

IMI共通語彙基盤

加藤 文彦¹ 武田 英明¹ 田代 秀一² 平本 健二³ 松澤 有三⁴

¹国立情報学研究所 ²情報処理推進機構 ³経済産業省 ⁴インディゴ（株）

オープンデータとして行政が保有するデータの利活用を推進していくためには、データで用いるさまざまな用語の表記や意味、構造を統一することが重要である。IMI共通語彙基盤は、共通で使われる語彙とその語彙同士の関係を示す仕組みで構成されるフレームワークである。本稿では、共通語彙基盤の各構成要素と、その中で現在主に整備が進んでいるコア語彙についての設計や構造、実装を述べる。また、コア語彙を中心とした共通語彙基盤の利用や実際の活用事例についても述べる。

1. はじめに

2017年5月に「世界最先端IT国家創造宣言・官民データ活用推進基本計画」[1]（以下、IT戦略）が政府により作成された。これは2016年12月に制定された「官民データ活用推進基本法」（以下、官民データ法）を受けたものであり、官公庁・自治体が行政データを積極的に公開し、民間セクターがそれを利活用するなど、データの保有主体を超えた円滑なデータ流通の促進により、新たなサービスやイノベーションの創出、データに基づく行政や産業等の改革を狙ったものである。これまではオープンデータのような自治体によるデータの公開・利活用は任意で行われていたが、この法律および基本計画により、都道府県は義務、市区町村は努力義務として官民データ活用を推進していかなければならない。

データ公開や利活用を推進していく上で大きな問題となるのが、データの標準化である。データの標準化というとデータ形式が思い浮かぶが、それだけではなく使われる語彙やコード、文字のレベルでも標準化が必要である。これらが交換可能でないと、形式だけ一緒でもデータの中身の解釈が異なることになってしまう。そのため、なるべく同種類のものを表すデータを同じ仕組みで共有する必要がある。

IMI (Infrastructure for Multilayer Interoperability, 情報共有基盤) は、電子行政分野におけるオープンな利用環境整備に向けた政府のアクションプランの一環で、データに用いる文字や用語を共通化することで情報の共有や活用を円滑に行うための基盤である。現在IMIは、データに共通で使われる語彙やその構造を共有するための仕組みである共通語彙基盤と、行政が必要な文字を扱うための仕組みである文字情報基盤から構成されている。本稿では2013年より取り組んでいる共通語彙基盤について取り上げる。

2. 共通語彙

共通語彙基盤の中核をなすのが共通語彙である。共通語彙基盤は、コンピュータ間のデータ通信を円滑かつ確実にを行うことを目的として、概念の代表的な表記としての1つの語の意味や構造、語によって表される概念と他の概念の関係などを明確にした概念の集合を指す。共通語彙基盤では、語で表される概念を用語と呼ぶので、語彙は用語の集合である。

共通語彙基盤は米国で開発が進められているNIEM[2] (7.1節参照) を参考に開発が開始された。NIEMに習い、語彙にはコア語彙とドメイン語彙があると想定している。コア語彙は、「人」「氏名」「組織」といった、分野を超えて使われる共通性の高い用語の集合である。ドメイン語彙は、防災、財務といった各分野内で共通な用語の集合であり、コア語彙を継承して定義することを想定している。

コア語彙については、NIEMも共通語彙基盤も基本的にはトップダウンで整備している。共通語彙基盤のコア語彙では、現在60近くの基本的な概念を定義している。一方でドメイン語彙については、

NIEMでは先にドメインの担当者を決め，そこを中心にドメイン語彙の開発が行われている。しかし，共通語彙基盤ではドメイン語彙はまだ存在しておらず，次のようにボトムアップで整備していくことを考えている。

データを構築するにあたってはコア語彙だけでは不足することがある。その場合，現場の必要に応じ，既存の語彙を継承した独自の語彙を定義する必要があるが，これを応用語彙と呼んでいる。応用語彙は，将来，分野に共通な語彙を洗い出すなどによりドメイン語彙へと整理されていくことを想定している。現状では共通語彙基盤にはドメイン語彙は存在していないが，それを含めると，語彙は3層の構造を持つこととなる（図1）。



図1 共通語彙の3層構造

3. コア語彙の設計

共通語彙の設計について，現在整備されているコア語彙2.4版を例に説明する。

3.1 クラス用語とプロパティ用語

コア語彙は共通語彙の基礎部分であり、氏名、住所、組織といったあらゆる社会活動で使用される用語の集合である。用語には、なんらかの事柄に関する概念を表すクラス用語（以下、クラス）と、それらの事柄の性質や、事柄と事柄の関係を指し示す概念を表すためのプロパティ用語（以下、プロパティ）がある。クラスは一つ以上の組み合わせ可能なプロパティを持つ。プロパティはデータの値を直接持つ場合と、ほかのクラスと関連付けられる場合がある。

図2はコア語彙の人型クラスから抜粋した例である。人型クラスが性別プロパティと氏名プロパティを持っており、性別プロパティの値型は文字列、氏名プロパティの値型は氏名型である。値型が文字列のようなテキストデータの場合は、主にXML Schemaで定義されているデータ型[3]を指定する。値型が氏名型のように共通語彙のクラスである場合は、さらに氏名型が姓名プロパティのようにさまざまなプロパティを持つため、階層構造となる。

