

データジャケットにおける変数ラベルの共起性を考慮した 変数ラベル推定手法の検討

Method for Estimating Variable Labels Considering Co-occurrence of Variable Labels in Data Jackets

早矢仕晃章*¹
Teruaki Hayashi

大澤幸生*¹
Yukio Ohsawa

*¹ 東京大学大学院 工学系研究科 システム創成学専攻
Department of Systems Innovation, School of Engineering, The University of Tokyo

It is important to discuss the variables included in datasets in considering the combination of data stored in different domains. A Data Jacket (DJ) is a method for sharing information about data and for considering the potential value of datasets, allowing data itself hidden, by describing the summary of data in natural language. In DJs, variables in datasets are described as variable labels, which is the name/meaning of variables. In the previous study, the information about sets of variable labels, which may help decision making, is not stored as knowledge for those who want to obtain new data. Moreover, due to the lack of information about the variables, some DJs could not have linkage with other DJs through variable labels. In this study, introducing the models based on the features of outlines and variable labels in DJs, we propose a method for estimating variable labels which may be included in DJs, from the information of data which does not include the information about the variables.

1. はじめに

近年、蓄積された膨大なデータを再利用可能な資源と見ること、様々な分析手法を用いて新しい価値を発見し、意思決定に役立てようとする動きが活発になっている。しかし、データ利活用の現場ではデータ管理コストやプライバシーなど様々な問題が指摘されており、領域や分野を横断したデータ利活用やデータ交換は容易ではない[経産省 15]。

データジャケット(以下、DJ)はデータ市場におけるデータ利活用を促すためのメタデータ記述方法として発展してきた[Ohsawa 13]。データに含まれる情報をメタデータとして記述することで変数を含むデータ自体を共有することなく「データに関する情報」の共有により、潜在的なデータの結合案が検討可能となる。DJを用いてデータ利活用に関わるステークホルダーがデータ利活用案を検討可能にするワークショップ手法として、Innovators Marketplace on Data Jackets(以下、IMDJ)[Ohsawa 13]及びアクション・プランニング(以下、AP)[Hayashi 13]が提案されている。IMDJ及びAPでは、データ所有者は自身が保有するデータに関する情報をDJとして提供し、データ利用者あるいはデータ利用方法の提案者は提供されたDJからデータ利活用方法を考案し、評価を行う。以上のデータ利活用案の検討により、データの売買及び交換が行われることが期待される。

データ利活用案検討の中で特に重要となるのがデータに含まれる変数についての議論である。DJでは、データに含まれる変数は、「変数ラベル(Variable Label)」としてメタデータ化されている。変数ラベルとは、データに含まれる変数の意味・名前を意味する。データ利活用方法の検討では、DJの変数ラベルの

組合せから適用する分析ツールや期待する分析結果に関する議論が行われる。

しかし、従来の手法では、新たにデータを取得したい人がどのような変数を取得することが、意思決定に役立つのかという情報は蓄積されてこなかった。また、DJではメタデータとしてデータのタイトル、概要説明、変数の名前、共有条件、データの保存形式などを記述する項目があるが、すべての情報を入力することは強制していない。つまり、DJ登録者(データ保有者)が公開したい情報のみがDJに記述されるため、DJには必ずしも変数名や変数に関する情報が含まれているとは限らない。そのため、変数に関する情報の不足により、本来結合する可能性のあるDJ同士が未結合となってしまうという問題がある。

本研究では、DJに含まれる概要と変数に関する情報(変数ラベル)から変数ラベルの共起性に着目し、2つのモデルについて考察する。そして、変数に関する情報を含まないデータの概要情報から、そのデータに含まれる可能性のある変数ラベルを推定する方法を提案する。変数名や変数に関する情報が含まれていなくても、データの概要情報から変数ラベルを推定することで、データ同士の潜在的な結合可能性を意思決定者に提示できれば、データ利活用を促すことが可能となる。また、新たにデータを取得し、意思決定に役立てたいと考えるプレイヤーに対し、どのような変数を取得することが意思決定において重要であるのかという知見を示すことができる。

