データの共通理解推進ガイド（概要版）

ー用語辞書や語彙を用いたデータの共通理解ー

2022 年 7 月 4 日 初版

独立行政法人情報処理推進機構(法人番号 5010005007126)

はじめに

本ガイドについて

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。「データ項目自体が同じものとして認識できるか」、「データの表記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。

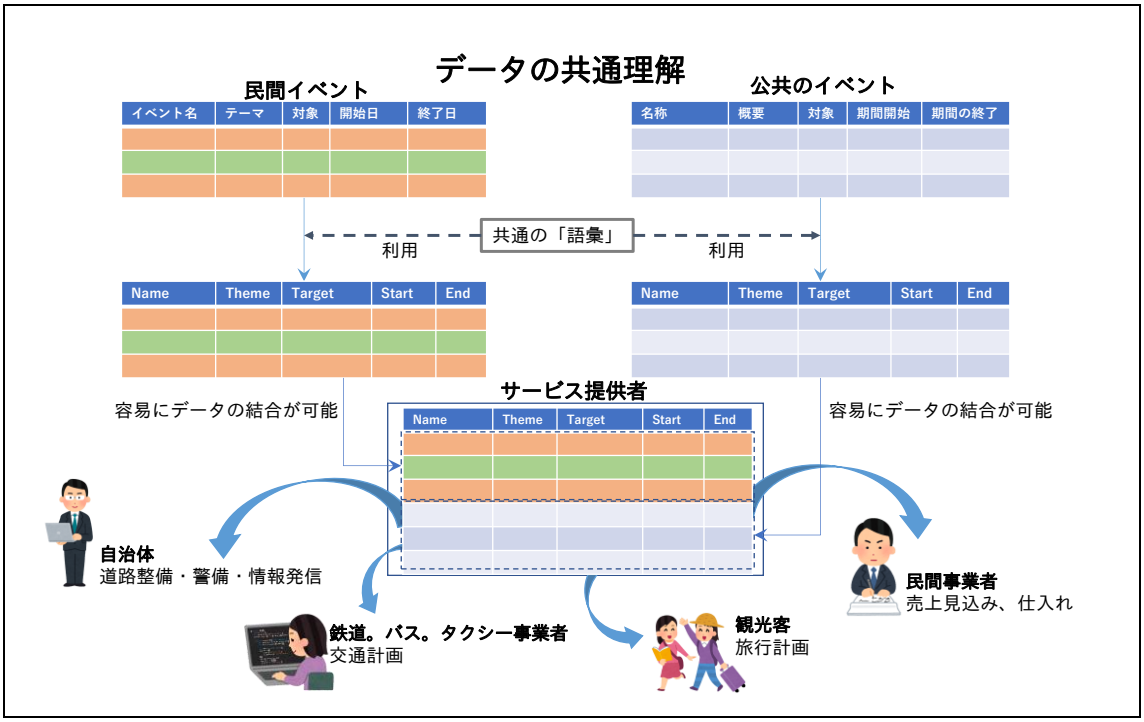

図 共通理解によるデータの価値向上

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・語彙の必要性、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。

本ガイドの読者には、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ような方々を想定している。

データが持つ意味の観点

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[1]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。

図は、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[2]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)として記録することを示している。

