データ辞書がなぜ必要か

独立行政法人情報処理推進機構
デジタル基盤センター
データスペースグループ

2024-03-29

IPA Better Life with IT

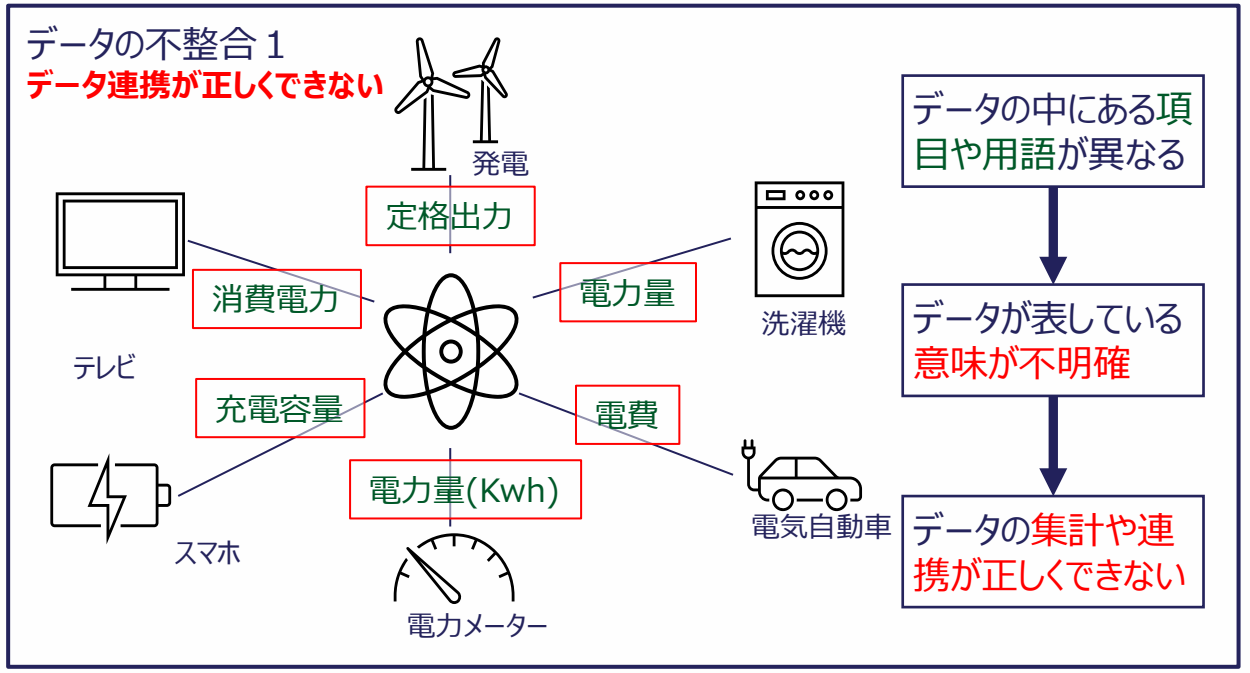
0．はじめに

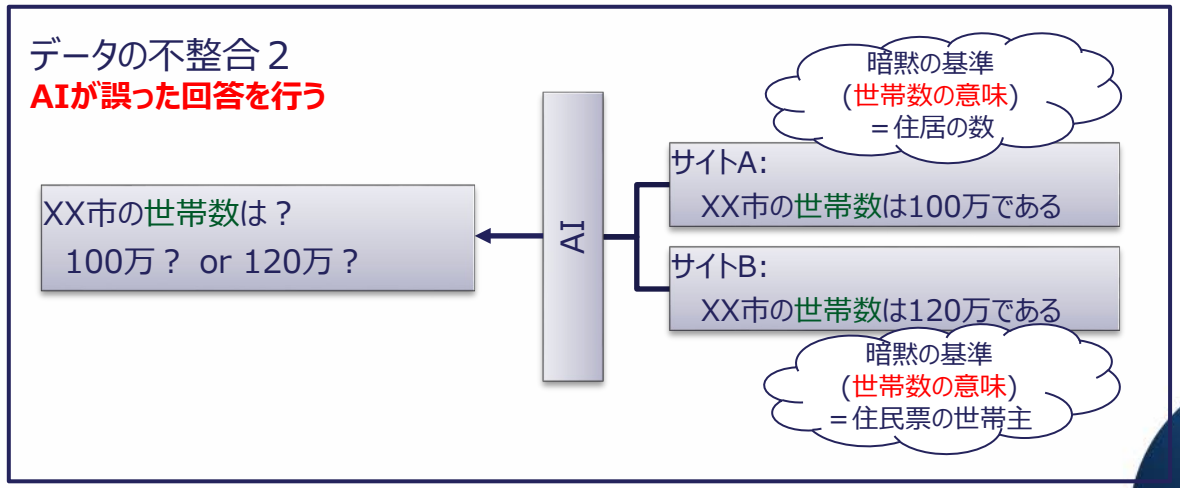
- 情報による経営環境(価値創造の仕組み)の変化
  - 近年インターネットの普及により、利用可能な情報(データ)が爆発的に増大しています。この増大したデータは、当初のデータの作成目的とは異なった目線を持つ第三者が利用する経営資源として扱われ始めています。身近な例としては、飲食店等に対する口コミ情報から、その店舗に直接関係のない周辺情報を抜き出し、地域のマーケティング情報として利用するなどが考えられます。
  - これからの経営環境は、様々なデータを経営資源として取り込み、それらを生かして意思決定のサポートを得、顧客満足度や市場競争力を向上し、新たなビジネスモデルの創出などを行うことが当たり前になってきています。
- 経営資源としてのデータ
  - 作成者が異なるデータは、同一の意味(概念)を持つ情報であっても、個々のデータの中で用いられる用語や項目名として違う言葉(ラベル)が割り当てられていることがあります。この不整合の存在は、データ連携、データを活用した分析、AI利活用などに対して問題を抱えます。言い換えると、共通した用語が使用されていないデータの散在は、データの利活用ができない状況を引き起こします。
  - データを経営資源として利活用するためには、後からデータを利用する際に不整合の起きない情報が必要です。

0．はじめに

