データの共通理解推進ガイド

ー用語辞書や語彙を用いたデータの共通理解ー

2022 年 3 月 18 日 初版

独立行政法人情報処理推進機構(法人番号 5010005007126)

目次

はじめに

本ガイドについて

近年、社会を取り巻く環境は目まぐるしく変化しており、将来の予測が困難となっている。

デジタル化の進展やイノベーションの推進に伴うデータ量の増大、AI 能力の向上などを背景にした、急速なデジタル化の進展や高度化も、社会全般における顕著な環境変化の一つである。こうしたデジタル社会による急激な変化に対応するため、企業においても競争力の維持・強化に優位性を示しながらデジタルトランスフォーメーション(DX:Digital Transformation)[1]を迅速に進めることが求められている。

多くの国では、データは国の豊かさや国際競争力の基盤であると捉え、新たなデータ戦略を策定し、データ利用の官民協業、ルール形成など、自国の状況に応じた施策を講じながら、その取り組みを強力に推進し、多分野において各仕組みに導入している。

我が国においても、データは智恵・価値・競争力の源泉であると同時に、日本国内にある社会的な課題を解決する切り札として位置づけられており、その利活用と社会実装の推進を図るために、個人・民間企業・国家のニーズを踏まえた新たな価値の創出を目指す戦略に取り組んでいる。

具体的には、21 世紀のデジタル国家にふさわしい「デジタル基盤構築」に向け、デジタル庁にて、我が国初となる「データ戦略 （データ戦略タスクフォース 第一次とりまとめ[2]） 」において、既存のデータ利活用に関わる課題を抽出・整理し、その後、課題に取り組むための方向性を示す「包括的データ戦略[3]」を策定した。その中に織り込まれた基本行動指針(データ活用原則)の第一項目として「データがつながり、いつでも使える」を掲げ、データの相互運用性の向上をうたっている。

本ガイドは、データの相互運用性に関する理解を深めるための導入書であり、その目的は、デジタル社会の実現に向けて、データ活用を社会全体でさらに推進するための一助となることである。

データの相互運用の達成には、データを共有する関係者（作成者と利用者）の間で、そのデータに関する共通理解[4]を得ることが不可欠である。また、データの価値を最大化するためにも、共通理解を得るには、データが持ついくつかの側面から理解が求められる。「データ項目[5]自体が同じものとして認識できるか」、「デ

---

[1] 企業がビジネス環境の激しい変化に対応し、データとデジタル技術を活用して、顧客や社会のニーズを基に、製品やサービス、ビジネスモデルを変革するとともに、業務そのものや、組織、プロセス、企業文化・風土を変革し、競争上の優位性を確立すること。;『デジタルトランスフォーメーションを推進するためのガイドライン(DX 推進ガイドライン) Ver. 1.0』https://www.meti.go.jp/press/2018/12/20181212004/20181212004-1.pdf より。

[2] https://www.kantei.go.jp/jp/singi/it2/dgov/dai10/siryou_a.pdf

[3] https://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20210615/siryou11.pdf

[4] 共通理解: common understanding （以降、他の文献等を参照する際の利便性を考慮し、いくつかのキーワードについて英語表現を記載する。）

[5] データ項目: data element; data item ではなく data element としたのは、ISO/IEC11179 において、本ガイドのデータ項目に相当するものが data element としてモデル化されているため。

ータの表記[6]やコードの利用などが共有されているか」、「データの精度や更新頻度などが伝わっているか」などである。

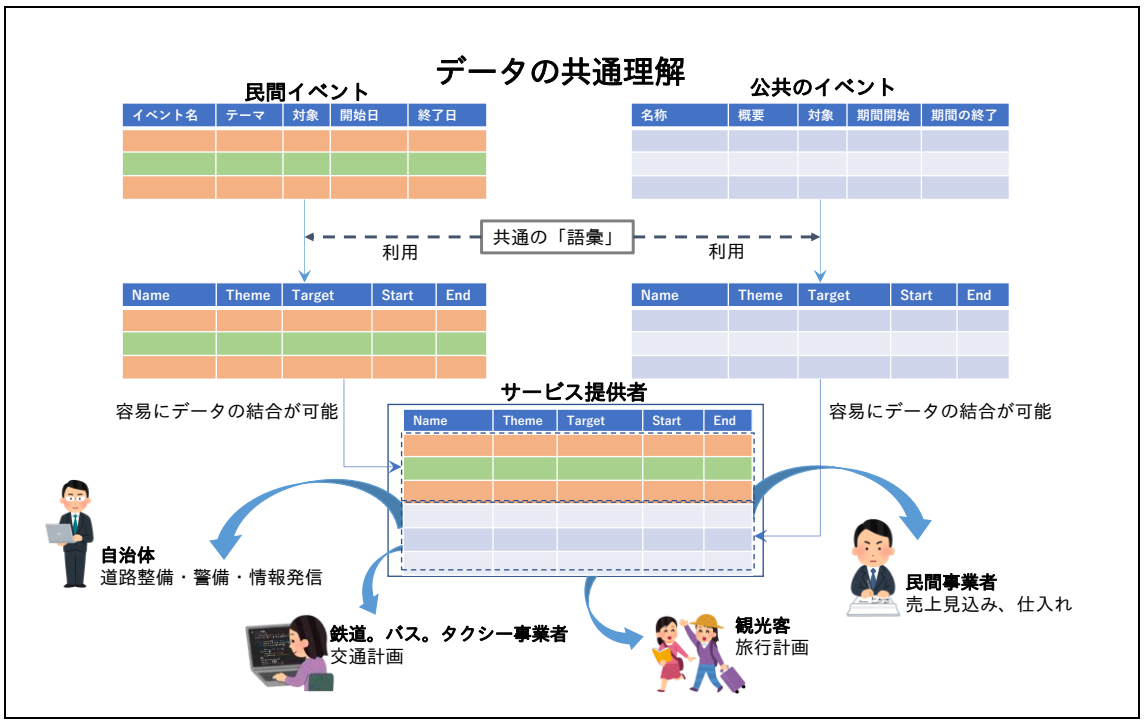

図 0-1 共通理解によるデータの価値向上

本ガイドでは、その中の「データ項目自体が同じものとして認識できるか」という部分を中心に解説し、データの共通理解がなぜ必要なのか、共通理解を得るための方法や用語[7]・語彙[8]の必要性、データの整備方法などについて、事例とともに具体的に提示しながら、データを共有することによって生まれるベネフィットについても説明していく。

ガイドの構成

第 1 章　データが持つ意味[9]

データの作成者と利用者が異なると、同じデータを使用しても、そのデータが待つ意味を取り違え想定通りに活用できないことがある。これはなぜ発生するのか。こうした意味の取り違えの発生を防ぐために、データが持つ意味を共通理解するとはどういうことか、また、共通理解を得るために利用する「用語辞書[10]」や「語彙」などについて解説する。

---

[6] 表記: designation; データ項目や値が指し示す概念をデータとして記述するために文字列で表すこと、または、表されたものの意。
[7] 用語: term
[8] 語彙: vocabulary
[9] 意味: meaning
[10] 用語辞書: terms dictionary

第 2 章　データが持つ意味を共通理解することで得られる効果

データが持つ意味を共通理解することで得られる効果について、事例をベースに紹介する。

第 3 章　語彙と用語辞書

データが持つ意味を共通理解するために、扱うデータを語彙や用語辞書などに関連づける方法が取られる。この語彙や用語辞書に関する整理方法や使い方について本ガイドとしての考え方を解説する。

ガイドの想定読者

データが持つ特性や意味を理解しその活用を図る立場にある方、行政や業界・団体・企業内の組織・業務を超えて相互運用を図る立場にある以下の方々を想定する。

- 業務システムの要件定義をする方
- DX を担当する方
- 行政機関でオープンデータを担当する方
- CDO (Chief Digital/Data Officer)を目指す方

第1章 データが持つ意味

本章では、複数の関係者間でデータを扱う際に意味の取り違えがなぜ発生するのか、その発生を防ぐためにデータが持つ意味を共通理解するとはどういうことか、共通理解を得るために利用する用語辞書や語彙などについて解説する。

1.1 データが持つ意味の観点

1.1.1 データの対象と表記について

本ガイドで扱うデータについて説明する。

データとは、コンピューターのファイルに記録され、コンピューターで処理できるもの[11]と本ガイドでは定義する。データ作成者である「主体」が「対象」を認識し、「主体」が理解できる「記号（文字、数値など）」でコンピューターに記したものである。

例として、倉庫に保管された商品の記録の事例を挙げる。

図 1-1 は、倉庫の商品(対象)を棚卸しする人(主体)が、商品(対象)を認識し、「在庫」という名称のデータセット[12]を作成し、商品の名称や数量といったデータ項目に在庫の現状をデータ(記号。ここでは値[13])として記録することを示している。

記録されるファイルは、スプレッドシートやリレーショナルデータベース（RDB）のデータ、スキーマが定義された XML や JSON など、一般的な形式で保存されたデータを想定している。

[11] メモ帳などに手書きした内容も認識した対象についての記述だが、本ガイドではコンピューターで処理できるデジタル情報として扱える状態のものを前提にする。

[12] 本ガイドでは、「データの集合、あるいはデータが埋め込まれた文書」として「データセット」を使用。「1.5.3 データの集合としてのデータセット」も参照のこと。

[13] 値: value

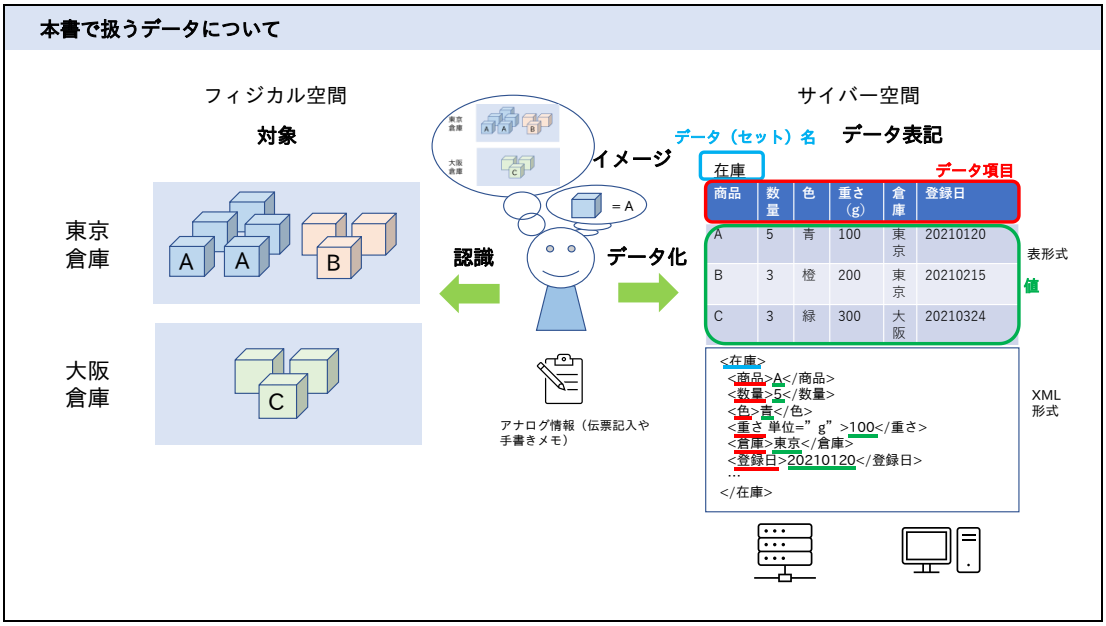

図 1-1 本ガイドで扱うデータについて

1.1.2 データの共通理解とは

データセットを受け取る担当者は、企業の場合、営業・生産・購買・物流など様々な部署の担当者が考えられる。さらに、サプライチェーンに含まれる組織外の取引先各部署の担当者が想定される。