2. 先行研究

メタデータからデータ同士の結合を行う試みはいくつか行われてきた。例えば、Linked Open Data(LOD)の分野では、行政が公開するデータをRDF(Resource Description Framework)で記述し、メタデータを介して各データベースに含まれるデータに統一的にアクセスする環境の構築を目指している[Berners-Lee 06, 岡嶋 15]。また、統計データとメタデータ交換関連研究では、データに統一的にアクセスするためのXMLベースの情報モデルを用いて、様々なデータに含まれる変数の組み合わせから統計解析を容易にするためのAPIやメタデータ記述方法を提供している(政府統計の総合窓口 e-Stat など)。以上のように、膨

連絡先: 早矢仕晃章, 東京大学大学院 工学系研究科 システム創成学専攻, teruaki.hayashi@panda.sys.t.u-tokyo.ac.jp
大澤幸生, 東京大学大学院 工学系研究科 システム創成学専攻, ohsawa@sys.t.u-tokyo.ac.jp

This research was supported by JST, CREST.

本研究を支援くださった構造計画研究所(KKE)の皆様にご感謝申し上げます。

大なデータを相互運用可能な形で取り扱い可能にする技術としてセマンティクスが期待されているが、それらはデータが公開されていること、あるいはデータ統合を前提とした議論に留まっている。本稿ではデータは公開されていない状況における領域間のデータ共有の促進、そして新たなデータ取得支援について言及している点で従来研究とは異なる。

また、LOD の分野では、変数ラベルは述語 (predicate) として、変数ごとに統一した語彙の利用を推奨している。しかし、実際に行政が公開しているオープンデータのほとんどは変数ラベルに自然言語を用いており、語彙の統一は図られていない。例えば、変数ラベルである「住所」と「所在地」、「人口」と「人数」は同じ意味であるが異なる語彙で表現されている。厳密に定義した語彙をデータに適用するには、データ取得者の背景知識及び取得意図を考慮する必要があり、これらの作業には膨大なコストと人手がかかることが予想される。また、日々生み出される膨大なデータすべての変数ラベルに統一した語彙を設定することはほとんど不可能である。また定量的なデータだけでなく、アンケートにおける自由回答のように、定性的なデータも数多く存在しており、定義されている語彙数以上の多様な種類の変数が世の中には存在している。

以上の理由により DJ は、データ概要及び変数ラベルなどのメタデータを、人間が意味を理解し表現可能な自然言語による記述で収集している。本稿で用いる DJ のデータ概要及び変数ラベルの例を表 1 に示す。

表 1 DJ に含まれるデータ概要と変数ラベルの一例

データ概要	変数ラベル
東京都が管理する都道における街路灯の設置・管理に関するデータ	緯度
	経度
	光束
	管理番号
	照明種別
毎週月曜日に更新する KSP-POS による全国食品スーパーの最新商品売れ筋ランキングレポート	商品画像
	商品名称
	出現日
	平均売価
	販売店率
	商品カテゴリ
	メーカー名

3. 変数ラベルの推定

3.1 変数ラベルとデータ概要の定義

DJ ではデータの概要情報、つまりメタデータである。データに含まれる変数は変数ラベルというメタデータで表現される。例えば「緯度」、「経度」、「住所」といった既知の変数に関する名前、あるいは「危険度(災害時の活動困難な度合いを項目ごとに評価した値を考慮した総合危険度)」といったデータ固有の変数に関する説明が自然言語によって記述される。本研究では、これらの「データに含まれる変数の名前及び意味」を以降、「変数ラベル」と定義して用いる。また、変数ラベルだけではそのデータが何を表しているのか理解することは難しい。そのため、DJ では、変数ラベルだけでなく、データがどのようなデータであるのかという情報を記述する説明文を収集している。本稿では、データの概要について説明した情報を「データ概要」と定義して用いる。

3.2 変数ラベルのモデル

本節では、データの概要情報からデータに含まれる可能性のある変数ラベルを推定するため、変数ラベルのモデル化について議論する。モデル化の目的は、変数ラベルが未知であるデータ概要から、そのデータに含まれる可能性のある変数ラベルを推定することである。つまり、変数ラベルが未知のデータ概要を入力とし、変数ラベル集合 \mathcal{V} から、関連する変数ラベルの集合 $\{vl \in \mathcal{V} | P(vl)\}$ を出力する機能の実現である。 $P(vl)$ は未知のデータ概要に関連する変数ラベルであるという条件を表す。以下、未知のデータ概要に対して $P(vl)$ を満たす変数ラベルの集合を出力するためのモデルについて、データに含まれる変数ラベルの特徴から考察を行う。本稿では、データ概要及び変数ラベルについて、以下の 2 つの特徴をモデル化する。