- 不整合の起きない情報を利用するために
  - データを経営資源として後から利用するためには、もともとのデータがもっているデータの意味を判別するための仕組みが必要になってきます。この仕組みがあれば、データを利用した価値創造が非常に容易になると考えられます。
  - この仕組みの一つがデータ辞書※です。企業や団体は、今後このデータ辞書を整備することで、経営環境の変化に追随し、さらなる事業拡大を目指すことが必要になってきます。
  - 本書は、このデータ辞書について、その必要性や考え方について、できるだけわかりやすく説明することを目指しています。これからデータの利活用を経営に取り入れたい方やデータ整備の企画を推進する方などにも一読いただきたい資料です。

※データ辞書は一般的に以下のようなデータに関する情報を管理します。
データの意味、他のデータとの関係、フォーマット、履歴や起源などのメタ情報など





1．言葉の辞書、データ辞書、データ定義の形式

辞書といっても色々な形があります。いわゆる辞書(言葉の辞書)とデータ辞書の違いを確認します。

- **言葉の辞書の形式**
  - ある用語に対して、その用語が利用される シーン(文脈)毎に意味が説明される

言葉の辞書の一般的形式

```
用語名：　意味1「利用シーン」
　　意味2「利用シーン」
　　意味3「利用シーン」
　　：
```

言葉の辞書の例(名前の場合)

```
(用語名)名前：(意味1-1) 人の姓名 「新入社員の～をおぼえる」
　　：(意味1-2) 姓に対しての、名 「子に～を付ける」
　　：(意味2-1) 事物の名称（一般の名称） 「草木の～」
　　：(意味2-2) 事物の名称（固有の名称） 「山の～」
```