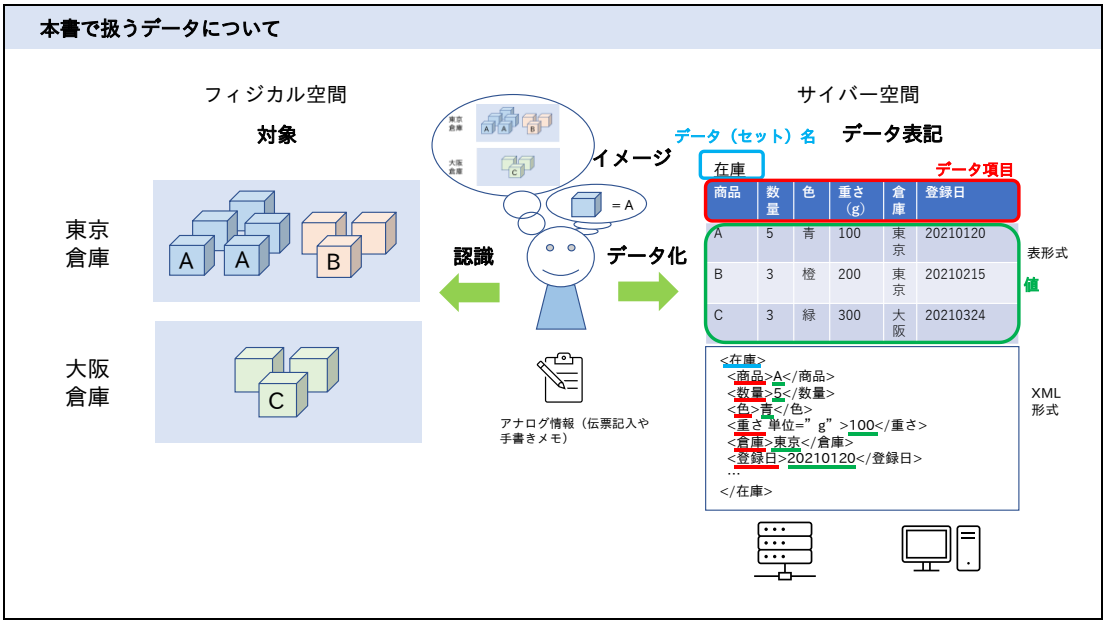

図 本ガイドで扱うデータについて

ここで、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いう前提があれば、”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。

しかしながら、上述したような前提がない場合、データの共通理解は難しくなる。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。また、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を推測する柔軟性があるが、コンピューターで処理する際に、その柔軟性を求めるのは難しい。

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

[1] 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。

[2] 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。

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。

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。

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[3]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際に、その代表用語として「商品」を選択し、【商品】と記述する。

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。

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、大きく二つの要素がある。

(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[4] 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などである。

この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。

「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。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し、それらを集合として管理/参照できるようにしたものが「語彙」として使われる。

---

[3] 概念そのものを表すための記述と捉えることもできる。

[4] 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。

また、「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(同義語など)が使われていることがある。

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。

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。

データ分析から語彙作成に至るまでの工程を示すと下図のようになる。

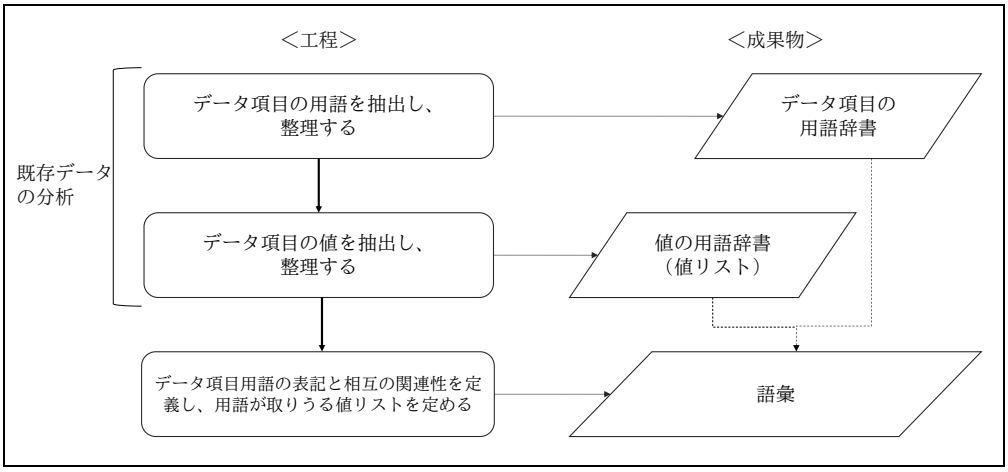

図 データ分析から語彙作成までの流れ

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し、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。

■本ガイドにおける「用語」「語彙」「用語辞書」について

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。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

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。語彙が備えるべき情報として、個別のデータ項目や値に関する定義情報、各データ項目間の関連に関わる定義情報、および、これらの定義情報を人が読んで理解できるように記述される解説書が求められる。

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。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。「語彙」の定義に、「データ項目」と「値」という２種類の用語が存在するが、「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種類が存在する。

(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。この説明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、図書館の書誌情報と同様、データセットの検索に使うこともできる。

「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おり、以下のような既存規格も存在する。

・「ダブリン・コア(Dublin Core)[5]」
・「Data Catalog Vocabulary (DCAT)[6]」

[5] 「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。

[6] World Wide Web Consortium(W3C)にて策定された、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の標準化。

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。

この具体的な例として、既存の語彙(Schema.org)を適用したウェブサイトの構築や、医療用医薬品添付文書に関する語彙の整備によるデータ活用の利便性向上などについて、ガイド本編で紹介しているので参照されたい。

この文書について

■ 表題

データの共通理解推進ガイド(概要版)　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 7 月 4 日

■ 監修　(各 50 音順, 所属は公開時のもの)

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

https://www.ipa.go.jp/

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。