- ① データの概要情報の類似度が高いデータ同士は、同じ変数ラベルを有している
- ② ある変数ラベルには同時に登場する頻度の高い変数ラベルが存在し、同時に登場する頻度が高い変数ラベルを有するデータ同士は類似している

例えば、「静岡県庁が提供する街路灯に関する設置データ (DJ_x)」というデータの内容を示す概要情報(変数ラベルは未知)が与えられているとする。ここで、「東京都が設置した街路灯の位置情報 (DJ_1)」及び「東京都における年齢別の人口割合 (DJ_2)」という二つのデータ概要が存在したとき、それぞれのデータ概要が DJ_x と似ている度合いを類似度と呼び、それぞれ $similarity(DJ_x, DJ_1)$, $similarity(DJ_x, DJ_2)$ とする。ここでは、 DJ_x と DJ_1 は街路灯の設置という共通のキーワードを有しているため、 DJ_2 と比較して類似度が高い ($similarity(DJ_x, DJ_1) > similarity(DJ_x, DJ_2)$) とすると、変数ラベルが未知である DJ_x は、 DJ_1 が保有する変数ラベルと同じ変数ラベルを保有している可能性が高いと考えられる。つまり、 DJ_x は街路灯設置情報として街路灯の設置位置を表す「緯度」、「経度」及び街路灯の明るさを表す「光束」という共通の変数ラベルを保有している可能性が高く、モデル①により変数ラベルが未知のデータ概要から、含まれている可能性の高い変数ラベルを推定できる(図 1)。

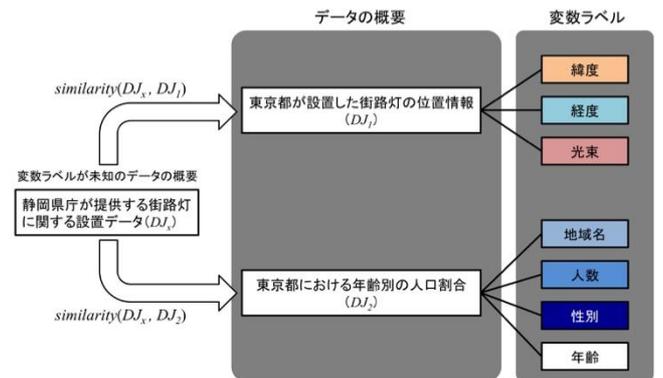


図 1 モデル①の例

続いて、②のモデルについて考察する。例えば、「東京都内のバス停の位置データ (DJ_3)」に含まれる変数ラベルは「緯度」、「経度」、「バス停名」であるとする。一方で、モデル①の例でみた「東京都が設置した街路灯の位置情報 (DJ_1)」に含まれる変数ラベルは「緯度」、「経度」、「光束」であった。すると、これらの 2 つのデータで共通しているのは、変数ラベル「緯度」及び「経度」を含んでいることである。つまり、「緯度」と「経度」の変数ラベルは同時に取得される可能性が比較的高い変数ラベルの組みであるということが出来る。同様に、「会社における男女の構成デ

ータ(DJ_4)というデータが存在し、含まれている変数ラベルは「会社名」、「性別」、「年齢」、「人数」とするとき、「東京都における年齢別の人口割合(DJ_2)」の変数ラベルを考慮すると、「性別」、「年齢」、「人数」は同時に登場する頻度が高い(共起度が高い)変数ラベルの組みであるといえることができる。さらに、 DJ_1 と DJ_3 はどちらも何かの位置情報を含んでいるデータであり、 DJ_2 と DJ_4 はある領域における人口に関するデータであることを考えると、同時に登場する頻度が高い変数ラベルを有するデータ概要は類似している可能性が高いといえる(図2)。