異なる担当者が同じデータセットを見たとする。各担当者がある商品(例:商品 A や B)に関する知識を事前に習得しているとすれば、”商品”というデータ項目に対して”A”という値が表記されていると商品 A について共通した理解が可能である。つまり、思い浮かべる形状や色の濃淡などに、人によって若干の違いはあっても、おおむね同じ商品と認識する。図 1-2 はそのことを図示したものである。

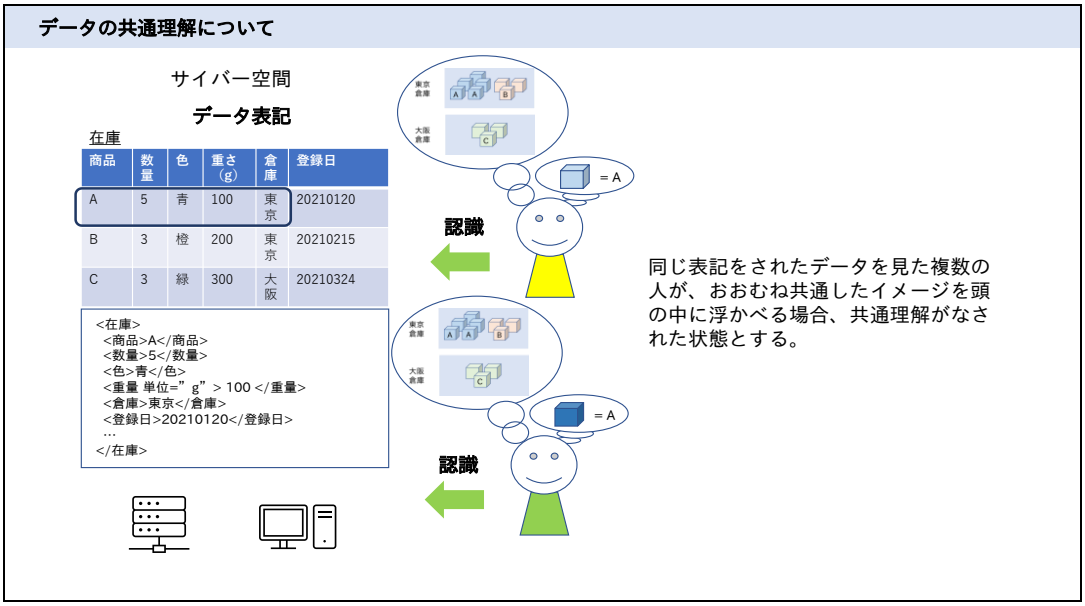

図 1-2 データの共通理解について

1.1.3 データの共通理解を阻むもの

(1) 各人が有する知識の差

共通理解を阻む原因の一つに、共有されたデータを使用する担当者により、その商品について持っている「知識の差」がある。

例えば、ある商品に関する同じデータを見た場合、商品知識が少ない研修中の新人、異動してから間もない現場経験が浅い担当者などは、その商品についてのイメージがそれぞれ異なる可能性がある。担当者によって商品の認識が異なることは、スムーズな顧客対応ができないなど、業務やサービスに支障が出ることがあるが、この対応の違いは、商品に関する知識の習得によって解消することができる。

(2) データに関する説明情報の欠落

共通理解を阻害する大きな要因として、データ項目や項目内に記載された値に関する説明情報が不足していることが挙げられる。この説明情報の不足は、関係者が共通した理解ができない、という問題を引き起こしている。ここで「説明」と言っているのは、データ項目や値がいったい「何を表したもの」なのかを、なるべく解釈の相違がない明確な方法で定義することを指す。

(3) データ項目や値の表記の揺れ

その他の要因として、データ項目や値の表記の揺れがある。

倉庫の商品を棚卸しする人が複数いれば、同じ対象(例:倉庫の保管商品)を認識しても、データセットを作成する際、 1 つの商品に対して異なる表記が複数存在するということがしばしば起きる。他には、担当者が商品 A の名称を”aaa”など当該企業や部署のみで通じる略語や業務上の「方言」に言い換えている場合もある。

データの定義者や作成者が、同じ対象を認識しているにも関わらず、データ項目や値の表記が揺れると、そのデータセットを受け取った側が頭の中にイメージするものが異なるというリスクが高まる。

図 1-3 は、2 人の担当者が同じ対象を見て、それぞれがデータを作成する際に、担当者によって表記が異なっており、”商品管理”と”在庫”や、”アイテム”と”商品”などと同義語が存在している状態を示している。また、データの値も「Blue」と「青」、「2021.1.20」と「20210120」のように異なっている。

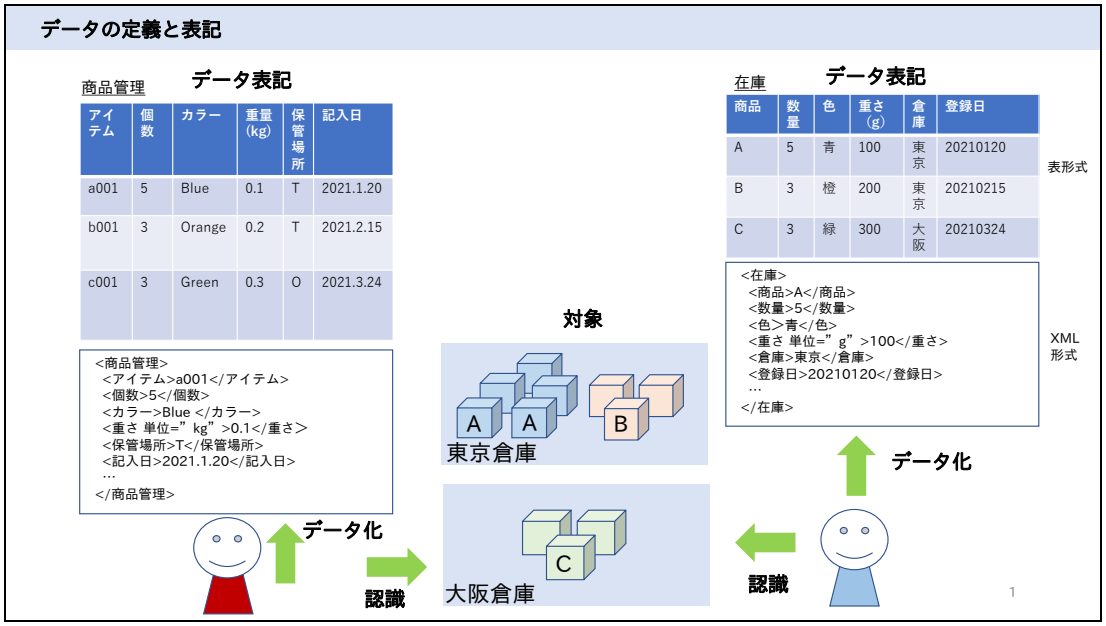

図 1-3 データの定義と表記

(4) コンピューター処理上の問題

コンピューターを用いたデータ処理は、人が共通理解を得るための助けになる。しかし、コンピューター処理の対象となるデータが、その説明情報との紐付けに不備などがある不適切なデータであった場合、共通理解が阻まれ、その結果、さらなるコンピューター処理がうまく行かなくなるという問題が生じる。

例えば、あるコンピューターで作成した、あるファイル形式で記述したデータや説明文書の記載の方法が関係者間で合意された標準的な形式に即していないことにより、別の環境におけるコンピューター処理の際に食い違いが生じ、処理の結果が誤った物(事故)となる場合がある。

(5) 事実との照合や確認が必要なデータの存在

データセットに含まれる値が、空欄になっている、または、間違っているなどのように、データに曖昧性がある場合、データの共通理解が損なわれる。

空欄になっている場合は、担当者の記入忘れなのか、それとも「ゼロ」(例:倉庫に当該商品が存在しない)という意味なのか、データセットを見ただけでは実際の状況がどうなっているのか分からない。

また、同じ名称でも異なる意味を持つ同音異義語だったり、氏名の場合には同姓同名の別人だったりするケースもあり、データセットの字面だけでは、作成の元になった対象が同一のものなのか、それとも別の異なるものなのかを判別できないことが少なくない。これも、部署を越えた横断的なデータの二次利用や共通理解を阻む要因になる。

不完全なデータが含まれるデータセットは、その作成者に確かめたり、現地の倉庫に問い合わせたりするなど、運用面でのフォローが求められる。

運用面での負荷を軽減するために、データの入力段階で空欄にするとエラーを通知してミスを防ぐ工夫や、

入力できる選択肢を絞り込んで勝手な値を入力しないようにする仕組みを、システム設計・開発の段階で組み込む対策が有効である。

1.1.4 データの共通理解を進めるために

データの共通理解を阻む課題がなぜ生じるのか、その理由について、さらに少し掘り下げて考えたい。

データ項目やその値に関する説明情報が欠けている場合にデータの共通理解が阻まれる大きな理由は、対象としてのデータセットに対して、人により思い浮かべるイメージが様々であるからである。

一方で、人の思考能力は、データセットのデータ項目や値に表記の揺れがあっても、頭の中で「名寄せ」を行い、作成の元になった対象が何であるか(どの商品か)を推測する柔軟性がある。とはいえ、個々人の推測が一致すれば共通理解は進むが、多少なりとも相違があると、かえって共通理解を阻害する。

例えば、異なる言語圏を流通する商品の名称や特徴などを記した多言語のデータを翻訳する場面では、表記が異なると、各担当者にかかる負担や労力が大きくなる。誤読によるトラブルが生じる可能性もある。無用なトラブルを避けるためにも、データセットを扱う相手の認識能力に依存し過ぎないよう、一般的な認識や表記に委ねることは極力しないことが肝要である。認識対象としてのデータ項目や値が何を表しているのかについて、誤解を避け、誰にとっても理解できるようにするためには、そのデータを共有する関係者間で、次のような取り組みが必要である。

① データ項目および値における表記の名寄せを行った上で、それらが同一の概念[14]を表していることを示すための整理を行う。

② ①の整理作業の成果物として、同義語リスト(用語表現のマッピング表)を作成する。

③ 概念を定義する説明情報を作成する。

人が頭の中に思い浮かべるイメージに名称を付してデータ化するとき、どんな同義語や略語を割り当てるのかについては、必ずしも正解はない。在庫または商品管理の例では、データ項目を”商品”とする人も、”アイテム”とする人もおり、異なる表記でも同じ対象として認識されている。ただし、同義語として名寄せができる場合は、同一の概念に基づいて行われた名寄せであるということができる。

③における定義とは、階層構造に基づく書式を用いながら参照関係などを表した概念を定義することを意味する。例えば、「商品の名称」という表記は、【商品】という概念が持つ【名称】という概念を属性として表現している。このような概念間の関係性を階層的な構造で示すと、図 1-4 のようになる。

なお、本ガイドでは、同一概念に対する様々な表記を意味する記述[15]として、その概念を代表する用語を【】(隅付き括弧)で囲んで記す。例えば、表記として「商品」「物品」「アイテム」「商物」などを持つ概念を表す

[14] 概念: concept

[15] 概念そのものを表すための記述と捉えることもできる。

際に、その代表用語[16]として「商品」を選択し、【商品】と記述する。

```
【商品】
 |____【品番】
 |____【名称】
 |____【製造年月日】
 |____【……】
```