シーンごとの意味

利用されるシーンの例

デジタル大辞泉(https://www.weblio.jp/content/%E5%90%8D%E5%89%8D)より

1．言葉の辞書、データ辞書、データ定義

データ辞書と呼ばれるものには様々なものがありますが、後から情報を利用するために必要なデータ辞書はデータの意味を中心に整備されることが一般的です。

- **データ辞書の形式**(本書の場合)
  - 対象となるものの意味に対して、その対象がデータとして扱われるときのラベル(項目名)等が説明される

データ辞書のイメージ

意味A：代表ラベル、他ラベル、類似する意味。。。
意味B：代表ラベル、他ラベル、類似する意味。。。
意味C：代表ラベル、他ラベル、類似する意味。。。
：

対象となるものの意味

対象がデータとして扱われる際のラベルや意味間の関係性などを説明

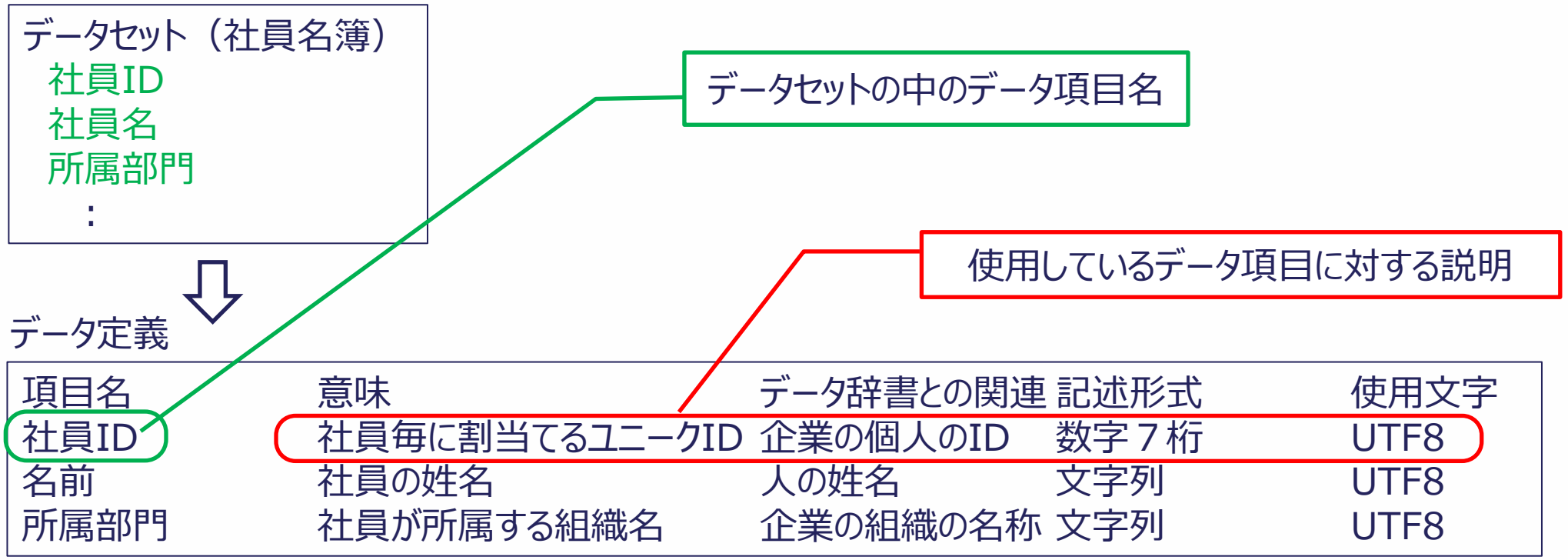
データ辞書のイメージ例

(意味A)人の姓名　：(代表ラベル)姓名；(他ラベル)名前，氏名；(類似する意味)意味B、。。。
(意味B)姓に対しての、名　：(代表ラベル)名；(他ラベル)名前；(類似する意味)意味A、。。。
(意味C)一般の事物の名称　：(代表ラベル)名称；(他ラベル)名前；(類似する意味)意味D、。。。
(意味D)固有の事物の名称　：(代表ラベル)固有名；(他ラベル)名前,名称；(類似する意味)意味C、。。。

1．言葉の辞書、データ辞書、データ定義

データ定義という言葉がデータ辞書の意味で扱われることがありますが、本書では異なるものとして扱います。 データ辞書はデータ定義から参照されることでより効果を発揮します。