変数ラベルが未知であるデータ概要から、変数ラベルを推定するだけであれば、モデル①のみで十分と考えられるかもしれない。しかしモデル①のみでは、推定性能がデータ概要の記述量に大きく影響されてしまうという問題がある。つまり、ほとんどの単語の特徴量が0であるデータ概要ベクトルは、記述量の多いデータ概要とのみ類似度が高くなり、それらに含まれる特定の変数ラベルのみが推定結果に現れてしまう。また、類似性の高い内容を表すデータであってもデータ概要の表記揺れなどにより、類似度の計算が困難となる問題も考えられる。例えば、「アメリカの子どもに関するデータ」と「日本における児童の統計情報」は、「年齢」、「性別」、「身長」、「体重」などの共通した変数ラベルを有している可能性が高い。しかし、この2つのデータ概要には共通の単語が含まれておらず、類似度を計算するためにはシソーラスなどのより高度な辞書が必要となってしまう。そこでモデル②を用いることで、共通する変数ラベルからデータ概要の類似性を評価することが可能となる。

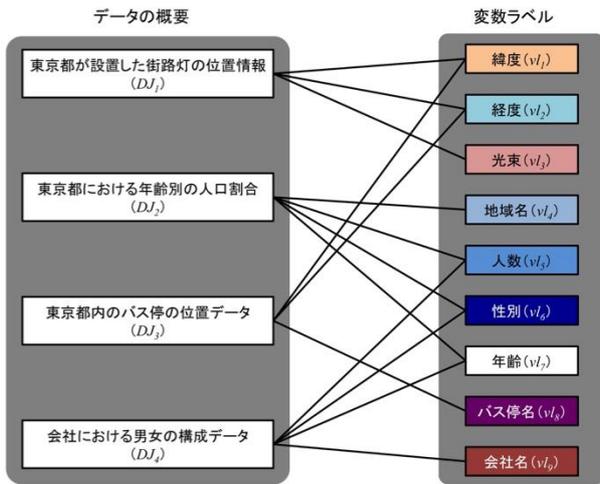


図2 モデル②の例

3.3 変数ラベル推定プロセス

本節では、前節の議論を踏まえ、モデル①及びモデル②を仮定し、データ概要の類似度及び変数ラベルの共起度から未知のデータに含まれる変数ラベルを推定する方法を説明する。

まず、モデル①に基づき、データ概要の類似度を計算するアルゴリズムについて考える。データ概要集合に出現する単語の集合に対して、任意の単語のデータ概要における出現頻度をベクトル空間モデルに基づいて表現する。具体的には、過去に登録されたDJをコーパスとして用い、形態素解析によりデータ概要内のW件の単語を行ベクトル、D件のデータ概要を列ベクトルとした単語-データ概要行列M(W×D行列)を作成する(式(1))。データ概要の列ベクトル \mathbf{dj}_j の要素 w_{ij} は、単語 t_i がデータ

概要 DJ_j に出現する回数から重み付けした値を表す(式(2))。なお、行列及びベクトルの右上の添字Tは転置を表すとする。

$$M = (\mathbf{dj}_1, \mathbf{dj}_2, \dots, \mathbf{dj}_j, \dots, \mathbf{dj}_D) \quad (1)$$

$$\mathbf{dj}_j = (w_{1j} \ w_{2j} \ \dots \ w_{ij} \ \dots \ w_{Wj})^T \quad (2)$$

続いて、データに含まれる変数ラベルの集合に対して、データ概要における変数ラベルの出現回数(0 または 1)をベクトル空間に表現する。図2で表したように、DJにはデータ概要とそれに含まれている変数ラベルが紐付いている。それらを用い、データ概要に紐付くV件の変数ラベルを行ベクトルとし、D件のデータ概要を列ベクトルとした行列R(V×D行列)を作成する(式(3))。データ概要の特徴ベクトル \mathbf{dj}'_j の要素 w'_{ij} は、変数ラベル vl_i がデータ概要 DJ_j に出現する回数を表す(式(4))。

$$R = (\mathbf{dj}'_1, \mathbf{dj}'_2, \dots, \mathbf{dj}'_j, \dots, \mathbf{dj}'_D) \quad (3)$$

$$\mathbf{dj}'_j = (w'_{1j} \ w'_{2j} \ \dots \ w'_{ij} \ \dots \ w'_{Vj})^T \quad (4)$$