図 1-4 概念の階層構造

この概念の定義を基に、XML の記法で”A”という値データを表記すると、以下のように記述できる。

```
<名称>A</名称>
```

図 1-5 XML の記法によるデータの表記

このように、概念および表記の集合を管理し、データの共通理解のために関係者で相互に参照できるように開示する環境が必要である。この管理された概念および表記の説明情報を、本ガイドでは「語彙[17]」と呼ぶ。

データの共通理解は、「語彙」と「語彙に基づくデータ」を組み合わせながらデータの受渡しを行うことにより促進される。

道具としてのコンピューターを使用しながら、効率よく、かつ、関係者間で共有するデータについて生じがちな誤解を最小限にする状態を作り出すことは、１つの組織内にとどまらず、社会全般においても様々なメリットが生まれる。そのメリットとして、円滑なコミュニケーションと業務の推進、多様なデータを利用した社会課題の解決、また、新たなサービスの連携や事業・産業の創出等が期待される。

[16]「1.3.2 概念自体や概念間の関係を整理するための情報～用語辞書の作成～」も参照のこと。

[17] 一般的に用いられる「語彙」の意味とは異なることに留意。「第 3 章 語彙と用語辞書」も参照のこと。

1.2 データが持つ意味を表現する要素

何らかのデータ項目で構成されるデータセットが存在するとき、その共通理解を図るための観点として、次の 2 つがある。

- データセットを構成するデータ項目と値の説明
- データセット全体を外から見たときの説明

以下、詳細に説明する。

(1) データセットを構成するデータ項目と値の説明

データセットに含まれる個々のデータが持つ意味を、データを使用する関係者が共通して理解するためには、データ項目やその値に関する説明情報[18] が必要である。説明情報に求められるのは、データ項目や値が指し示すものが何であるのかの情報、また、それらがデータ利用者の業務の中でどのような関係や位置づけを持つのかを認識・整理するための情報、の 2 つである。

言い換えると、データを共有する関係者が、データセットに含まれる個々のデータが持つ意味を共通して理解するために必要な情報には、次の 3 つの要素があると言える。

- データが表しているもの(概念)の定義情報
- 概念間の関係を整理するための情報
- データ項目の値に関する情報

以下に、詳細を解説する。

①データが表しているもの(概念)の定義情報

本ガイドでは、「データが表しているもの(概念)の定義情報」を、データ項目と値の組、あるいは、その集まりが持っている情報としての定義と考える。あるモノやコトの概念をデータで表現しようとすると、その概念は一つのデータ項目とその値として表現される、あるいは、複数のデータ項目とその値で表現されると考える。これが「1.1 データが持つ意味の観点」でも述べた「語彙」の実体である。

この定義情報の一つに、概念の表記がある。データ項目や値など、データが表しているもの(概念)は、データになった時点で文字列として表記される。本ガイドでは、この「概念」をデータ化したときの文字列表記を「用語[19]」と呼ぶ。

②概念間の関係を整理するための情報

概念間に存在する様々な関係を定義する情報が、「概念間の関係を整理するための情報」である。この概念間の関係は、データ項目と値のそれぞれに存在する。

---

[18] ここで「説明」と言っているのは、データ項目や値が、いったい“何”を表したものなのかを解釈の相違がないような方法で明確に定義するという意味である。

[19]「3.1 本ガイドにおける考え方」も参照のこと。

データ項目間の関係としては、あるデータ項目に対して、そのデータ項目を説明するための詳細化したデータ項目がある、といった階層関係がある。例えば、【会社】というデータ項目は、一つの【会社名】というデータ項目と複数の【部署】というデータ項目から構成されるという関係があり、さらに【部署】は一つの【部署名】というデータ項目と複数の【社員】から構成されるという関係によって、【会社】というデータ項目に関する階層的な関係が構築される。

値の間の関係としては、ある概念が他の概念を包含するという包含関係がある。例えば、【動物】という概念には【犬】や【猫】という複数の概念が含まれている。

この【会社】【会社名】【部署】や【動物】【犬】【猫】といった概念の関連性を定義した情報が、概念間の関係性を整理するための情報である。

③データ項目の値に関する情報

データ項目の値として、入れることができるものに関する情報が、「データ項目の値に関する情報」である。コンピューターを用いたデータ処理で一般に使われる文字列や数値に加え、①と②で整理した「値」に関する概念をリストとして用意して使うこともできる。

本ガイドでは、「人が識別するために、ある対象を何らかの文字列で記したもの」を「表記」と呼ぶ。したがって、「用語」は「概念」の一つの「表記」である。また、本ガイドで考える「概念」はデータ項目と値の両方を表すものであるため、「用語」には、データ項目としての用語と、値としての用語という、2 種類の用語があることになる。

これらを【都道府県】を例に、以下に説明する。

出身地を表すデータ項目に【都道府県】という概念を用いることを考える。【都道府県】には”都道府県”という表記の用語や”prefecture” など別の表記の用語もある。また、【都道府県】という概念をデータ項目とすれば、その値として【東京都】、【大阪府】、【北海道】、【神奈川県】…という概念が割り当てられ、値の表記には”東京都”、”大阪府”、”北海道”、”神奈川県”…という値としての用語を使うことになる(他に都道府県コードを使った”13”、”27”、”01”、”14”…など)。

データ項目の「概念」は、それ自体の表記としてどのような用語を使うのかだけでなく、どのような値の用語を割り当てるかという組合せを決めることになる。本ガイドでは、これを「概念」の「表現」と呼ぶ。

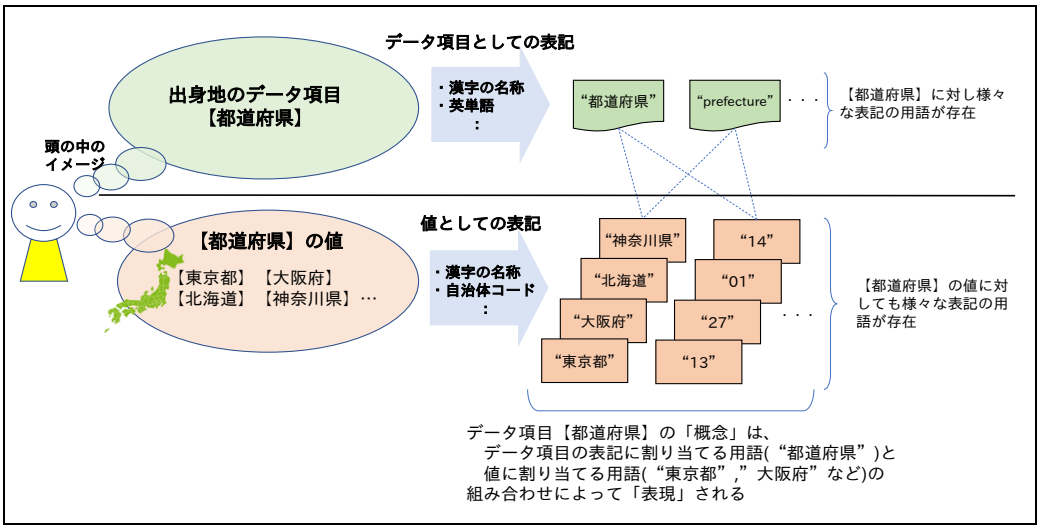

図 1-6【都道府県】を例とした概念の表現

データが表しているもの(概念)は、語彙としての表記は統一されるが、既存のデータセットや文書では、その「概念」に対する呼び方や言い方が様々であることから、表記も様々に使われている。これらは同一物(同一の概念)に対して存在する複数の「同義語」とみなされる。

ここで、同一物であるかどうかを特定するには、同義語を整理してリスト化した情報が必要である。これを、本ガイドでは「用語辞書」と呼ぶ。

(2) データセット全体を外から見たときの説明

データセット全体を外から見たときの説明とは、データセットが全体として何を記述したものであるかを説明する情報のことである。

この情報から、データセットの利用者がそのデータセットについての共通理解を得ることができる。この情報は、図書館の書誌情報[20]と同様、データセットの検索に使うこともできる。

[20] 書誌情報: bibliographic information

1.3 データセット内のデータ構成

1.3.1 データが表しているもの(概念)の定義情報～語彙の構築～

データの共通理解の基として、データが表しているもの(概念)を定義した語彙を作成することが求められる。語彙を作成するためには、データに表現されている様々な用語が同一物であることを特定するための整理作業を行う必要がある。ここでは、先に最終的に必要な語彙の構築について説明を行い、用語の整理については後述する。

データ項目やその値が表している「概念」について、それが何であるかを明確に説明する方法としてまず考えられるのは、自然言語を使用した文としての「定義」を与えることである。それに加え、「概念」をデータとして表現し解釈する場合には、辞書的な説明だけではなく、データ項目と値の対応関係、データ項目の成り立ちや由来などの構造情報も定義する必要がある。

また、個々のデータ項目とその値について構造情報を含めて定義するとともに、それらを集合として管理し、参照できるようにしたものが「語彙」として使われる。

データ項目の成り立ち(由来)を定義する例として【人の居住国】というデータ項目を考えると、定義としては「【人の居住国】というデータ項目は、【人】というものが持つ【居住国】という属性(プロパティ)を表現したものである」などが考えられる。

このように定義することで、コンピューターはデータ項目としてファイルに記述されたものを誤解なく認識できる。つまり、概念の定義をすることで、共通理解できるようになる。

なお、データ項目やその値は、具体的には何らかの記法を使って表記される。例えば、XML という記法では、<名前>や<住所>というタグと呼ばれるもので値を挟んで下記のように記述する。また、それらが人に関する情報ということで、<人>というタグでそれらを囲み、図 1-7 のようにデータ項目の階層を作ることもできる。

```
<人>
  <名前>〇山口男</名前>
  <住所>東京都文京区・・・</住所>
  ・・
</人>
```

図 1-7 XML 記法で記述した人に関する情報

このような、データ項目、値、データ項目間の階層構造などは、それらを表記するデータ記法(例えば XML)ごとに定義されることになり、スキーマ定義言語と呼ばれるものを使って定義を行う(XML 用のスキーマ定義言語としては、XML Schema[21]がある)。

[21] https://www.w3.org/XML/Schema

■コラム:データ項目や値などの定義情報を登録するための国際標準 ISO/IEC 11179

データ項目およびデータ項目間の構造、そして値を定義した情報を登録する汎用的な仕組みとして、『ISO/IEC 11179 Metadata registries (MDR)』という国際標準がある。これは複数パートからなる標準規格であり、Part3 が『JIS X 4181-3:2004 メタデータ登録簿(MDR)－第３部:登録簿メタモデル及び基本属性』として JIS 規格にもなっている。

ISO/IEC 11179 は、名称に含まれる「registries(登録簿)」という言葉からも分かるように、データ項目(ISO/IEC 11179 では「データ要素」)や値をデータベースに登録して管理することを目的としている。また、ISO/IEC 11179 は、XML Schema などと同列のスキーマ定義言語ではなく、任意のスキーマ定義言語を使って定義されたデータ項目などの定義情報を参照し、より汎用的に表現する仕組みとなっている

1.3.2 概念自体や概念間の関係を整理するための情報～用語辞書の作成～

前述したような方法で「概念」の定義が行われていても、実際のデータ交換においてデータの共通理解が阻まれることがある。その要因の一つに、一つの「概念」に対して多様な表現が使われていることがある。同一のモノ(概念)に対する表現の違いの例として、正式名称、略語、同義語、異なる言語での表記、などが挙げられる。