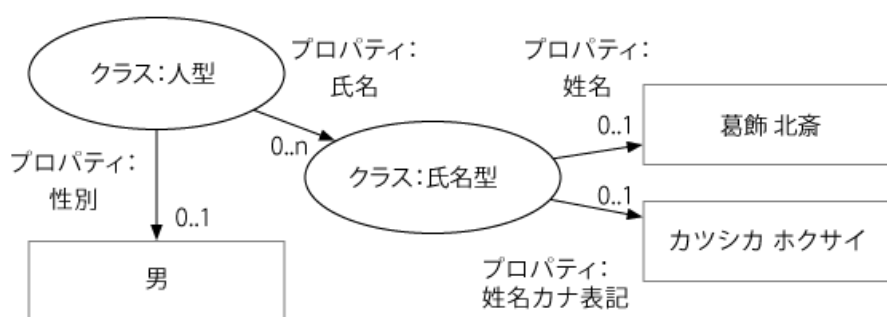


図2 クラスとプロパティ

プロパティはクラスごとに出現回数を指定することができる。人型クラスでの性別プロパティは0から1回（0..1）、氏名型は0からn回（0..n）である。コア語彙はさまざまな要求に応じられるように、プロパティの数制約は必要がない限り0..nにして、制約が少ないように設計されている。

3.2 クラス階層

あるクラスを継承して、プロパティの追加などを行うことで新たなクラスを定義できる。共通語彙におけるクラス継承は、既存のクラスからプロパティをそのまま継承して、さらにプロパティを追加することが可能である。既存のクラスのプロパティを削除したり制約を加えたりすることはできない。コア語彙もこの継承の仕組みを用いて設計されている。また、ドメイン語彙や応用語彙のように新たに語彙を作成する場合には、コア語彙のクラスを親や祖先として継承する形で定義することを求めている。これにより、相互運用性を高めることができる。

図3は人型から見た周辺のクラスの関係である。点線矢印は継承関係を表す。コア語彙では、全てのクラスのベースとなるクラスとして概念型クラスを定義している。その下位クラスとして、識別可能なことやものを表すクラスとして事物型クラスや、IDやコード等を表現するための特別なクラスがある。

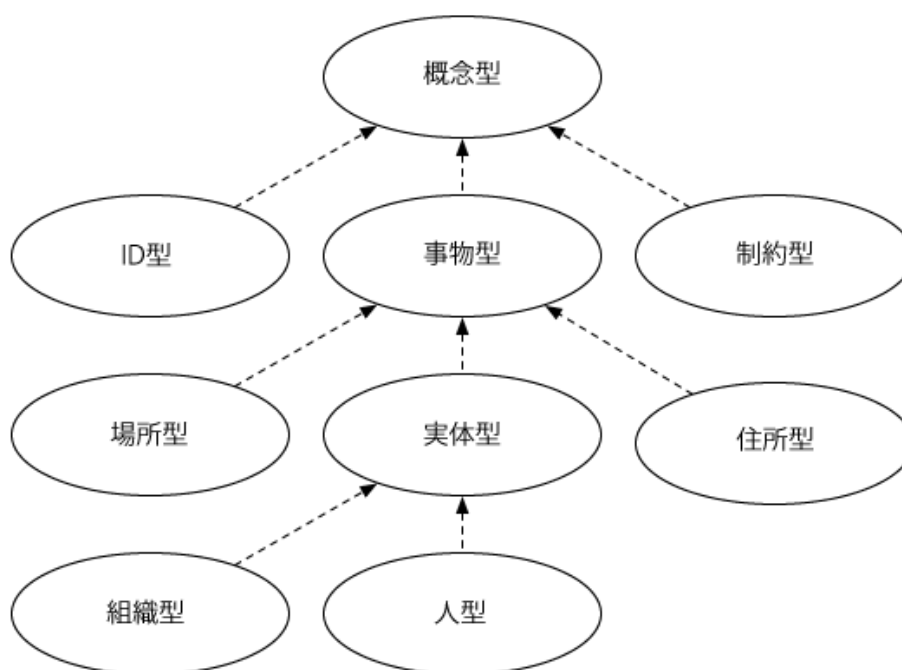


図3 クラス階層

共通語彙基盤ではデータの一塊になるものは識別子をつけられると想定しているので、通常のクラスは事物型を継承して定義する。コア語彙は、人型クラスや組織型は実体型クラスを継承しており、