- **データ定義の形式**(参考)
  - データセットに含まれているデータ項目についての説明。



２．物や事の情報をデータとして伝える

情報を記録するとは（データの生い立ち）

- データとは、対象となる事象について観測したりされたりしたときの情報を記録したもので、記録には様々な手段(方法や方式)があります。
- 原始的な時代には、洞窟の壁画に絵で表現した情報があり、その後、情報を伝達するための文字が開発されました。
- 文字を木簡や紙などに記録することで、情報を第三者に伝えることに関し、或る程度の共通認識ができる手段が確立してきました。
- これが、第三者に伝えるための情報が記録されたデータが生み出された最初の過程であると考えます。
- テクノロジーが進んだ今、情報は人や計測器が認識した観測の結果ですが、その結果を人が文字にしたり計測器が示す値などを利用して電子的に記録することでデータとして利用することができます。

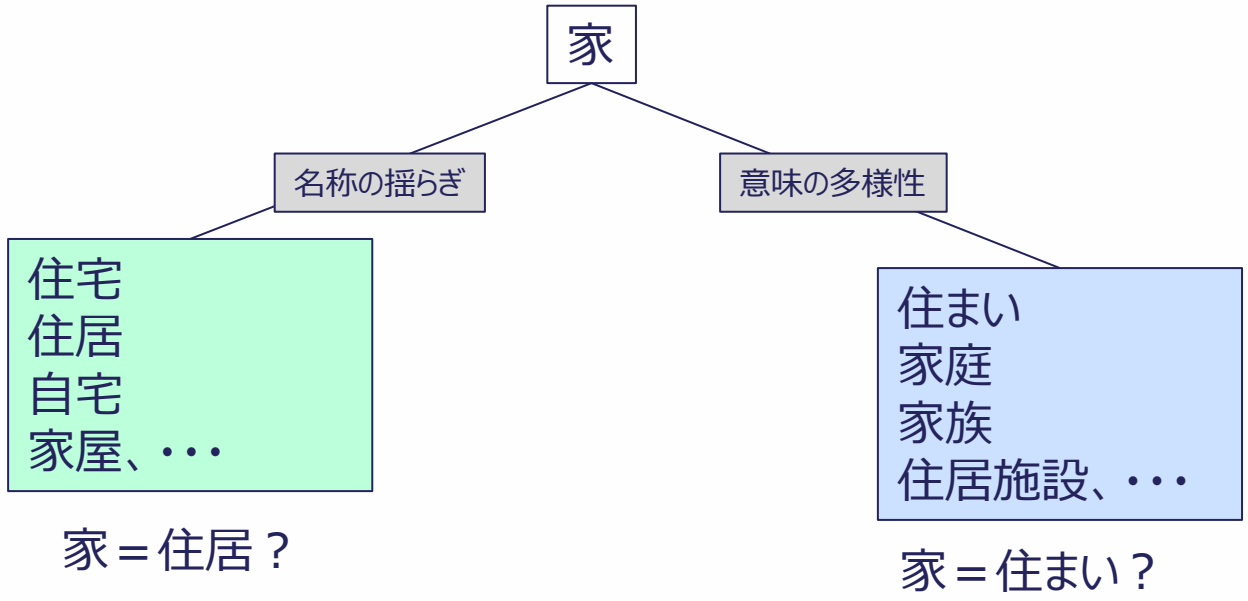
2．物や事の情報をデータとして伝える

情報伝達の難しさ

言葉が持つ意味には多様性があり、単語だけで伝えても何を意味するのかが解りません。

言葉の多様性

- Ex)”家”と一言で表現したデータでは、複数の名称や意味が存在するため何を指しているのかがわからない。



(参考)データの社会活用のためのセマンティックウェブ技術 / 国立情報学研究所 武田英明教授　発表資料

2．物や事の情報をデータとして伝える

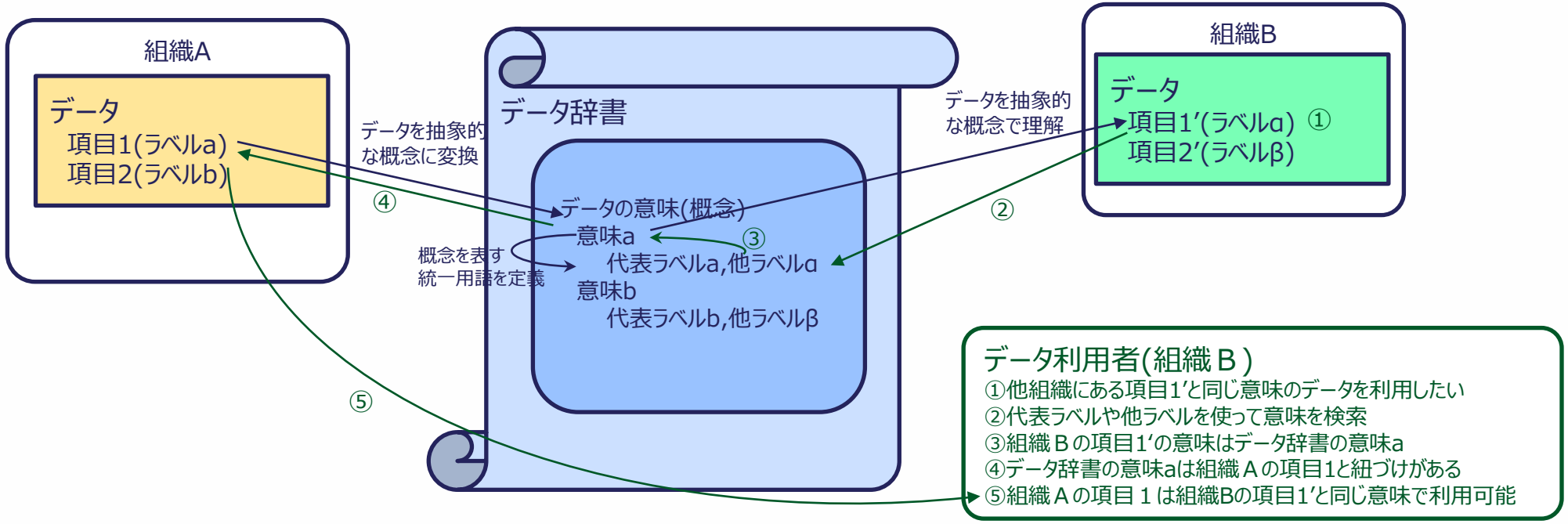
データの活用範囲について

- ある目的をもって作成されたデータは、その個別事情(作成者や作製組織等の枠組み)の中で留まっている限り、当初の目的に限定された断片的な情報の活用のみを実現します。