これは、「概念」を定義して語彙を作成する前に、データ交換を行いたい業界や組織に存在する様々なデータや文書を分析し「同一のモノ(概念)を表現していると思われるものを名寄せする」という整理作業が必要であることを示している。その作業の結果として同義語リスト(用語表現のマッピング表)が出来上がり、それが「用語辞書」になる。

用語辞書には、名寄せした複数の用語表現の中から、対応する概念を表す「代表用語」を決めて記述することになる。なお、その他は「同義語」として記述されることになるが、具体的にそれらをどのように構成するかは本ガイドでは定めない。

1.3.3 語彙作成の意図と用語辞書の役割

データ項目や値の意味(それが何を表しているか)を誰もが誤解なく認識する(共通理解する)には、用語表現の統一が必要である。

語彙は、同一のモノやコト(概念)に対して、一つの代表用語を使って定義されたものである。語彙を作成し、それを使うということは、用語表現が統一されているという意味を持つ。言い換えると、語彙はデータの共通理解のために用語表現を統一することを意図して作成される。

語彙が出来上がった後は、データを活用する当事者がその用語を使ってデータを作り、読むことになるため、全てのデータにおいて用語の揺れがない状態で解釈できる。

既存のデータは、語彙作成のための名寄せを行う前の様々な用語を含んでいるため、そのデータを語彙に基づくデータに移行する際には用語辞書を利用する。名寄せを行う前の用語に対し、語彙として定義した名寄せを行った後の用語を用語辞書から見つけ、マッピングや書換え作業を行うことができる。

このように用語統一されたデータを使いながらデータ交換を行うことにより、データの共通理解を図るというのが、語彙作成の意義である。

ただし、一つの語彙を作成するだけで世界のデータ全てが記述可能になるわけではない。データ項目にしても、値にしても、同義語を使って作成される別の語彙が作成されていることが考えられる。このような場合にデータの共通理解を得るためには、語彙間のマッピングを取る必要がある。マッピング作業の結果として、「語彙間の同義語リスト」としての用語辞書も出来上がる。これらの用語辞書は、異なる分野・業種の語彙を使って書かれたデータを交換する場合に、他の分野・業種のデータ項目や値の意味について、人が理解することを可能にし、かつ、コンピューターが処理するためにも使用することができる。

1.3.4 語彙利用のユースケース

語彙が整備されていた場合、その利用方法は様々である。ここでは、語彙の利用形態についていくつかのパターンを示す。

最初に、語彙を使わない場合の組織(A,B)間のデータ交換について図 1-8 に示す。

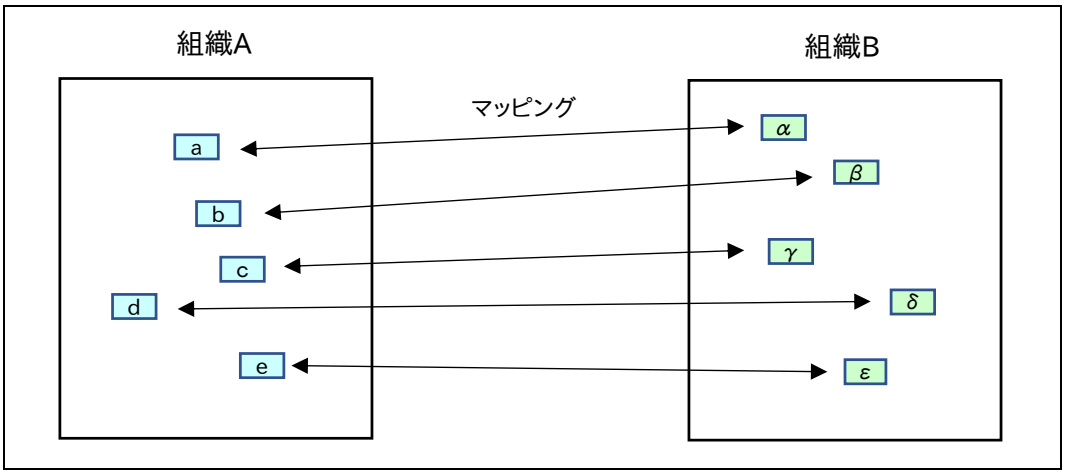

図 1-8 個別にマッピングを取るパターン

図 1-8 では、組織 A のデータ項目名”a”と、組織 B のデータ項目名”α”は、同じ意味を持つものとする。この場合、異なるデータ項目名のマッピングを取るために用語辞書の同義語リストを使うことが考えられる。

次は、関係する全ての組織が共通の語彙を使ってデータを運用するパターンである。

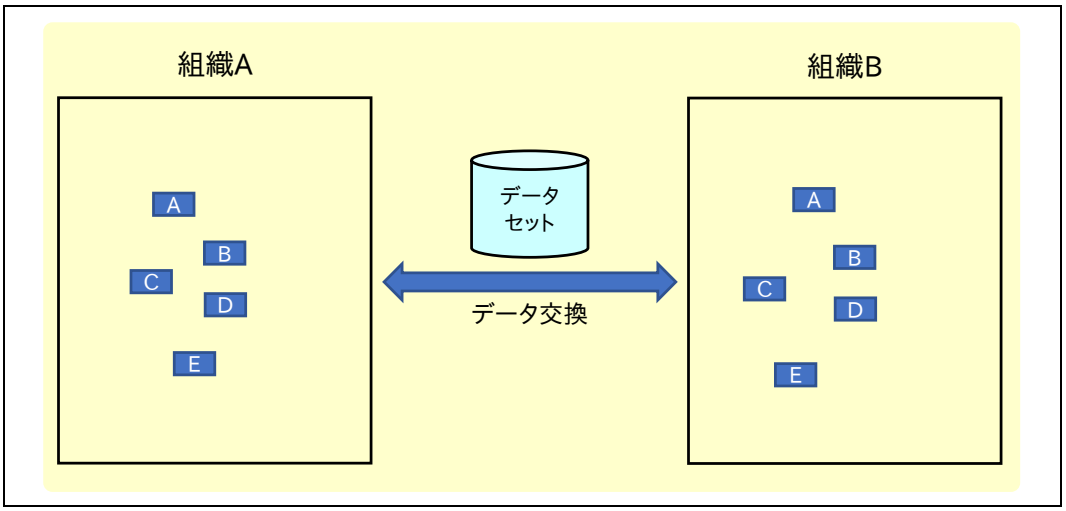

図 1-9 共通の語彙をあらゆる組織内で使うパターン

図 1-9 のように共通の語彙を項目名に使用して運用すれば、完全にデータの相互理解を図ることができる。これは、最も理想的な利用パターンであるが、何年にも渡って蓄積されたデータの全てを書き換えるのは困難な場合もある。その際は、図 1-10 のように、データを外に出すときと受け取るとき、つまり、データ交換の時点でのみ共通の語彙を使うというパターンが考えられる。

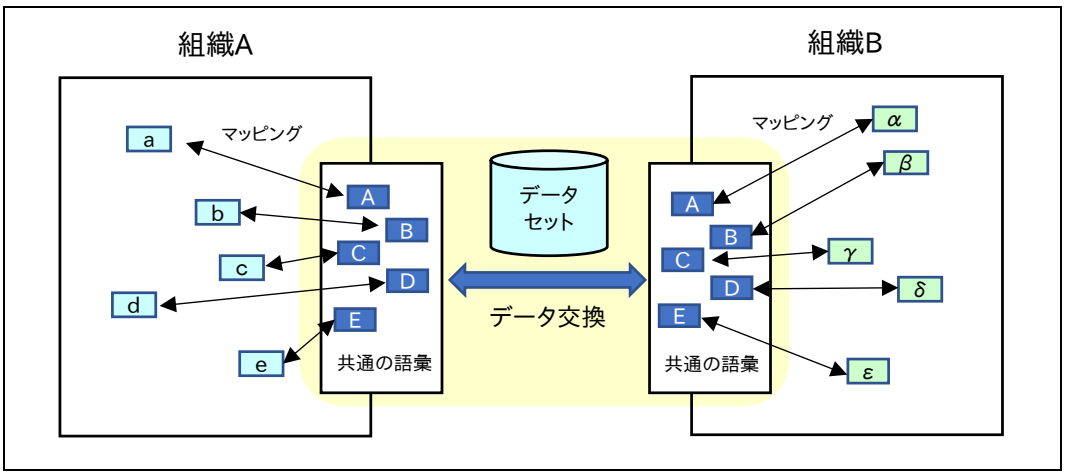

図 1-10 交換時のみ共通の語彙を使うパターン

この場合、データの送信時、受信時に、それぞれの独自データ項目名と共通語彙のデータ項目名のマッピングにより変換する仕組みを作る必要がある。

1.4 データセット全体を外から見たときの説明

1.4.1 データセットを説明する情報の付与

データの共通理解を得るための別の観点として、そのデータが全体として「何を表すものか」「どのように作成されたのか」などの説明を添付し、それを適正に活用することができるような仕組みを作ることが挙げられる。そのような説明情報を本ガイドでは「書誌情報」と呼ぶ。

「書誌情報」を構成する基本的な項目として、以下のようなものが考えられる。

・表題
・内容説明
・作成者
・日付

このような説明の対象となるのは、データセットや文書など、語彙に基づくデータの集合体であり、コンピューター上では物理的に一つのファイルで表されることが多い。

「書誌情報」に書かれた説明を読むことによって、データセットや文書についての理解を得ることができる。また、このような情報がデータカタログとして記述されて参照できるようになっていれば、求めるデータセットや文書を見つけ出す際にも役立つ。

「書誌情報」を活用してデータの共通理解につなげるためには、「書誌情報」の形式(データ項目や内容)がデータを共有する関係者に分かる方法で整備されている必要がある。これは、「書誌情報」のデータ項目や値も、語彙の実現例の一つであるということを示している。

「書誌情報」を表すデータ項目に関する既存規格として、「ダブリン・コア(Dublin Core)」があるが、このような既存の規格にあるデータ項目を基準とし、必要に応じて、独自のデータ項目を追加するなどの拡張を行って「書誌情報」の書き方を定めることが推奨される。

なお、「書誌情報」は対象となるデータセットや文書とは別に作成することが多いが、データセットや文書の先頭などに記述するという方法もあり、「書誌情報」の定義方法は様々である。

■コラム:「書誌情報」を表すデータ項目集 「ダブリン・コア(Dublin Core)」

「書誌情報」を表すデータ項目集(語彙)として、技術標準「ダブリン・コア」が存在する。

これは、「The Dublin Core Metadata Initiative (DCMI)」(https://dublincore.org/)によって開発され、ISO の国際標準、および、JIS 規格(日本産業規格)としても定められている。

ダブリン・コアでは、まず基本的な 15 個のデータ項目を「ダブリンコアメタデータ基本記述要素集合

(DCMES:Dublin Core Metadata Element Set)」として定め、その後、それを拡張した「DCMI メタデータ語彙(DCMI Metadata Terms)」を定めた。これらは、現在の ISO15836 の Part1、Part2 となっている。

● 国際標準

・ISO 15836-1:2017 The Dublin Core metadata element set — Part 1: Core elements
・ISO 15836-2:2019 The Dublin Core metadata element set — Part 2: DCMI Properties and classes

● JIS 規格

・JIS X 0836:2005 ダブリンコアメタデータ基本記述要素集合(旧版の ISO 15836 に基づいたもの)