実体型クラスは事物型クラスを継承している。施設型等の上位である場所型も事物型を継承している。図3の各クラスのより詳しい説明は表1を参照していただきたい。

表1 図3の各クラス説明

概念型	すべてのクラスのベースとなるクラス。
事物型	識別可能なことやものを表すクラス。IDや表記、説明といったデータの基本的なプロパティを持っており、通常のクラスは事物型を継承する。
ID型	IDを表す特別なクラス。識別値やその識別値が属するID体系を指すプロパティを持つ。
制約型	年齢制限のように、対象となる事物の制約条件を表す特別なクラス。さまざまな制約条件を記述できるように制約型の下位クラスとして範囲制約型や期間制約型等が用意されている。
場所型	場所を表現するクラス。場所の名称や地理識別子、住所、地理座標といった、場所表現の基本的なプロパティを持つ。
住所型	住所を表すクラス。郵便番号や都道府県、市区町村、丁目といったプロパティを持つ。
実体型	組織や人を抽象化するための便宜的なクラス。下位のクラスとして組織型や人型がある。
組織型	組織を表すクラス。代表者や設立日、関連組織、連絡先といったプロパティを持つ。
人型	人を表すクラス。氏名や生年月日、出生地、年齢といったプロパティを持つ。

3.3 設計での知見

コア語彙は、2014年4月にコア語彙2.0検証版として公開した。その後、自治体の参加を得た実証実験での経験、パブリックコメントの意見、NIEMやISA²等類似プロジェクトの担当者との意見交換を受け、用語の追加、構造の整理など改版を続けた。

共通語彙の設計方針で重要なのが、「人」のように日本語で概念定義をすることである。通常の語彙設計では英語で定義することが多く、パブリックコメントでも同様の意見をいただく。しかし共通語彙の目的は日本語の意味交換なのであえて日本語で通しており、実際に効果がある。たとえばイベント型は催し物を表すクラスであり、英語でEventと言うには問題があるが、言語間の意味の差異に悩む必要がない。

2.1検証版では、ベースの型を基本型から事物型に変更した。これは語彙定義のためによく用いられるオントロジー言語OWLにおける、ベースの型であるThing（事物）クラスに寄せたためである。また、テキスト型や真偽値型のように末端のデータ型を独自作成していたが、階層が無駄に深くなり使いにくいとの意見により、XML Schemaのデータ型を直接使うように変更した。

2.2版は初の正式版として公開した。2.2版ではクラスとプロパティを分離してプロパティを汎用化した。たとえば、「名称」というプロパティを、「組織」や「施設」のそれぞれに属するものではなく、独立した汎用的なものと考えようになった。

もう1つ大きな変更は地理空間情報関係のクラスである。地理空間情報コミュニティからの意見により、地点型を地物型に変更し、場所型、地物型、施設型等の関係を整理した。また、座標は点だけではなく線や面にも対応するようにした。

2.3版では単位コード型を追加した。重量や面積など数値で単位が必要な場合は多いが、単位を共通化する仕組みがなかった。また、関係する人や組織を表現するための関与型や、人や組織の活動を表現するための活動型も追加した。

2.3.1版は設計上の変更ではなく、英語表記や説明を追加した。海外連携のための変更である。また、2.3.2版はドメイン変更のみであり、実装上の問題なので4.5節にて後述する。

現行の2.4版における変更は、ほとんどのクラスで必要とされるプロパティである「ID」「表記」「参照」「画像」「説明」を事物型クラスに集約したことである。事物型は抽象クラスであり、プロパティは持たなかったが、事物型以下で常に必要なプロパティは事物型で持つという設計変更を行った。一方で事物型に当てはまらないID型等のクラスを整理するために、トップのクラスとして概念型を新たに設けた。

また、文書型とサービス型、さまざまな制約型を追加した。文書型は文書情報を表現するクラスであり、プロパティは文書のメタデータとしてよく使われるDubline Core Metadata Element Set[4]に準拠した。サービス型は、保育サービスのよう、行政におけるサ

ービスを表現するクラスである。サービスやイベントは利用対象や期間が限定されていることが多いため、範囲制約型や期間制約型といったさまざまな制約を記述するためのクラスも追加した。

4. 共通語彙基盤の実装

共通語彙基盤は多様なニーズに応えられるように各種データ交換形式および構造化項目名という文字列表現で実装している。また、DMDと呼ばれるパッケージによって、語彙利用に関する情報を交換できる仕組みを用意している。さらにこれらに共通する内部表現として、IMI語彙記法を用意している。

4.1 データ交換形式

語彙やデータモデルは抽象的なものだが、実際にコンピュータでデータを扱うためには、なんらかの形式で表現する必要がある。共通語彙基盤ではXML, RDF, JSONを用意している。最初はXMLだけだったが、ニーズに応じて検討追加をしてきた。各形式での用途に限定はないが、XMLはコンピュータ間の厳密なデータ交換、RDF及びJSONはオープンデータのようなインターネット上のデータ交換を想定している。

共通語彙基盤サイトでは、XMLとRDFで記述したコア語彙のスキーマを公開している。JSONは、RDFの派生であるJSON-LDで実装しているので、RDFによるスキーマを参照できる。また、コア語彙用のJSON-LDコンテキストも提供している。以下は、図2をJSONで表した例である。

```
{
  "@context": "http://imi.go.jp/ns/core/context.jsonld",
  "@id": "http://example.org/# 人 1",
  "@type": "人型",
  "氏名": {
    "姓名": "葛飾 北斎",
    "姓名カナ表記": "カツシカ ホクサイ"
  },
  "性別": "男"
}
```