- また、個別事情によって記録されるデータは、同じ事柄を記録したつもりでも、作成者等がもつ個々の背景が異なることで、表現が異なるデータとして作成されます。そのため、これらのデータは、正しく認識できる情報として双方に伝わらなくなります。
- 現在、この個別事情で記録されたデータであっても、個々にとどまらず枠組みを超えて有効に使うためには、どのように考えれば良いのかという課題に対する議論が進んでおり、さまざまな工夫、方式を試みている状況があります。
- その一つの解としてデータ辞書を介した情報の伝達があると考えられています。

2．物や事の情報をデータとして伝える

データ辞書を利用してデータを定義すると、データ辞書に定義されている意味から内容を理解することができます。

データ辞書を介した情報の伝達イメージ



(参考)抽象化した概念で伝える
ある分野の情報は緻密で詳細に表現され、そのままでは他の分野には理解できないことがあります。
それらの情報を判別するためには、情報に抽象化した概念を結びつけることによって判別の糸口を持つことができます。
辞書には、それらの概念からたどって物事の本質を表現した情報が読み取れる必要があります。

3．データ辞書に必要なこと

誰もが理解できる辞書であること

- 辞書といっても、さまざまなレベル・種類が存在します。しかしながら、ほとんどの人が違いを理解せず、区別せずに使うため、判別に混乱が生じています。
- 分野・組織・国境を越えてデータをつなぐことが辞書には求められています。
- そのため、みんな(国際社会)が合意して決めた辞書を使うことが大前提になります。
- 各国・各社が独自に決めた表現は誰もわかりません。
  - 表現してあることが理解できないので、結局つながらない世界を再生産することの繰り返しになる。

「事・物」の関係や意味にラベルを付けて整備された方法で表現し、辞書として整理する。決められた方法で表現を計算機がわかる言葉(owl, RDFなど)で表記し、判別できるようにすることが必要です。