以上の手順により求めた単語-データ概要行列M(W×D行列)及び変数ラベル-データ概要行列R(V×D行列)から、単語-変数ラベル行列I(=MR^T)(W×V行列)を作成する。これは、行列R^T内のD次のデータ概要空間にある変数ラベルの特徴ベクトルを行列Mによって、W次の単語空間に写像することに相当する。行列Mによる写像をV件のすべての変数ラベルの特徴ベクトルに適用し、単語-変数ラベル行列Iを得る。つまり、単語-データ概要行列Mに変数ラベル-データ概要行列Rを転置して右からかけることによって得られる(式(5))。単語-変数ラベル行列Iの要素 v_{ij} は、単語 t_i と変数ラベル vl_j の両方を含むDJの数を表している(式(6))。

$$I = MR^T = (vl_1, vl_2, \dots, vl_j, \dots, vl_V) \quad (5)$$

$$vl_j = (v_{1j} \ v_{2j} \ \dots \ v_{ij} \ \dots \ v_{Wj})^T \quad (6)$$

続いて、モデル①とモデル②を組み合わせることを考える。以下、同じDJ内に含まれる任意の2つの変数ラベルは1回共起しているものとする。モデル②では、変数ラベルの共起度を考慮するため、変数ラベル vl_i と変数ラベル vl_j が同時に登場するDJ数を表す変数ラベル共起行列C(=RR^T)(V×V行列)を定義する(式(7))。変数ラベル共起行列Cの要素 v'_{ij} は変数ラベル vl_i と変数ラベル vl_j を含むDJの数を表している。図3にて、図2で例示したデータ概要と変数ラベルの紐付けから変数ラベル共起行列を作成する例を示す。

$$C = RR^T = \sum_{k=1}^D \mathbf{dj}'_k \mathbf{dj}'_k{}^T \quad (7)$$

変数ラベル共起行列C									変数ラベル-データ概要行列R				データ概要-変数ラベル行列R ^T											
v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	dj_1	dj_2	dj_3	dj_4	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9			
v_1	2	2	1	0	0	0	0	1	0	v_1	1	0	1	0	dj_1	1	1	1	0	0	0	0	0	0
v_2	2	2	1	0	0	0	0	1	0	v_2	1	0	1	0	dj_2	0	0	0	1	1	1	1	0	0
v_3	1	1	1	0	0	0	0	0	0	v_3	1	0	0	0	dj_3	1	1	0	0	0	0	0	1	0
v_4	0	0	0	1	1	1	1	0	0	v_4	0	1	0	0	dj_4	0	0	0	0	1	1	1	0	1
v_5	0	0	0	1	2	2	2	0	1	v_5	0	1	0	1										
v_6	0	0	0	1	2	2	2	0	1	v_6	0	1	0	1										
v_7	0	0	0	1	2	2	2	0	1	v_7	0	1	0	1										
v_8	1	1	0	0	0	0	0	1	0	v_8	0	0	1	0										
v_9	0	0	0	0	1	1	1	0	1	v_9	0	0	0	1										

図3 図2の変数ラベル共起行列

単語-変数ラベル行列 I ($W \times V$ 行列) に変数ラベル共起行列 C ($V \times V$ 行列) を右からかけることで、変数ラベルの共起度を考慮した $W \times V$ の単語-変数ラベル行列 IC を得る (式(8)). 行列 IC の列ベクトルは W 次元の変数ラベルの特徴ベクトル vl'_j を表している. 単語-変数ラベル行列 IC の要素 v''_{ij} は、データ概要の類似度(I による作用)と変数ラベルの共起度(C による作用)を考慮し、ある単語 t_i が与えられたときの変数ラベル vl'_j の重みを表している(式(9)). 具体的には、データ概要、変数ラベルを頂点集合とする2部グラフにおいて、単語 t_i を含むデータ概要から到達可能な変数ラベルの頂点に至り(I による作用)、続いてそれらの変数ラベルの頂点からデータ概要の頂点を経由して変数ラベル vl'_j の頂点に至る(C による作用)経路数が v''_{ij} である. つまり、モデル①は単語 t_i を含むデータ概要の頂点から変数ラベルの頂点への到達可能性と経路数を計算している. モデル②はモデル①で計算された到達可能な変数ラベルの頂点からデータ概要の頂点を経由して自身を含む他の変数ラベルの頂点への到達可能性と経路数を計算していることと同義である.