4.2 構造化項目名

共通語彙は階層構造を持つが、表形式データでデータの項目名を記載するヘッダ部分に共通語彙を使うというように、階層構造を1つの文字列として表現したいという必要がある。構造化項目名はそのような用途を想定して設計した。似た仕組みとしてXPath[5]があるが、XPathのように文書中のノードを特定するのが目的ではなく、構造を保ったまま用語を記述するのが目的である。そのため、途中の用語を省略することはできない。

構造化項目名は、特定の項目群の構造を定めたスキーマを持つわけではなく、単なる文字列として表現する。基本構造は1つのクラスと任意のプロパティを半角の「>」で区切って並べたものになる。たとえば図2の断片を構造化項目名で表すと以下になる。

人型 > 氏名 > 姓名
人型 > 氏名 > 姓名カナ表記

これは最も単純な例であるが、プロパティにグループ名を指定することで複数のプロパティが同一のインスタンスに結びつくことを明示したり、値に制約を加えたりすることができる。

4.3 DMD

コア語彙は汎用性重視の設計のため、特定のデータに対しては不必要なプロパティが含まれている。たとえば人型には継承も含めると現在21種のプロパティがあるが、常に全てが必要とされるわけではない。住民データを記述するためには本籍は重要かもしれないが、芸術作品の作者データを記述するのには必要ない。一方、作者データでは本名や作者名のように氏名複数が当たり前である。このように、データ構築の現場の必要に応じ、プロパティの取捨選択などによるデータモデルを記述し、共有する手段が必要とされる。

共通語彙基盤では，その目的のため，DMD（Data Model Description）を用意している．これはNIEMにおけるIEPD（Information Exchange Package Document）に相当するもので，当初共通語彙基盤でもIEPと呼んでいたが，日本独自のものに発展したので，改めてDMDと呼ぶこととした．現在共通語彙基盤サイトでは11件のDMDが共有されている[6]．

DMDを活用してデータモデルを共有し，使いまわすことができると，データ作成者は新たにデータ項目やデータモデルを検討する手間を省ける．また，DMDで指定されているデータモデルに合わせてアプリケーションを開発することで，そのDMDに従って作成されたすべてのデータを扱うことができる．たとえば公共施設用DMDがあれば，データ作成者が公共施設データをそのDMDに従って作成することで，同じDMDを用いて公開されているほかの公共施設データと互換性を保つことができる．そのDMDに対応しているアプリケーションはどちらのデータも使うことができる．

DMDにおけるデータモデル記述には，4.4節で述べるIMI語彙記法を用いる．また，DMDにはデータモデルと代表的なラベルを結びつけるためのマッピングを含む．代表的なラベルとは，多くの場合DMDを作成するときにベースにしたデータの項目名である．ラベルに対応するデータモデルの断片は構造化項目名で指定する．以下の例では元データの項目名である「名前」を「人型>氏名>姓名」に対応付けている．表形式データからデータモデルにマッピングすることは良く行われるため，マッピングの記述がDMDパッケージから独立して使われることも想定している．

```
{
  "targetDMD": "http://imi.go.jp/dmd/xxxxxxxx",
  "description": " マッピング例 ",
  "defaultVocab": "http://imi.go.jp/ns/core/2#",
  "mapping": {
    "名前": "人型>氏名>姓名",
    "フリガナ": "人型>氏名>姓名カナ表記",
    "性別": "人型>性別"
  }
}
```

4.4 IMI語彙記法

3章で設計した抽象的な概念としての語彙を、XMLやRDFといった特定の形式に依存しない中立的な形式で記述できると便利である。コア語彙2.3.2版までは表として語彙の定義を記述していたが、2.4版からは開発を進めているIMI語彙記法を試験的に用いて記述している。

この記法を用いることで、語彙定義やデータモデル記述をシンプルに行うとともにその機械処理が可能となる。XMLやRDFのスキーマファイル生成も容易となる。構造化項目名によって構造の断片が形式的に記述できるようになったのが、記述性の高さに大きく貢献した。以下は、人型の定義例である。

```
class ic: 人型 {@ic: 実体型};
use ic: 人型 >ic: 氏名 {0..n};
use ic: 人型 >ic: 性別 {0..1};
```

IMI語彙記法の基本構造は、ディレクティブとその値の対である。1行目はクラス宣言で、ic:人型はic:実体型を継承して定義することを表す。ic:はコア語彙の接頭辞である。2-3行目はクラスに対するプロパティ宣言で、氏名プロパティと性別プロパティの数制約を含んだ構造化項目名が値となっている。

IMI語彙記法は共通語彙の定義だけでなく、DMDの記述にも用いている。

4.5 実装での知見

実装レベルで大きな問題となったのは、XMLとRDFの差異である。コア語彙では当初名前空間に版番号を入れるようになっていた。名前空間はXMLの要素の集合を識別するために用いられるので、XMLでは重要となる。しかしRDFにおいては、名前空間が変わるとデータそのものが変わるので、互換性を保つためにはなるべく名前空間は変わらないほうが良いという、真逆の要求があった。