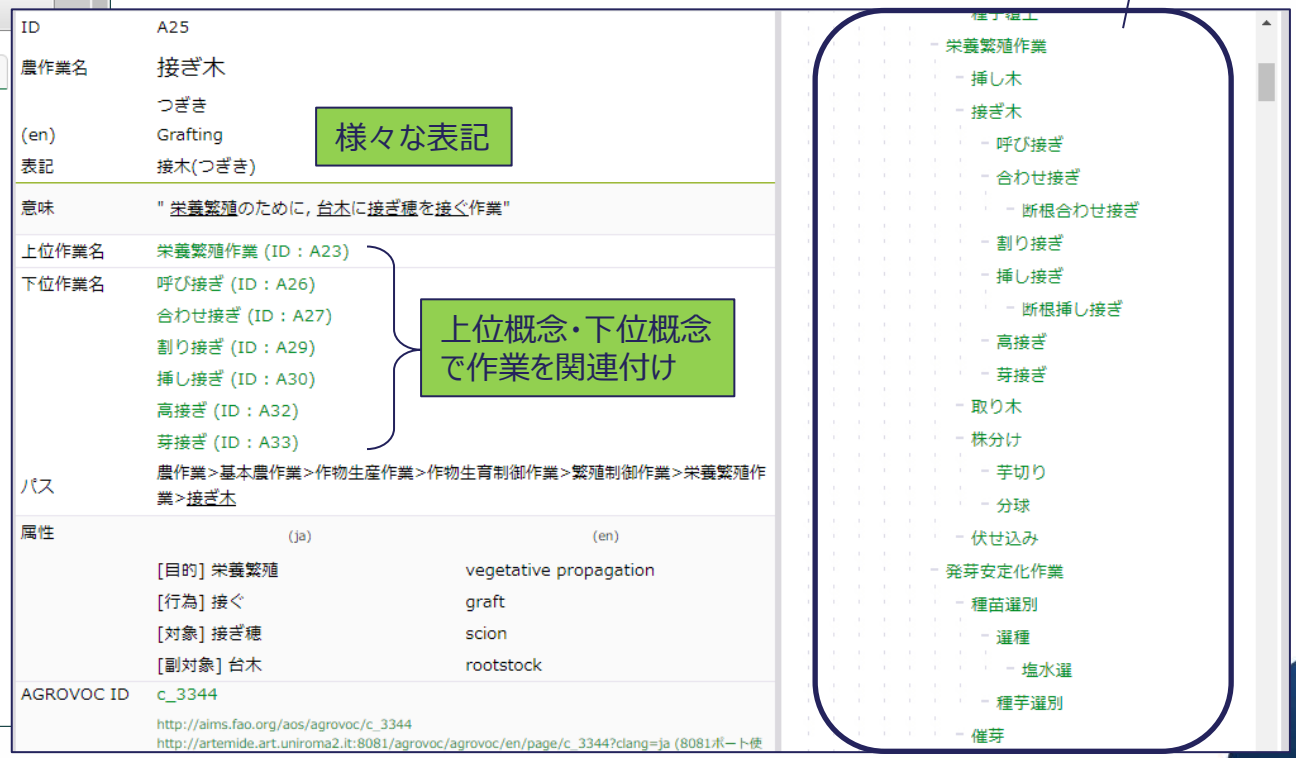
※参考: (一財)エンジニアリング協会: 苑田氏の資料

4．データ辞書の実例

農研機構と国立情報学研究所が中心となって、農業分野におけるシステム間の横断的なデータ連携を促進するために、コンピュータが参照できるデータ辞書(オントロジー)を構築。