ダブリン・コアによって、書誌情報を表現するデータ項目の語彙がある程度決まっているため、書誌情報については共通理解が図られやすい状況にあると言える。

1.4.2 データカタログを記述するための語彙

データカタログとは、様々なデータセットの書誌情報を集め、データセットのカタログとしてウェブサイトなどに公開することにより、求めるデータセットを見つけて利用しやすくするための仕組みである。

データカタログの利便性を高めるため、W3C[22]では、ウェブで公開されるデータカタログ間の相互運用性を促進するためのデータ項目を標準化した「Data Catalog Vocabulary (DCAT)」を策定している。標準化されたモデルと用語を使うことによって、複数のカタログサイトの書誌情報を集めることができるようになり、複数のカタログサイトを横断してデータセットを検索でき、求めるデータセットが見つけやすくなる。

DCAT の仕様は、2014 年に初版[23]が公開され、その後新たなユースケースなどに対応した第 2 版[24]が 2020 年に公開されている。

DCAT の初版の仕様では、データカタログを記述する以下の 7 つのクラスを規定している。

・Catalog
・Catalog record
・Dataset
・Distribution
・Concept scheme
・Concept
・Organization/Person

「Catalog」クラスを使って、様々なデータセットを登録するカタログを表し、個々のデータセットは「Dataset」クラスを使って表す。

---

[22] World Wide Web Consortium の略称。
[23] Data Catalog Vocabulary (DCAT); https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
[24] Data Catalog Vocabulary (DCAT) - Version 2; https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/

「Dataset」クラスは、下記のようなプロパティを持っており、それによってデータセットの書誌情報を記述する。

・表題(dct:title)
・内容説明(dct:description)
・発行日(dct:issued)
・使用言語(dct:language)
・発行者(dct:publisher)
・テーマ(dcat:theme)
・キーワード(dcat:keyword)
・　：
・　：

ここで、「dct:*****」とあるのは、前述の「ダブリン・コア」が規定した書誌情報記述のデータ項目を表し、「dcat:*****」とあるのは DCAT の仕様で規定したデータ項目を表す。

DCAT は、ウェブにおけるデータセットの検索性を向上させることを目的の一つとしているため、データセットの配信に関する「Distribution」クラスが規定されている。これは、データセットをどのように配信するか(ユーザーが入手する方法)について、ライセンス情報、入手先 URL、データ形式、データサイズ、などの情報を記述するためのものである。

このように、DCAT は、ウェブに存在するデータセットの書誌情報や入手方法を記述するデータ項目を定め、それを使って記述したデータセットの情報を集めてカタログとして公開するための仕様である。なお、「一般社団法人　データ社会推進協議会[25]」では、DCAT をベースとしたデータカタログを構築するための「データカタログ作成ガイドライン V2.1[26]」を公開している。

DCAT のデータ項目を使ったデータカタログの例としては、「e-Stat 統計 LOD[27]」がある。

[25] https://data-society-alliance.org/
[26] https://data-society-alliance.org/survey-research/datacatalogguideline/
[27] https://data.e-stat.go.jp/lodw/

1.5 データが持つ意味を共通理解するための考え方

データが持つ意味について明確に定義された状態を前提とした上で、データの意味について「共通理解する」に至るまでに何が必要かをここで考える。

1.5.1 データの共通理解のために必要な語彙の共有

せっかく語彙(語彙体系)を作っても、限られた人の範囲での使用にとどまってしまうと、その意味をより多くの人に理解してはもらえない。データの「共通理解」といった場合、前提となる環境として、複数の利用者・ステークホルダーが存在している。

そのような環境には、例えば、ビジネス観点の環境であれば、「一つの会社/組織」、「一つの業種」、また、地理・法制度観点の環境では、「自治体などの一つの地域」、「一つの国」、「世界」など、様々なレベルが考えられる。

このように、目的に応じた環境を想定し、できるだけ多くの人が同じ「語彙(語彙体系)」を共有しながら使うことができるようにすることが、データの共通理解につながる。(どの範囲・レベルで語彙を共有するかは、語彙を利用する目的によって異なる)

1.5.2 語彙に基づくデータの作成

データの共通理解とは、人がデータを見てそこに書かれたデータ項目の意味を理解する、あるいはコンピューターがデータを処理してデータ項目の意味に合った(プログラムが想定した)処理を行うことを指している。

この実現の一歩として、語彙として明確に定義されたデータ項目を利用してデータを記述することが挙げられる。例えば、「人の居住国」というデータ項目が定義されていれば、XML 記法を用いて以下のような形式でタグを付けることにより、「その人の居住国は日本である」という意味を持つデータが出来上がる。

```
<人の居住国>日本</人の居住国>
```

図 1-11 XML 記法を用いたタグ付きデータ

データ交換において、データの送り手は、上記のように何らかの記法で語彙に従ったデータを作成し、データを送る。データの受け手は、そのデータがどのような語彙を使っているかを知ることができれば、その語彙の定義を参照することによって、データを作成者の意図通りに理解・処理することができる。これがデータの共通理解である。

データの共通理解は、「語彙」と「語彙に基づくデータ」によって実現することができる。

1.5.3 データの集合としてのデータセット

データセットは、データの集合、あるいはデータが埋め込まれた文書として存在し、語彙に基づくデータ（データ項目と値）で構成される。これが関係者間におけるデータ交換の対象となる。

前述したように、データセットは、その内容としてのデータに語彙を使用することによって共通理解が図られる。これに加え、「1.4 データセット全体を外から見たときの説明」で述べたように、データセットが何を表したデータなのかを記述した書誌情報があれば、データセットをより理解しやすくなる。

第2章 データが持つ意味を共通理解することで得られる効果

データ定義者とデータ作成者、さらにデータの(二次)利用者の間でデータの説明情報を共有できれば作成・流通するデータの品質が上がる、というシナリオが、意味を共通理解する効果に挙げられる。本章では、データが持つ意味を共通理解するために、「語彙」と「語彙に基づくデータ」を組み合わせて受渡しすることで得られる効果について、事例をベースに紹介する。

2.1 標準化された語彙の活用による効果

標準化された語彙の活用事例として、料理レシピ動画サイト「DELISH KITCHEN」を運営するエブリー社[28]のケースを紹介する。エブリー社では、ウェブサイト訪問者であるユーザーが、探したいレシピ情報にストレスなく検索できるように、レシピ情報のデータ項目や書式を Schema.org に準拠して整理し、構造化データとして整備している。

(1) 取り組みの背景

多くのユーザーが利用する Google での検索結果をより効果的に表示する SEO[29]対策として、現場部門の主導で構造化データの整備を推進した。構造化データの書式については Google 公式ページに開示されているリッチリザルトの情報や事例を参照している。

(2) 構造化データの仕組み

Google 検索では、検索結果の表示において対象ページのコンテンツ作成者の意図を伝える明示的な手がかりとして Schema.org のスキーマに沿った構造化データの利用を推奨している。構造化データを用いることで、ページに関する情報をより明示的に提供することができる。

例えば、レシピに関するページの場合であれば材料、加熱時間と加熱温度、カロリーなどを分類するために標準化されたデータ形式を利用して構造化データを提供する。検索エンジンは、ウェブ上で検出した構造化データに基づいてページのコンテンツを分類、整理している。

下の表はパーティ向けコーヒーケーキのレシピの一部抜粋である。1 行目のデータ項目名は、データ作成日(dataPublished)であり、作成日(実データ)は 2018 年 3 月 10 日であることを示している。日付は「yyyy-mm-dd」という書式にそろえている。

---

[28] 株式会社エブリー; https://corp.every.tv/

[29] Search Engine Optimization(検索エンジン最適化)の略。検索サイトにおいて自サイトに有利な検索結果を導くための各種調整。

```
"datePublished": "2018-03-10",
    "description": "This coffee cake is awesome and perfect for parties.",
    "prepTime": "PT20M",
    "cookTime": "PT30M",
    "totalTime": "PT50M",
    "keywords": "cake for a party, coffee",
    "recipeYield": "10",
    "recipeCategory": "Dessert",
    "recipeCuisine": "American",
    "nutrition": {
      "@type": "NutritionInformation",
      "calories": "270 calories"
    },
    "recipeIngredient": [
      "2 cups of flour",
      "3/4 cup white sugar",
      "2 teaspoons baking powder",
      "1/2 teaspoon salt",
      "1/2 cup butter",
      "2 eggs",
      "3/4 cup milk"
    ],
```

”データ項目”は正体文字、”*入力値(実データ)*”は斜体文字で示す。

図 2-1 Google 検索用構造化データの例

（パーティ向けコーヒーケーキのレシピの一部抜粋。実装形式は JSON）

出典:https://developers.google.com/search/docs/advanced/structured-data/recipe

ユーザーに対し知りたいレシピの検索を円滑に行わせるためには、大きく次の 2 点が重要である。

①レシピを集積したサイトに誘導できること

②知りたいキーワードでヒットさせること

①はデータ項目の定義が重要になる。エブリー社ではレシピのマスターデータを検索用構造化データとして利用できるように管理、整備している。レシピのデータについては、関係者は常に項目定義が参照できる状態にしている。

②は入力値(実データ)の記述における表記揺れへの対応が必要である。検索されるキーワードはログデータなどから収集し、検索用語辞書として整備、運用している。具体的には、同義語をまとめるほか、入力されたキーワードに対する除外レシピを作成している。ユーザーが入力したキーワードで期待するレシピがヒットしない場合を少なくするための作業であり、その都度メンテナンスを行いながら検索精度の向上を図っている。

(3) 活用の実績と効果

構造化データの使用開始前と後で変化があったのかについては、実測値での効果は測定していない。ただし、Google 検索の結果表示が、他と比較してより詳細に表示されていることは確認できているため、構造化データを使用する効果は認められている。辞書の登録、整備運用は、ウェブサイトを訪れ情報にアクセスするユーザーの快適度、満足度につながっていると考えられる。

2.2 新規語彙体系の確立による効果

新規語彙体系の確立により効果を得ている事例として、医療用医薬品添付文書記載要領改正に伴う添付文書情報の XML 化対応のケースを紹介する。

(1) 導入の経緯

医療用医薬品添付文書とは、「医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律[30]」に規定された、医療用医薬品の使用および取扱い上の必要な注意等を記した文書である。その記載要領は、厚生労働省の通知により定められている。

1997 年に定められた記載要領（以降、旧記載要領）は、医療の進歩や高齢化、IT 技術の進歩など、医療を取り巻く状況が大きく変化していることを踏まえ、20 年後の 2017 年に改訂された(以降、新記載要領)。旧記載要領のデータ記述には SGML[31]記法が使われていたが、新記載要領になったことを機に、データ記述が XML 記法へ変更されることになった。

新記載要領による運用は、添付文書の届出及び安全性情報の掲載を行うシステム(以降、医薬品医療機器情報提供システム)の改修を経て、2019 年 4 月に開始し、2024 年 3 月までの 5 年を移行期間としている。移行期間の終了後は、XML で記述された新記載要領に基づく添付文書を公開しなければならない。このため、既存の SGML データは XML に変換する必要がある。

医薬品医療機器情報提供システムを運用する PMDA[32]では、2016 年度に添付文書の XML 化の検討と関連システムの改修要件定義を開始した。XML 化の検討では、添付文書に必要なデータ項目、値、そしてそれらのデータ構造について議論し、XML Schema(語彙)を策定した[33]。システムの実装は 2017 年度から 2018 年度に掛けて行い、2019 年 4 月に並行運用を開始している。5 年後の 2024 年 3 月までに既存 SGML データの XML への変換(再提出)を完了させ、以後は XML による完全運用へ移行する計画である。