結論としては、コア語彙2.2からはXML用とRDF用で名前空間を別にして、XML用には版のある名前空間、RDF用には版のない名前空間を用いることにした。XMLは内部のデータベースなどで厳密に使われるであろうというのと、RDFは公開用に使われるであろうという想定のもとの決断である。一方で現在も2つの名前空間があるのはわかりにくいとの意見はいただいている。

また、各形式における制約も存在する。たとえば、XMLではプロパティに順序があるので姓名カナ表記の後に姓名を書けない。RDFには順序がないため姓名と姓名カナ表記を入れ替えて書ける。名前空間の別問題として、ドメイン変更問題があった。共通語彙基盤は経済産業省およびIPAのプロジェクトとして開始したため、imi.ipa.go.jpというドメインを利用していた。しかし、コア語彙2.3.2版では組織名を含まない純粋なサービス名のみからなるimi.go.jpドメインを併用することとなり、2.4版からはimi.go.jpへ移行した。これは長期安定性を目指した変更であったが、既存の公開データが新しいデータから孤立してしまうという問題を起こした。これに対して新旧両方のクラスを多重に継承してデータの互換性を維持するといった対応をとる利用者が現れている[7]。名前空間の永続性については初期の段階から特に注意深く考慮すべき課題である。

5. 共通語彙基盤の運用

5.1 支援ツール

共通語彙基盤の運用に重要なのが、語彙やDMDの作成や利用の支援である。共通語彙基盤では、支援ツールやデータベースの開発をプロジェクトとして進めている（図4）。現在はDMD作成・利用ツールである「表からデータモデル」を公開しており、150件以上のデータモデルの登録がある[8]。

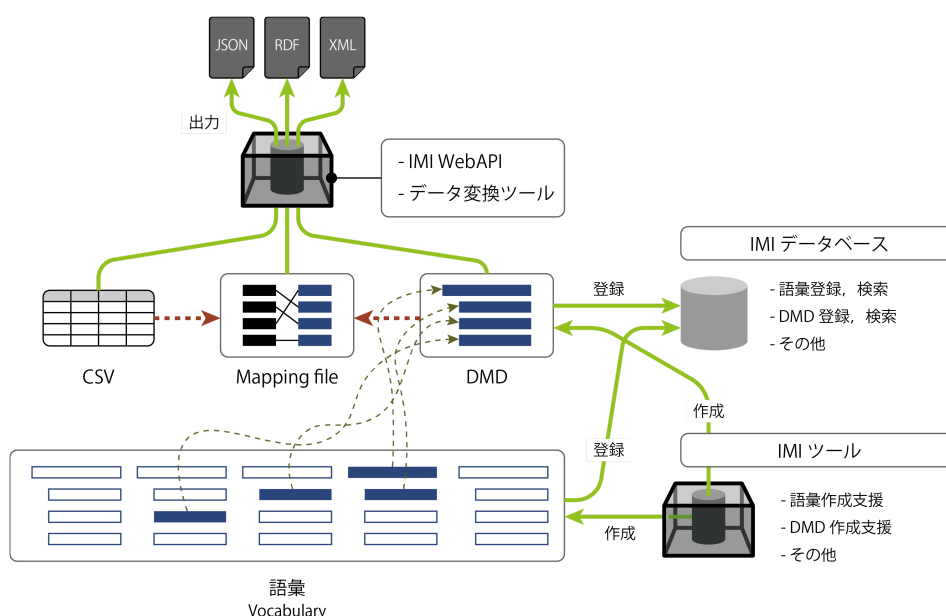


図4 共通語彙基盤プロジェクト

「表からデータモデル」では、作成公開されているDMDごとに対応する表形式データのテンプレートを提供する。そこへデータを埋めてアップロードすることで、DMDの中身を理解していなくても、データモデルに沿ったXMLやRDFのデータファイルを変換生成できる。

「表からデータモデル」では、1からDMDを作成するだけでなく、第三者が作成したDMDを改変して作成することができる。DMDが共有されはじめると、あるデータモデルが基本的には合っているので使いたいが、少しプロパティを足したいということがよく起きる。そのときに、DMDを改変することでニーズを満たせる。これを繰り返すと一時的に似たようなDMDが乱立してしまうが、その中でよく使われるものが出てくる可能性もあるし、すべてのニーズをまとめた決定版DMDが出てくる可能性もある。そのため、現在はDMDで共有する数を増やす段階と考えており乱立については気にしていない。

また、作成したDMDを登録・検索したり、DMDで用いた応用語彙を登録・検索するためのデータベースもプロジェクトとして開発中である。

5.2 パートナー制度

語彙開発やデータの整備を行っている団体と協力関係を結ぶための仕組みとして、共通語彙基盤ではIMIパートナー制度を設けている。現在8団体がIMIパートナーとなっている。また、官公庁や自治体は、宣言がなくともパートナーと同等の協調作業を進めている。