↓

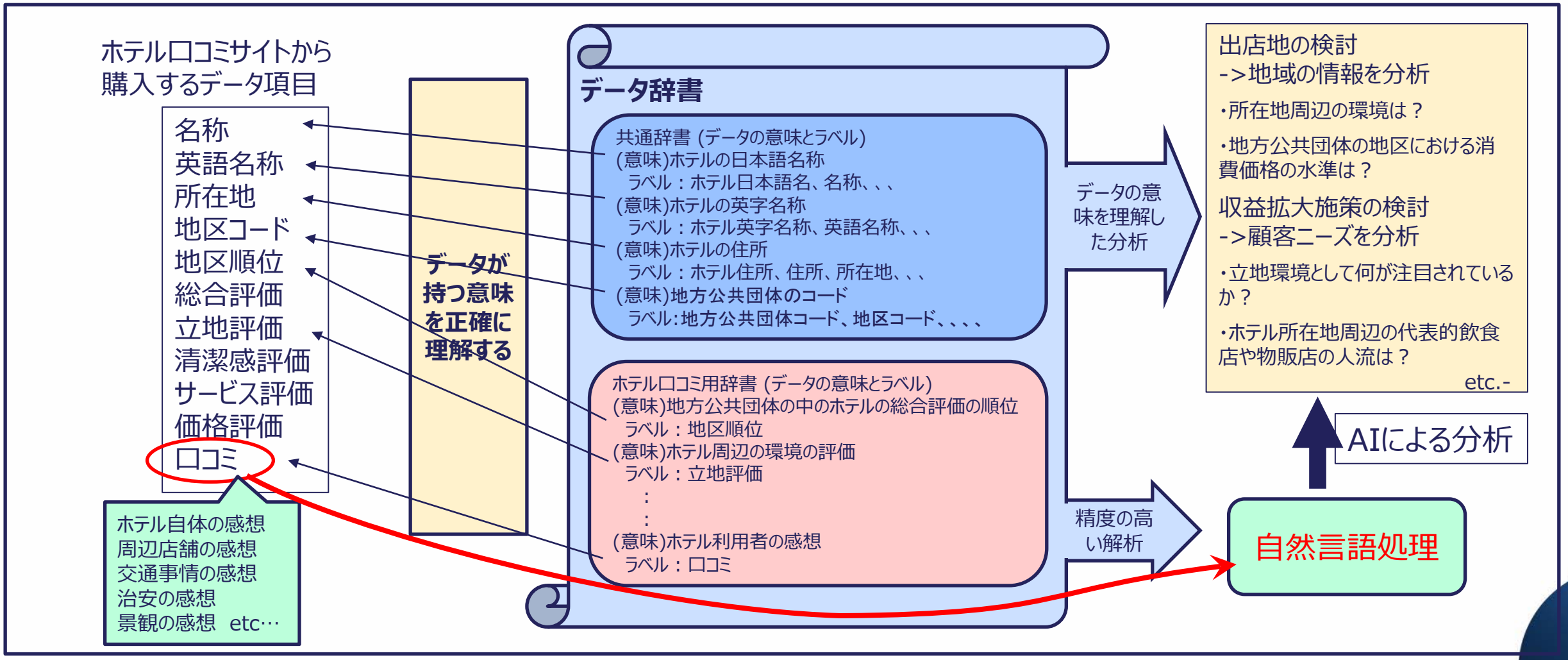
現状システム毎に個別に管理されていたデータの共用を可能にし、貴重なビッグデータである営農情報の有効活用が可能に。



5. データ辞書の活用例

他社から口コミ情報を入手して自社のマーケティングや収益拡大に活用

他社から購入したデータをデータ辞書を用いて正確に理解し、また、自然言語処理へ適用することで、より確かな情報を取得できます。



5. データ辞書の活用例

自然言語処理へのデータ辞書活用

データ辞書は言語処理を行う際にデータ理解促進を促す重要な要素の一つです。

解析対象のテキスト　暖冬により暖房用燃料の販売が低調だった。

↓ 文章を品詞に分割する

一般的な解析

各単語に対して品詞（名詞、動詞、形容詞など）を予測してラベル付けする

暖冬	に	より	暖房	用	燃料	の	販売	が	低調	だっ	た	。

暖冬 → 暖房 or 燃料の販売が減る

精度を高めた解析

各単語へのラベル付けに独自に整備した個別データ辞書も利用

暖冬	に	より	暖房用燃料	の	販売	が	低調	だっ	た	。

暖冬 → 暖房用燃料の販売が減る

個別データ辞書を用いると、より目的に沿った精度の高い言語解析が可能

データ辞書

データの意味とラベル
(意味)冬の平均気温が平年より1度高い
　ラベル：暖冬
(意味)燃やしてその熱を利用するための材料
　ラベル：燃料、たきぎ・木炭・石炭・石油・ガス等

＋

個別データ辞書
データの意味とラベル
(意味)暖房器具で用いる燃料
　ラベル：暖房用燃料、石油・ガス

(参考)金融市場・経済テキストアナリティクス/東京大学大学院　坂地泰紀氏


IPA