(2) 医薬品医療機器情報提供システムの概要

医薬品医療機器情報提供システムは PMDA によって整備・運用される。同システムは、医薬品・医療機器等の製造販売業者(製薬会社など)自らがシステム操作を行い、添付文書の届出やその他の安全性情報の掲載を行うことができる。同システムにより掲載された添付文書等の情報は、最終的に PMDA ウェブサイトに公開される。

---

[30] 昭和 35 年法律第 145 号

[31] Standard Generalized Markup Language:ISO8879; XML の前身となるデータ記述記法。

[32] 独立行政法人 医薬品医療機器総合機構; Pharmaceuticals and Medical Devices Agency の略称。

[33] 「3.5.2 医療用医薬品の添付文書情報の電子化書式(XML)」も参照のこと。

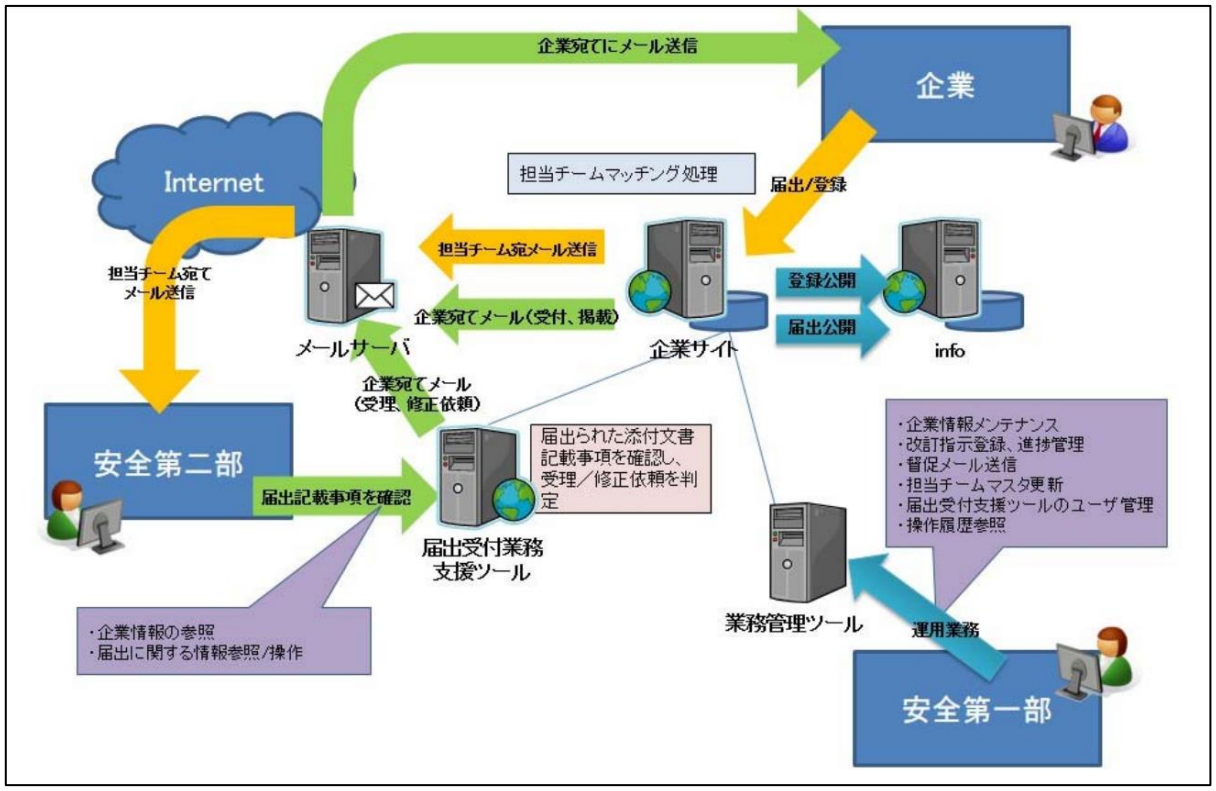

図 2-2 添付文書届出受付システム関連図（平成 29 年 3 月時点）

医療用医薬品添付文書記載要領改正に伴う医薬品医療機器情報提供システムの機能追加及び改修業務(平成 29 年 3 月)の仕様書[34]より

(3) 活用の実績と効果

現在、医療用医薬品の添付文書情報に関するデータは、新記載要領により統一された語彙を用いて記述することが関係者の間で合意されている。これにより、以下のような効果が得られている。

- 添付文書届出受付・確認業務の効率化(PMDA)
- 届出業務の利便性向上(医薬品・医療機器の製造販売業者)

[34] https://www.pmda.go.jp/files/000217037.pdf

第3章 語彙と用語辞書

本章では、「語彙」と「用語辞書」について、その構築方法や使い方などを具体的に説明する。

3.1 本ガイドにおける考え方

第 1 章でも述べた通り、本ガイドでは「用語」、「語彙」および「用語辞書」について以下のように定義している。

(1) 用語

人は、様々なモノやコトの概念に対して言葉を当てはめてコミュニケーションを取る。言葉は、紙の文書やコンピューターのデータとして扱う際に文字列として表記される。このモノやコトの概念を表すために用いられる言葉を文字列として表記したものを「用語」と呼ぶ。

(2) 語彙

概念を記述するために用意されたデータ項目と値の集合が「語彙」である。語彙のデータ項目や値は、それが表す同一のモノやコト(概念)に対して一つの代表用語を使って定義され、データ項目間の関連性(階層構造など)の定義も行う。語彙はデータの共通理解のための用語表現を統一することを意図して作成されるものであり、データの交換を行うユーザー間においてデータの意味を共通理解するための土台となる。

(3) 用語辞書

データ交換を行う業界や組織に存在する様々なデータや文書を分析し、同一のモノやコト(概念)を表現していると思われる用語を名寄せした後に、整理したものが「用語辞書」である。用語辞書の整理は、「概念」を定義して語彙を作成するための前作業としても必要となる。この整理の方法として、単に用語をリスト化するだけではなく、用語の分野に即して上位概念・下位概念という形で体系的に用語を分類しておくことが望まれる。

3.2 語彙について

3.2.1 語彙の定義情報

「語彙」からは、データ項目に関連した 2 つの観点での定義情報を得ることができる。個別のデータ項目に関する定義情報、そして、各データ項目間の関連に関わる定義情報である。

(1) 個々のデータ項目について

- データ項目の表記は何か
- 取りうる値(データ型や値リストなど)

例えば、ある語彙の定義の中に【人の居住国】という概念がデータ項目として定義されている場合、その表記が”Person_CountryOfResidence”であり、取りうる値は ” JP”、” KR”、” US”などのコードであるということが定義されている。定義は XML Schema などを使ったスキーマで行われ、それに基づく XML データは以下のように記述される。

```
<Person_CountryOfResidence>JP</Person_CountryOfResidence>
```

図 3-1 語彙定義に基づく XML データの例

(2) データ項目間の関連について

- データ項目が他のどのようなデータ項目と関連があるか
- データ項目同士の関連性の意味(包含、構成、属性、など)

一つのデータ項目が、複数のデータ項目によって構成されることがある。例えば、一般的に【会社】というデータ項目は、一つの【会社名】というデータ項目と複数の【部署】というデータ項目から構成され、【部署】は一つの【部署名】というデータ項目と複数の【社員】というデータ項目から構成される。このように、データ項目の間には、それぞれ他のデータ項目との関連性がある。この例では、その関連が階層構造を形成しており、末端にある【会社名】【部署名】が値を持つデータ項目ということになる。

「語彙」の定義は、データの記法に応じたスキーマ定義言語で記述する。具体的な例として、データの記法として XML を用いる場合は、そのスキーマ定義言語としての XML Schema で記述することになる。

「語彙」を定義したファイルは、コンピューターがデータをチェック(構文解析)するときに使う。そこで定義されている内容は、データ項目と値に関する非常に本質的な情報であり、データの共通理解のために欠かせない。

3.2.2 語彙の解説書

語彙の定義内容を関係者に周知するために使うのが「語彙の解説書」である。語彙を定義したコンピューター向けのスキーマ定義言語で記述されたファイルだけでは、何が「語彙」の定義として記述されているか人間には理解が難しい。したがって、「語彙」を解説した文書が必ず必要であり、その解説書を読むことによって、「語彙」を定義したファイルの中で記述されているデータ項目と値について人間も理解することができる。

語彙の解説書には、「3.2.1 語彙の定義情報」で説明したデータ項目や値に関する定義情報を人が読んで理解できるように記述される。

3.3 用語辞書について

3.3.1 データやデータ項目の概念を整理

データの共通理解の土台となる「語彙」を作るためには、既存のデータを分析し、同一概念に対応する様々な表記を同義語として整理(名寄せ)するという作業が必要となる。その作業の結果として出来上がる同義語リストが「用語辞書」である。

「用語辞書」には以下のような項目が含まれる。

・用語名
・意味
・略語(複数可)
・同義語(複数可)

「用語辞書」によって名寄せされたデータ項目に対して、相互の関連情報の付加などを施すことにより「語彙」としてのデータ項目の構造が定義される。また、用語辞書を整理する際に、用語が表す概念に対して上位概念・下位概念の関係などを体系的に定義する場合もある[35]。

3.3.2 用語の使われ方による辞書の違い

「3.2.1 語彙の定義情報(1)個々のデータ項目について」で示したように、「語彙」の定義には、「データ項目」と「値」という 2 種類の用語が関係する。このことは、既存データの分析によって名寄せする対象が 2 つあることを示しており、「用語辞書」においても「データ項目の用語辞書」と「値の用語辞書」の 2 種類に分類される。

例えば、料理レシピで使用する”材料”と、材料として使われる”人参”を考える。この場合、データ項目が”材料”であり、”人参”はそのデータ項目が取ることができる一つの値となる。

また、「(2)データ項目間の関連について」では、「語彙」には、データ項目間の関連性の定義があることを説明した。個々のデータ項目の用語に対して、データ項目間の関連性の情報を付与することにより、「語彙」におけるデータ項目の構造が定義される。

一方、値の用語辞書は、「値リスト」と呼ぶべきものであり、データ項目の定義時に参照される。この「値リスト」も、「語彙」のデータ項目の定義において指定されることになり、「語彙」の定義と同様にスキーマ定義言語による記述によって定義される。

このように、「語彙」は、「データ項目の用語辞書」と「値の用語辞書(値リスト)」の両方を基に整備される。

---

[35] 考え方の例として「3.5.3 共通農業語彙(CAVOC)」を参照のこと。

3.4 語彙の作成と利用

3.4.1 用語辞書と語彙の作成

ここまで説明した内容に基づいて、データ分析から語彙作成に至るまでの工程を示すと次のようになる。

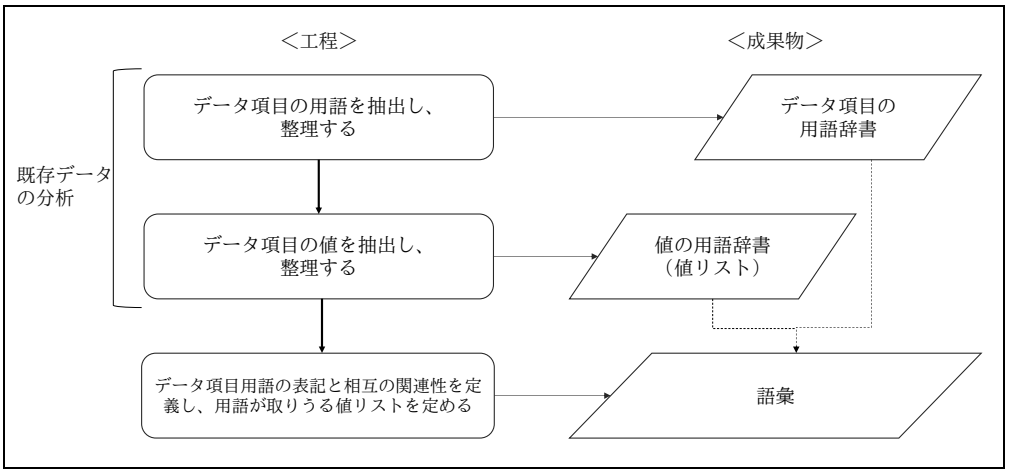

図 3-2 データ分析から語彙作成までの流れ

用語辞書には、名寄せの結果として選んだ代表的な名称が用語名として含まれており、語彙の定義には、その用語名を使う。

このような手順で作成した語彙を用いながら、データを共有する人々の間で統一された用語を使うという合意を形成する。データの共通理解は、この合意の上に実現されることになる。