各IMIパートナーは、データ構築の現場で、そのデータの記述に必要な項目名の整理、コア語彙と関係付け、応用語彙としての新規語彙の作成などの作業を進めている。その過程で生み出された項目名の整理結果や検討中の語彙等は順次、公開ドラフトとして共通語彙基盤サイトから公開している。これは共通語彙基盤がこのような作業の過程で生じたノウハウの共有を含め、データの相互運用性を確保するために必要な基本情報を共有するための場として活用されることを目指しているからである。

図5にさまざまな語彙を示す。「語彙の素」は、項目名の整理結果等をまとめたドラフトであり、その構造を整理してスキーマなどが整ったものは共通語彙基盤に沿った「語彙」となる。語彙は最初は応用語彙として作られると想定されるが、公開ドラフト語彙として公開して共有し、利用者の意見を反映しつつドメインに共通な語彙等が整理統合されてゆく。さらにそのドメインの運用体制が確立されるとドメイン語彙として確立されることになる。

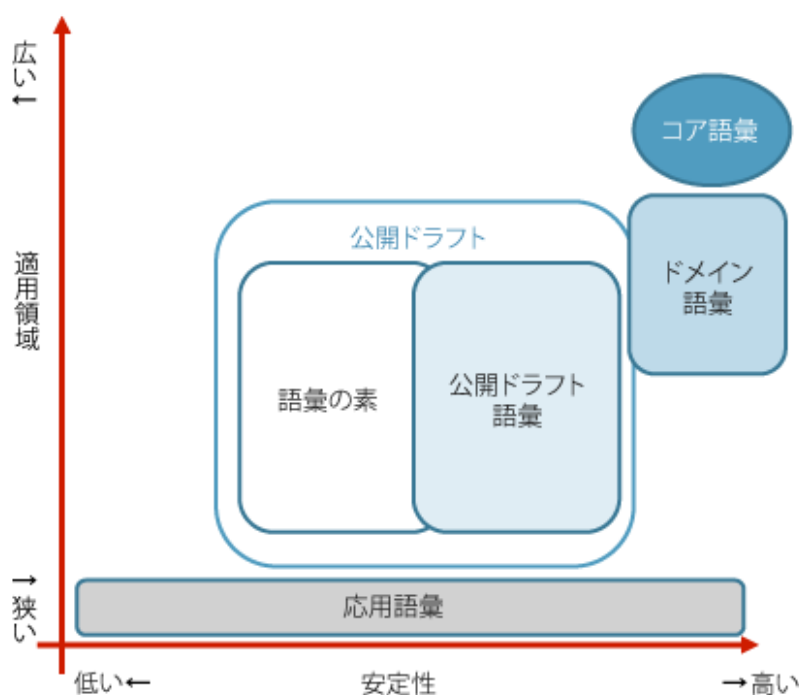


図5 さまざまな語彙

5.3 運用上の課題

運用上での課題の1つはコードである。共通語彙基盤ではコードの記述方法は提供しているが、その中身にはふれてこなかった。コードは、データを明示的に区別、同定、共通化していく上で重要な役割を担うが、管理主体が複数ある、版が複数ある、有料提供であるといった問題があり、共有して用いるのは意外と難しい。しかし、実際に運用しているとコード共通化の要望は多い。

そこで、共通語彙基盤側では都道府県コードのように広く使われる既存のコードについてのリストを提供することや、各コードについて具体的な記述例を提供することを検討している。また、外部で既存のコードを管理している団体と協力関係を築くことも同時に検討している。一方で、新たに作成して共有したいコードがある場合は、それをIMIデータベースで提供可能にすることも検討している。

6. 活用事例

共通語彙基盤は、トップダウンで設計するだけでなく、パートナー制度のように活用事例を増やすことに注力している。活用事例からのフィードバックによって共通語彙基盤も改良されている。本章では官公庁と自治体での事例を取り上げる。

6.1 法人インフォメーション

法人インフォメーション[9]（以下、法人インフォ）はIT戦略の下で経済産業省を中心として運用されているサイトである。法人登記されている約400万社を対象として、法人番号や法人名といった法人基本情報に加えて、府省との契約情報や表彰情報といった法人活動情報が一括で検索閲覧可能である。

法人インフォは共通語彙基盤 RDF版をベースに開発されており、SPARQLエンドポイントも提供している。法人インフォ用にコア語彙2.3.2を拡張して応用語彙が設計されている。コア語彙の法人型を継承して法人基本情報型を定義している。ここでは株主や決算情報、事業内容、業種コード等のプロパティが追加されている。さらに、法人基本情報型を継承した法人活動情報型も定義している。これは認定番号や認定日、部門、金額等のプロパティがある。

法人インフォで作られた応用語彙は共通語彙基盤にもフィードバックされており、公開ドラフトのPD2342 法人情報に関する語彙[10]として公開されている。また、法人インフォで用いられているデータモデルについても、法人基本情報DMDと法人活動情報DMDとして共通語彙基盤サイトで公開している[6]。これにより、第三者が法人インフォのデータモデルに合わせたデータやアプリケーションの作成が容易となる。

6.2 埼玉県

埼玉県は、2016年に県および県内市町村58団体で共通のデータを共通形式で公開するためのワーキンググループを立ち上げて、推進をしている[11]。2017年6月時点で58団体中42団体がすでに公開をしている。共通形式の提供は、公共施設やイベントカレンダーといった10種類に絞って実施している。

