてデータ要素を登録、検証、認証して標準化する手続きについて説明している。データ要素の標準化は、MDR を中心に多くの利害当事者が関係して成り立っていて、各利害当事者の役割と責任、データの状態に対する定義を明確にしなければならないことを明示している。

登録関連の概念図は次の通りである。

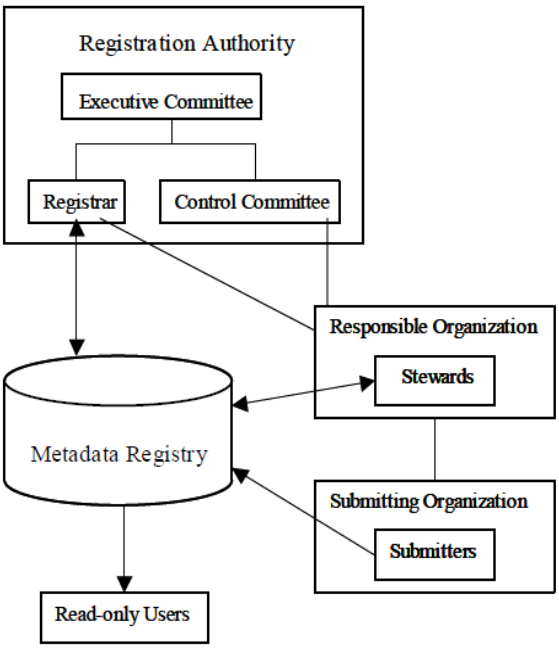

図 4　MDR 登録手続き概念図

概念図に明示された各責任者(組織)の役割は以下の通りである。

- 管理委員会(Executive Committee):各専門分野別機関が提案するメタデータを標準化されたデータ要素として登録するための議決および承認
- 登録責任者(Registrar):管理責任者の検証が完了したメタデータを認証
- 統制委員会(Control Committee):登録責任者が認証したメタデータを標準化

・管理責任者(Stewards):提出者が提出したメタデータを検証
・提出者(Submitters):メタデータのデータ要素を登録、修正、削除
・利用者(Read-only Users):MDR で標準化されたデータ要素を検索およびダウンロード

4. 結論

情報技術の発達は、目録情報の変化、多様な形態の知識情報の管理/サービス、デジタル機器を基盤にしたサービス形態の変化など、内的、外的に多くの変化を図書館に要求している。

しかし情報技術を基盤とした知識情報社会において図書館がもつ最大の長所は、他のどんな機関よりも多様な形態の知識情報を収集、管理、保存してきた経験と、原資料にアクセス可能なように構成した 2 次情報源すなわち目録、メタデータの存在と、主題名標目、著者典拠など長い間に構造化された分類システムを維持管理しているという点である。

情報化時代に量的な面での爆発的な知識情報の増加は、「どれくらい正確なのか？」「信頼できる情報なのか？」といった問題の重要性を浮き彫りにしており、また「関連する他の情報」へのアクセス提供を要求している。考えてみると、このような必要性と課題は、図書館において新しいものではない。ただ、それを実現させるために必要な新しい発想とアクセス方法を要求しているだけなのである。

多様な分野、多様な必要性によって生じるメタデータを非効率的だと考えることはできない。図書館が既存の MARC の持つ限界を超えるために努力するのと同じ論理だ。いま重要なのは、このように散らばっている情報をどのように結びつけるかということではないか？どうすれば情報を関連付けて活用できるようにしていくのか？であり、人間と同じように機械が理解できるようにするという目標の下に、図書館が持っているメタデータを再度振り返ってみなければならない。

業務で毎日のように出会う題名 Title、著者 Author が本当に意味するところが何であるのか司書として人間として理解するのと同様に、私たちと情報が会う媒体である多様なプラットホーム(ウェブ、モバイルなど)が私たちと同じように理解して自動で情報を伝達できるようにすることが、結局のところセマンティックウェブ、機械翻訳のような概念、技術などを作り出したのだ。目標は抽象的であり、RDF、OWL、SPAQLE のような新しい理解を必要とするが、最も基本になるのはやはり徹底したメタデータの管理だ。意味基盤としてのウェブ、情報流通を願うならば、まずはメタデータ自体を意味のある個体として扱うことが可能でなければならない。

[参考文献]

・コ・ヨンマン、ソ・テソル、イム・テフン「意味互換のためのメタデータマッピング研究」、韓国情報管理学会、情報管理学会誌、第 24 巻 4 号(2007.12),pp. 223 ～ 238
・コ・ヨンマン「メタデータ標準化とメタデータレジストリ」、国会図書館、国会図書館報第 42 巻第 11 号(2005.11),pp. 18-26
・コ・ヨンマン「ISO/IEC 11179 標準ファミリー」、第 4 回メタデータ標準化セミナー、産業資源技術標準院、pp. 3-30,2005.11.24
・イ・ジェムマ「ISO 11179(Metadata Registry)の概念と適用方策」、国家記録院、記録保存第 18 号(2005),pp. 117-132
・ナ・ホンソク. 2006.メタデータレジストリ互換、2006.「第 3 回メタデータ標準化ワークショップ」、2006 年 5 月 25 日. [ソウル:成均館大学校].
・ ISO/IEC JTC1.2004.ISO/IEC 11179 Information Technology - Metadata Registries:Part1
・ ISO/IEC JTC1.2005.ISO/IEC 11179 Information Technology - Metadata Registries:Part2
・ ISO/IEC JTC1.2003.ISO/IEC 11179 Information Technology - Metadata Registries:Part3
・ ISO/IEC JTC1.2004.ISO/IEC 11179 Information Technology - Metadata Registries:Part4
・ ISO/IEC JTC1.2005.ISO/IEC 11179 Information Technology - Metadata Registries:Part5
・ ISO/IEC JTC1.2005.ISO/IEC 11179 Information Technology - Metadata Registries:Part6
・韓国標準協会. 2006.KS X ISO/IEC 11179-1 情報技術-メタデータ レジストリ(MDR) -第 1 部:フレームワーク