3.4.2 語彙の利用

ここでは、出来上がった語彙を実際に使うときに必要な考え方や仕組みについて説明する。

(1) 語彙のバージョン管理

語彙を作成した後に業務内容の変化やデータの利用環境などの変化があると、新たなデータ項目や値を使う必要が出てくる場合や、逆に既存のデータ項目や値が不要になる場合がある。これにより、語彙に含まれるデータ項目や値の増減が生じる。語彙は時間とともに変化するため、ほとんどの場合、その内容に応じた複数のバージョンが存在する。

語彙に基づくデータをコンピューターで処理する場合、使われているデータ項目や値が正しく書かれているかをチェックする必要がある。このため、どのバージョンの語彙を使ってデータが書かれているかを指定することが不可欠となる。

(2) 表示ラベル[36]の指定

実際の語彙では、データ項目に関して、代表用語に対して英数字を使用した英語名称を使うことが多い。これは、国際的なデータ交換やコンピューターでプログラミングする際の容易性を考慮したものである。このため、データ項目を日本語などで表示する際には、代表用語に対応する表示ラベルを使う必要が生じる。

コンピューターのデータとして書かれた XML 要素などに何らかの表示ラベルを対応付けるためには、スタイルシートなどデータ自体とは別の仕組みを使って指定することが多い。

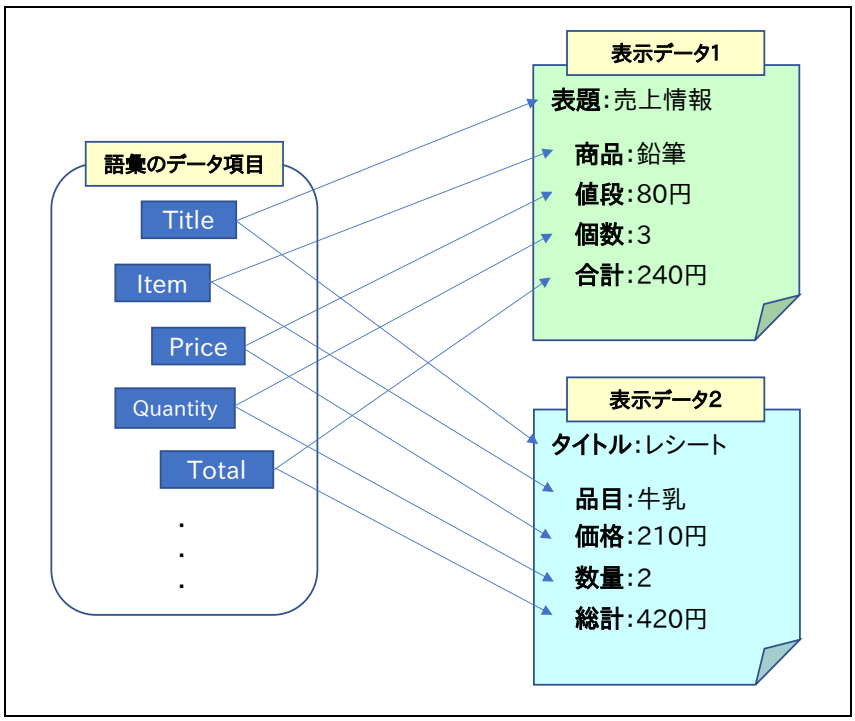

図 3-3 語彙のデータ項目とその表示

[36] 表示ラベル: text to display

3.5 語彙体系の例

以下に、日本で利用されている語彙の例を示す。

3.5.1 XBRL を使った財務報告用 EDINET タクソノミ

金融庁は、EDINET[37]において、有価証券報告書などの財務報告関連情報を開示するための電子データ形式として、XBRL という XML ベースの標準規格を使って情報を提供するよう企業に求めている。

この XBRL はデータの構文を定義した規格であり、企業開示制度で利用する場合には、そこに記述する報告項目は、適用する財務報告基準または国ごとの金融関連法制度に合わせて作成することになる。XBRL ではそのような報告項目の集合を「タクソノミ(taxonomy)」と呼んでおり、これが本ガイドで考えている語彙に相当する。また、日本の財務報告では、EDINET 用のタクソノミが整備されている。

第一部【企業情報】

第１【企業の概況】

１【主要な経営指標等の推移】

(1) 連結経営指標等

回次		第104期	第105期	第106期	第107期	第108期
決算年月		平成29年３月	平成30年３月	平成31年３月	令和２年３月	令和３年３月
売上高	(百万円)	231,282	273,802	303,080	316,934	323,609
経常利益	(百万円)	2,546	8,632	10,898	10,646	15,263
親会社株主に帰属する当期純利益	(百万円)	1,235	2,807	3,392	7,558	8,056
包括利益	(百万円)	2,107	3,543	3,967	9,409	6,780
純資産額	(百万円)	81,290	98,045	100,435	225,225	229,563
総資産額	(百万円)	286,829	294,251	298,813	496,837	509,039
１株当たり純資産額	(円)	243.41	295.50	302.94	699.94	702.29
１株当たり当期純利益金額	(円)	3.84	9.04	10.55	23.50	25.05
潜在株式調整後１株当たり当期純利益金額	(円)	3.55	8.94	10.23	23.42	24.88
自己資本比率	(%)	27.3	32.3	32.6	44.7	44.4
自己資本利益率	(%)	1.58	3.06	3.48	3.40	3.57

図 3-4「開示府令 第三号様式 有価証券報告書」のサンプル

(出典)金融庁ホームページ 2021 年版 EDINET タクソノミの公表について[38]

サンプルインスタンス(https://www.fsa.go.jp/search/20201110/1h-1_Sample.zip)より抜粋

[37] Electronic Disclosure for Investors' NETwork; 金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システムの略。

[38] https://www.fsa.go.jp/search/20201110.html

タクソノミは以下のように定義されている。

	B	H	I
33	連結経営指標等	jpcrp_cor	BusinessResultsOfGroupLineItems
34	売上高	jpcrp_cor	NetSalesSummaryOfBusinessResults
35	売上収益	jpcrp_cor	RevenueKeyFinancialData
36	営業収益	jpcrp_cor	OperatingRevenue1SummaryOfBusinessResults
37	営業収入	jpcrp_cor	OperatingRevenue2SummaryOfBusinessResults
38	営業総収入	jpcrp_cor	GrossOperatingRevenueSummaryOfBusinessResults
39	経常収益	jpcrp_cor	OrdinaryIncomeSummaryOfBusinessResults
40	正味収入保険料	jpcrp_cor	NetPremiumsWrittenSummaryOfBusinessResultsINS
41	経常利益又は経常損失(△)	jpcrp_cor	OrdinaryIncomeLossSummaryOfBusinessResults
42	親会社株主に帰属する当期純利益又は親会社株主に帰属する当期純損失(△)	jpcrp_cor	ProfitLossAttributableToOwnersOfParentSummaryOfBusinessResults
43	包括利益	jpcrp_cor	ComprehensiveIncomeSummaryOfBusinessResults
44	純資産額	jpcrp_cor	NetAssetsSummaryOfBusinessResults
45	総資産額	jpcrp_cor	TotalAssetsSummaryOfBusinessResults
46	1株当たり純資産額	jpcrp_cor	NetAssetsPerShareSummaryOfBusinessResults
47	1株当たり当期純利益又は当期純損失(△)	jpcrp_cor	BasicEarningsLossPerShareSummaryOfBusinessResults

図 3-5 データ項目とタクソノミ要素のリスト

(出典)金融庁ホームページ 2021 年版 EDINET タクソノミの公表について
タクソノミ要素リスト(https://www.fsa.go.jp/search/20201110/1e_ElementList.xlsx)より抜粋

上図の I 欄の”NetSalesSummaryOfBusinessResults”などが、最終的に XBRL の語彙(タクソノミ)として決めたデータ項目名である。B 欄の”売上高”などは、「3.4.2 語彙の利用」で説明した表示ラベルの指定に当たる。(H 欄の”jpcrp_cor”は、XML でデータを作成する際、データ項目がどの語彙のものなのか特定するためにデータ項目の前に付ける接頭辞であるが、ここでは詳しくは触れない。)

この語彙を使って、有価証券報告書を記述した XBRL データを次に示す。

```
    </xbrli:divide>
</xbrli:unit>
<jpdei_cor:NumberOfSubmissionDEI decimals="0" unitRef="pure"
    contextRef="FilingDateInstant">1</jpdei_cor:NumberOfSubmissionDEI>
<jpcrp_cor:NetSalesSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior4YearDuration">2312[illegible]2000000</jpcrp_cor:NetSalesSummaryOfBusinessResults>
<jpcrp_cor:NetSalesSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior3YearDuration">273802000000</jpcrp_cor:NetSalesSummaryOfBusinessResults>
<jpcrp_cor:NetSalesSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior2YearDuration">303080000000</jpcrp_cor:NetSalesSummaryOfBusinessResults>
<jpcrp_cor:NetSalesSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior1YearDuration">316934000000</jpcrp_cor:NetSalesSummaryOfBusinessResults>
<jpcrp_cor:NetSalesSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="CurrentYearDuration">323609000000</jpcrp_cor:NetSalesSummaryOfBusinessResults>
<jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior4YearDuration">2546000000</jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults>
<jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior3YearDuration">8632000000</jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults>
<jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior2YearDuration">10898000000</jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults>
<jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior1YearDuration">10646000000</jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults>
<jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="CurrentYearDuration">15263000000</jpcrp_cor:OrdinaryIncomeLossSummaryOfBusinessResults>
<jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior4YearDuration">1235000000</jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfB
<jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior3YearDuration">2907000000</jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfB
<jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfBusinessResults decimals="-6" unitRef="JPY"
    contextRef="Prior2YearDuration">3392000000</jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfB
<jpcrp_cor:ProfitLossAttributableToOwnersOfParentSummaryOfBusinessResults decimals="-6" unitRef="JPY"
```

図 3-6 XBRL データ

(出典)金融庁ホームページ 2021 年版 EDINET タクソノミの公表について
サンプルインスタンス(https://www.fsa.go.jp/search/20201110/1h-1_Sample.zip)より抜粋

3.5.2 医療用医薬品の添付文書情報の電子化書式(XML)

厚生労働省は、患者の安全を確保し医薬品の適正使用を図るために、医薬関係者に対して必要な情報を提供する目的で、製造販売業者が医療用医薬品の添付文書を作成することとしているが、その添付文書に記載するデータ項目(「警告」「製法の概要」「組成」など)が XML で規定されている。