ここで言う共通形式とは、各データ項目から共通語彙へのマッピング関係を独自に表にしたものである。埼玉県で検討された10種類は、共通語彙基盤の公開ドラフトとして提供されている。公共施設一覧やイベントカレンダー等はコア語彙で十分にカバーできるが、たとえば広報紙については文書型にはない公開終了日が必要といったように細かいプロパティが足りない部分がある。コア語彙で足りない応用語彙部分は、明示的に語彙定義をしているわけではなく、公開ドラフトの範囲で独自作成したものを用いている。

埼玉県下の団体が共通形式で作成した表データは、埼玉県のオープンデータポータルに集約して公開している。更にそれらを埼玉県側が共通語彙基盤RDF版に合わせて変換しており、SPARQLエンドポイントとして公開している。これによって、一つのクエリで埼玉県下のご当地キャラデータを横断して問い合わせるといったことが可能となっている。

独自の応用語彙部分をRDF版として扱うためにはなんらかの名前空間が必要となるが、埼玉県側で用意するのが難しいという事情があった。これはほか事例でも発生することが想定されたため、公開ドラフトではドラフトごとに固有の名前空間を共通語彙基盤側から割り当てることで、その名前空間内では自由に用語を作れるということにした。これにより、データ作成のハードルが下がったといえる。

7. 関連事例

グローバルなデータ交換の機会が増えていることから、海外との相互運用性も考慮した仕組みにしていく必要がある。本章では、IMIとの間でデータモデルの相互マッピングの試作[12]に取り組んでいるNIEM、ISA²と、Web上のデータへの普及が進んでいるSchema.orgを事例として取り上げる。

7.1 NIEM

NIEM (National Information Exchange Model) [2]はアメリカ連邦政府が2005年より推進している、情報共有を円滑にするための共通モデルである。当時司法省で作成していたGJXDM (Global Justice XML Data Model) を基礎として、国土安全保障省と共同で作業を開始した。その後ほかの省も加わり、2017年現在にはNIEM 4.0を公開している。

NIEMはシステムレベルでの共通の辞書を果たすもので、個々のシステムから参照して使われることを想定している。NIEMのモデルは、人、場所、組織といった行政機関のどこでも利用する共通語彙をNIEM Coreとして、司法、入国管理、農業といった各分野ごとの共通語彙をNIEM Domainsと呼んでいる。NIEMはXML Schemaで設計されるが、抽象的な表現としてUML Profile[13]も作成している。

NIEMを用いて実際に交換可能な情報を作成するために、IEPD (Information Exchange Package Document) という形でXML Schema群を共有する。NIEMのモデルやIEPDといった考え方は、IMIでも大きな影響を受けている。また最近はJSON-LDベースでJSONによる表現を提供しはじめている。

7.2 ISA²

EU各国間での行政やビジネス、市民間の相互運用性を確保するために、欧州委員会では情報技術総局の下に2010年からISA、2016年から後継のISA²のプロジェクト組織を設置し、そこで相互運用性のためのフレームワークや戦略、構造等が検討されている。その中で情報共有のために進められているプラットフォームがJoinup[14]であり、その中のSEMIC (Semantic Interoperability Community) において、データ標準に関わるさまざまな検討がされている。

データモデルは、コアモデル、ドメインモデル、データ交換モデルの3層としており、コアモデルとドメインモデルがNIEMで言うコアとドメイン、データ交換モデルがIEPDに似た構成となっている。コア語彙についてはビジネス、場所、人等6つの語彙が定義されている。コア語彙はUMLと表による概念モデルと、XMLスキーマ、RDFスキーマを提供する。ドメイン語彙はまだないようである。

7.3 Schema.org

Web上で用いる構造化データの共通語彙として現在注目されているのがSchema.org[15]である。Schema.orgはGoogle, Microsoft, Yahoo, Yandexという検索大手4社によって2011年から開発されており、2015年からはSchema.org Community Group[16]で議論されているが、最終的な決定は4社の代表と貢献度の高いエキスパートからなる運営委員会が行う。

Schema.orgの主な目的は検索エンジンの結果を向上させるために、Web開発者が構造化データを提供しやすくすることである。語彙は英語で作成されており、比較的緩いデータモデル設計なのが特徴である。クラスは多重継承を許し、プロパティのドメインとレンジが複数指定されることもある。また、実際にはプロパティはどのクラスで使っても良いということにしている。プロパティのドメイン複数指定については共通語彙も影響を受けている。

もう1つ興味深い点は、コア語彙と拡張語彙の関係である。http://schema.orgで定義されている語彙はすべてコア語彙だが、コミュニティによって使われる語彙を拡張語彙として作ることができる。拡張語彙にはSchema.orgがホストする拡張と、外部の組織・団体が運営する拡張の2種類がある。前者には審査があり、通った拡張はhttp://bib.schema.orgのような情報提供用のサブドメインが与えられる。拡張語彙自体は、サブドメインではなくhttp://schema.orgとしてコア語彙と同じ空間で扱う。

一方で、schema.org外で拡張を公開して、コア語彙にリンクする方法もある。たとえば製品の外部拡張語彙としてGS1 Web Vocabulary[17]がある。拡張の考え方は、共通語彙基盤でも影響を受けている。

8. おわりに

IMI共通語彙基盤は全体の設計とコア語彙の設計実装が一段落して、法人インフォや埼玉県のように、さまざまなコア語彙および応用語彙の応用事例や、それに伴う公開ドラフトが出てきている段階である。官民データ法の推進に伴い、今後はより幅広い適用事例が増えていくと考えられる。ボトムアップ・トップダウン両方において、より詳細な語彙の共通化が行われていくことになることを期待している。

謝辞 IMI検討部会委員並びにIMIパートナーの皆様に深く感謝いたします。

参考文献

- 1) 首相官邸：世界最先端IT国家創造宣言・官民データ活用推進基本計画,
<http://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20170530/honbun.pdf> (2017年8月6日現在)
- 2) アメリカ連邦政府：NIEM, <https://www.niem.gov> (2017年8月9日現在)
- 3) Biron, V, P., Malhotra, A. : XML Schema Part 2 : Datatypes Second Edition, W3C Recommendation (2004).
- 4) Dublin Core Metadata Initiative : Dublin Core Metadata Element Set, Version 1.1, DCMI Recommendation (2012).
- 5) Clark, J., DeRose, S. : XML Path Language (XPath) Version 1.0, W3C Recommendation (1999).
- 6) DMD一覧, <http://imi.go.jp/dmd/> (2017年8月11日現在)
- 7) jig.jp : オープンデータプラットフォーム,
<https://odp.jig.jp> (2017年8月15日現在)
- 8) IPA : 表からデータモデル, <https://imi.go.jp/goi/dmd-editor.html> (2017年8月9日現在)
- 9) 内閣官房 情報通信技術総合戦略室, 経済産業省 : 法人インフォメーション, <http://hojin-info.go.jp> (2017年8月9日現在)
- 10) 法人情報に関する語彙,
PD2342, <http://imi.go.jp/pd/2342/index.html> (2017年8月9日現在)
- 11) 埼玉県 : 共通形式によるオープンデータの公開について,
<https://opendata.pref.saitama.lg.jp/events/news20170119.html>
(2017年8月9日現在)
- 12) SEMIC : Core Data Model Mapping Directory,
<http://mapping.semic.eu> (2015)

- 13) Object Management Group : UML Profile for NIEM Version 3.0, <http://www.omg.org/spec/NIEM-UML/3.0> (2017)
- 14) 欧州委員会 : Joinup, <https://joinup.ec.europa.eu> (2017年8月9日現在)
- 15) Google, Yahoo, Microsoft and Yandex : Schema.org, <http://schema.org> (2017年8月9日現在)
- 16) W3C : Schema.org Community Group, <https://www.w3.org/community/schemaorg/> (2017年8月9日現在)
- 17) GS1, GS1 Web Vocabulary, <https://www.gs1.org/voc/g> (2017年8月11日現在)

加藤 文彦 (非会員) fumi@nii.ac.jp

2004年慶應義塾大学大学院政策・メディア研究科修士課程修了。同年同大学院助手。2007年同大学院助教。(株)未来技術研究所,(共)情報・システム研究機構を経て2016年より国立情報学研究所特任研究員。LODやオープンサイエンスの研究開発に従事。IMI検討部会委員。

武田 英明 (正会員) takeda@nii.ac.jp

1986年東京大学工学部卒業。1988年同大学院工学系研究科修士課程修了。1991年同博士課程修了。工学博士。ノルウェー工科大学, 奈良先端科学技術大学院大学を経て2000年4月から国立情報学研究所助教授, 2003年より, 同教授, 総合研究大学院大学教授。2008年～2010年, 東京大学特任教授。知識共有, Web情報学, 設計学等の研究に従事。人工知能学会, 電子情報通信学会, 精密工学会, AAAI各会員。

田代 秀一 (正会員) tashiro@acm.org

1987年筑波大学大学院工学研究科博士課程修了。工学博士。通商産業省工業技術院電子技術総合研究所, 産業技術総合研究所を経て2011年より現職(独立行政法人情報処理推進機構参与/国際標準推進センター長)。ISO/IEC JTC1 SC2議長。

平本 健二（正会員） hiramoto-kenji@meti.go.jp

政府CIO上席補佐官。経済産業省CIO補佐官。デジタル技術による行政サービス改革を担当。国・自治体を通じた調達情報、支援制度情報総合サイトの構築・運用をするとともに、文字、語彙、コード等の基盤整備、Webサイトの見直し等、行政サービス改革を総合的に推進。本会シニア会員。

松澤 有三（非会員） yuzo@indigo.co.jp

2001年東京大学工学系研究科修士課程修了。2011年よりインディゴ（株）シームレス空間基盤研究開発センター主席研究員。地理空間情報およびLOD分野の研究開発・標準化活動に従事。IMI検討部会委員。

採録決定：2017年10月25日

編集担当：平林元明（（株）日立製作所）