添付文書の記載要領については厚生労働省の通知により定められており、1997 年に定められた記載要領(以降、旧記載要領)は、医療の進歩や高齢化、IT 技術の進歩など、医療を取り巻く状況が大きく変化していることを踏まえ、20 年後の 2017 年に改訂された(以降、新記載要領)。旧記載要領のデータ記述には SGML 記法(Standard Generalized Markup Language:ISO8879; XML の前身となるデータ記述記法)が使われていたが、新記載要領になったことを機に、データ記述が XML 記法へ変更されることになった。

これに伴い、PMDA が管理する医薬品医療機器情報提供システムを改修し、2019 年 4 月 1 日に運用を開始、2024 年 3 月 31 日までの 5 年を移行期間とした。移行期間の終了後は、新記載要領に基づく添付文書を XML データで届出しなければならない。このため、既存の SGML データは XML に変換する必要がある。

要素名	項目名	Occurs min	Occurs max
GenericName	キ.一般的名称	1	1
SupplementaryInformaiton0fVaccineStrain	ワクチン株の補足情報	0	1
SpeciallyDescribedItems	特殊記載項目(※ 当該欄は行政が指示した	0	1
Wamings	1.警告	0	1
ContraIndications	2.禁忌(次の患者には投与しないこと)	0	1
CompositionAndProperty	3.組成・性状	0	1
0verview0fRecipe	製法の概要	0	1
Composition	3.1 組成	0	1
0verview0fComposition	組成の概要	0	1
CompositionForBrand	薬品毎の組成	0	unbounded
CompositionForConstituentUnits	構成毎の組成	0	unbounded
ConstituentUnits	構成	0	1
CompositionTable	組成テーブル	0	unbounded
CompositionAndPropertyTblTitle	組成・性状テーブルのタイトル	0	1
ContainedAmount	有効成分	0	unbounded
ActiveIngredientName	有効成分名	0	1
ValueAndUnit	含有量	0	1
ActiveIngredientAdditionalInfo	有効成分の追加情報	0	1
ActiveIngredientName	有効成分名	1	1
ValueAndUnit	含有量	0	1
Additives	添加剤	0	1
xs:choice 以下のいずれかを選択。		1	1
List0fAdditives	添加剤リスト	1	1
IndividualAdditives	個別添加剤情報(繰り返し)	1	1
InfoIndividualAdditive	個別添加剤情報	1	unbounded

図 3-7 医療用医薬品添付文書のデータ項目の整理と XML 要素の対応付け

(出典)PMDA ホームページ 添付文書改訂等の安全対策に関連する通知等[39]の
「平成 30 年 11 月 22 日 薬機安一発第 1122001 号」「(別添 4)入力項目一覧」(https://www.pmda.go.jp/files/000236333.xlsx)より抜粋

図 3-7 の”要素名”欄の”Warnings”などが、添付文書の新記載要領の語彙として決めたデータ項目名に当たる。また、”項目名”欄の”1. 警告”などは、「3.4.2 語彙の利用」で説明した表示ラベルの指定に当たる。(”min 0ccurs”欄の”0”,”1”、 ”max 0ccurs”欄の”1”,”unbounded(無限)”は、それぞれのデー

[39] https://www.pmda.go.jp/safety/consultation-for-mah/0002.html

タ項目に関する出現回数の下限と上限を表している。）

この語彙を使った医薬品の添付文書を記述する XML データのテンプレート(値は入っていない)を次に示す。

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- 添付文書 -->
- <PackIns xml:lang="ja" containedLang="ja" drugType="Medicine" xmlKind="Packins"
xmlns="http://info.pmda.go.jp/namespace/prescription_drugs/package_insert/1.0"
referenceOfPrecautionsForHandling="" referenceOfPrecautionsForUse="" version="">
        <!--添付文書番号-->
    <PackageInsertNo/>
        <!--企業コード-->
    <CompanyIdentifier/>
        <!-- ア.作成又は改訂年月 -->
  + <DateOfPreparationOrRevision id="HDR_DateOfPreparationOrRevision">
        <!-- イ.日本標準商品分類番号 -->
  + <Sccj id="HDR_Sccj">
        <!-- オ.薬効分類名 -->
    <TherapeuticClassification id="HDR_TherapeuticClassification"/>
        <!-- 承認等 -->
  + <ApprovalEtc>
        <!-- キ.一般的名称 -->
    <GenericName id="HDR_GenericName"/>
        <!-- ワクチン株の補足情報 -->
    <SupplementaryInformaitonOfVaccineStrain/>
        <!-- 特殊記載項目 -->
    <SpeciallyDescribedItems/>
        <!-- 1.警告 -->
    <Warnings id="HDR_Warnings" heading="fixing"/>
        <!-- 2.禁忌 -->
    <ContraIndications id="HDR_ContraIndications" heading="fixing"/>
        <!-- 3.組成･性状-->
  - <CompositionAndProperty id="HDR_CompositionAndProperty" heading="fixing">
            <!-- 3.1 製法の概要 -->
      <OverviewOfRecipe id="HDR_OverviewOfRecipe" heading="fixing"/>
            <!-- 3.1 組成-->
    - <Composition id="HDR_Composition" heading="fixing">
        <OverviewOfComposition/>
            <!-- 薬品毎の組成 -->
      - <CompositionForBrand ref="BRD_Drug1">
                <!-- 構成毎の組成 -->
```

図 3-8 医療用医薬品添付文書の XML データのテンプレート

(出典)PMDA ホームページ 添付文書改訂等の安全対策に関連する通知等の
「平成 30 年 11 月 22 日　薬機安一発第 1122001 号」「(別添 2)テンプレートインスタンス」(https://www.pmda.go.jp/files/000228344.zip)より抜粋

図 3-8 にあるように、データ項目に対応する XML 要素は、"Warnings","OverviewOfRecipe","Composition"のような英語で定義されている。このため、このデータを使って医薬品に添付されている印刷物を作る際には、対応する表示ラベル("Warnings"→"警告"、"OverviewOfRecipe"→"製法の概要"、"Composition"→"組成"など)をスタイルシートなどの何らかの仕組みを使って組版(印刷物作成のためのフォーマット付け)を行うことになる。

3.5.3 共通農業語彙(CAVOC)[40]

農作物(「カカオ」など)や農作業(「稲刈り」など)の「値/用語」を整理して体系化したもの。

■ 農作物語彙体系(Crop Vocabulary): http://www.cavoc.org/cvo.php

農薬情報管理における標準語彙としての利用を意図して、農作物の名称をリスト化したものである。それぞれの農作物に対して図 3-9 のように同義語(Synonym)、科名などの分類情報が記述されている。

図 3-9 農作物の名称のリスト(値リスト)

(出典)http://www.cavoc.org/cvo/ns/3/C1309

右の欄で”オクラ(総称)”の下に”オクラ(果実)”,”オクラ(花)”が位置づけられているように、農作物の名称の上位・下位の関係も表現されている。なお、”アオイ科”などの科名はデータ構造として階層化されているわけではなく、それぞれの科目の情報を基に科目ごとに農作物をまとめて表示したものである。

■ 農作業基本オントロジー: http://www.cavoc.org/aao.php

農作業の名称を、階層的な構造で論理的に意味づけし整理したものである。農作業管理における標準語彙としての利用を意図して構築されている。

[40] http://www.cavoc.org/cavoc.php

図 3-10 階層化された農作業の名称のリスト(値リスト)

(出典)http://cavoc.org/aao/ns/4/A295

図 3-10 のように、個々の農作業に対して、農作業名、意味、上位作業名、下位作業名などの情報が記述されており、上位・下位の関係によって農作業用語の階層構造が定義されている。

おわりに

本ガイドは、「何らかの目的を持って作成したデータを、当初の目的を超えて広く利活用するには、関係者間、あるいは、コンピューター間で、そのデータが持つ意味について互いの理解に共通性がある状態で利用する必要がある」という考えの基、データの共通理解をするための標準的な方法の一つを示している。

本ガイド作成にあたって議論した代表的な課題に以下のようなものがある。

- データが持つ意味を共通理解しているとはどのような状態か
- データカタログにあるようなデータの説明情報を一般的には「メタデータ」と呼ぶが、ISO/IEC の一部の規格において「メタデータ」を異なる意味で使っている場合がある
- 概念間の関係性を語彙に定義してコンピューター処理することで何かメリットがあるのか

「データが持つ意味を共通理解する」というのは、人の場合は「その対象の概念が関係者の頭の中で一致すること」というのが概ねの回答となる。コンピューターの場合は「その対象が同一であるのか異なるものなのかが判別できること」だと本ガイドでは考えている。ここで、「概念」という言葉が、読者のガイドに対する印象を難解なものとしてしまわないよう丁寧に解説を試みたつもりであるが、ご理解いただけたであろうか。

「メタデータ」という言葉については、国際標準などを参照した際に、個々の標準の中で異なる意味合いで使われており、明確な定義として「これだ」と言えるものが見当たらなかった。このため、読者の混乱を避ける意味で、本ガイドでは使用しないこととした。

「語彙」の持つ情報に概念間の関係性を示すことが必要とよく言われている。一般的な業務処理を想定した際に、「この関係性の情報が実際のコンピューター処理に役立つのか」という課題は「将来的に AI で活用」といった解以外に当初メリットが想定できず、本ガイド記載の妥当性が問われた。しかしながら、出所が異なったり、集計レベルが異なったりしているデータを統合して統計的に処理する際に、それらのデータ項目や値と語彙が紐づけられていれば、統合したデータを語彙の概念階層などに基づいて容易に処理することができるといった気づきから、ガイドへの記載が確定した。これは一見すると当たり前のことであるが、意味を共通理解するための語彙があるからこそデータの連携が可能となり、容易に相互利用が可能になることをご理解いただきたい。

本ガイドは、上記のように様々な議論を経て作成された。これからのデータ活用に向け、データの意味を共通理解することの重要性を認識いただき、今後のデータ作成やデータ整備などを実施される際に、語彙や用語辞書を活用しながら、データを共有する組織や集団の枠を超えて、その意味を伝えるための方法を採り入れることを検討いただけることを願っている。

この文書について

■ 表題

・ データの共通理解推進ガイド　ー用語辞書や語彙を用いたデータの共通理解ー

■ 公開履歴

・ 初版　　2022 年 3 月 18 日

■ 監修・協力　(各 50 音順, 所属は公開時のもの)

・ (監修協力)	頃末和義	独立行政法人情報処理推進機構専門委員
	武田 英明	独立行政法人情報処理推進機構専門委員、国立情報学研究所
・ (事例協力)	堀田敏史	株式会社エブリー
・ (編集協力)	奥井康弘	株式会社ティージェイ総合研究所
	柏崎 吉一	合同会社エクリュ
	三浦まゆみ	フリーライター
・ (監修)	斉藤 浩	独立行政法人情報処理推進機構
	萩原 正規	独立行政法人情報処理推進機構
	堀越 秀朗	独立行政法人情報処理推進機構
	森貞 夏樹	独立行政法人情報処理推進機構
	我妻 浩子	独立行政法人情報処理推進機構

■ 編集・発行

独立行政法人情報処理推進機構(IPA)（法人番号 5010005007126）
東京都文京区本駒込 2-28-8 文京グリーンコートセンターオフィス

この文書のご利用にあたって

本ガイドの内容を適用した結果生じたこと、また、適用できなかった結果については、IPA は一切の責任を負いかねますのでご了承ください。