$$IC = MR^T RR^T = (vl'_1, vl'_2, \dots, vl'_j, \dots, vl'_V) \quad (8)$$

$$vl'_j = (v''_{1j} \ v''_{2j} \ \dots \ v''_{ij} \ \dots \ v''_{wj})^T \quad (9)$$

以上より、未知のデータ概要 DJ_x が与えられたとき、データ概要の類似度(モデル①)及び変数ラベルの共起度(モデル②)を考慮して関連する変数ラベル集合 $\{vl \in \mathcal{V} | P(vl)\}$ が取得できる. 実際には、 DJ_x を特徴ベクトル dj_x (W 次元)に変換し、単語-変数ラベル行列 IC に含まれる変数ラベルの特徴ベクトル vl'_j との類似度($\text{similarity}(vl'_j, dj_x)$)を算出し、未知のデータ概要 DJ_x に対して類似度の高い変数ラベルを取得する仕組みとなる.

4. 実データによる推定例

提案手法を用い、「日本のある地域における年ごとの人口の推移を表すデータ」というデータ概要が与えられたときに推定された変数ラベルの例を示す. 比較のため、モデル①のみ(単語-変数ラベル行列 I)を用いて推定された変数ラベルも提示する. モデル①のみから変数ラベルは、入力したデータ概要に対して類似度の高いDJの集合を取得し、それらに含まれる変数ラベルの集合を取得することを意味する.

変数ラベルの推定のための訓練データには、データ概要及び変数ラベルの両方を含む622件のDJを用いた. DJの変数ラベルは3201種類あり、1件のDJは平均して6.69件(標準偏差6.08)の変数ラベルを持つ. 形態素解析器にはMeCab¹を用いた. データ概要内の特徴的な単語の重み付けにはtf-idfを用い、類似度の計算にはコサイン類似度を用いた. なお、訓練データには「日本のある地域における年ごとの人口の推移を表すデータ」というデータ概要は含まれていない.

表2は変数ラベルが未知のデータ概要に含まれる可能性の高い変数ラベル上位10件のリストである. 提案手法では、入力したデータ概要に含まれる可能性の高い変数ラベルが提示されていることが分かる. 一方、モデル①のみを利用した推定方法では、「農業人口」に関するデータなどの類似度の高い訓練データのデータ概要が影響し、入力したデータ概要に含まれる可能性が高くない変数ラベルが提示されている. データ概要の類似度だけでなく、変数ラベルの共起性を考慮することによって、データに含まれる可能性の高い変数ラベルを推定することができるという示唆が得られた.

表2 「日本のある地域における年ごとの人口の推移を表すデータ」というデータ概要から推定された変数ラベル

提案手法 (モデル①と②の組み合わせ)	モデル①のみ
年齢区分(5歳毎)	人口(女)
人口(女)	人口(男)
人口(男)	人口(総数)
人口(総数)	農家総人口
出生数	農業就業人口
死亡数	農業就業人口(男)
転入者数	農業就業人口(女)
死亡者数	専業農家数
転出者数	兼業農家数
人口	年(5年毎)
年月日	前回増減
世帯数	流入人口

5. まとめ

本研究ではデータジャケットに含まれるデータの変数の名前・意味を変数ラベルと定義し、変数ラベルが未知のデータ概要から、そのデータに含まれている可能性の高い変数ラベルを推定する方法を提案した. データ概要の類似度だけでなく、保有する変数ラベルの共起性を考慮したモデルを導入することで、含まれている可能性の高い変数ラベルを推定することができた. 今後は様々なデータ概要から変数ラベルが推定し、本手法の性能を評価する実験を行う予定である.

本論文では変数ラベルの共起度を考慮したモデルを提案したが、変数ラベルの類似度は考慮しなかった. 表2に示したように、「人口」と「人口(男)」、「人口(女)」は非常に近い関係にあり、「死亡数」と「死亡者数」は同義であると考えられる. 変数ラベルの類似度を考慮することが今後の課題として挙げられる.

参考文献

- [経産省 15] 経済産業省: 平成26年度経済産業省委託事業、我が国経済社会の情報化・サービス化に係る基盤整備(データ駆動型イノベーション創出に関する調査事業)調査報告書, 2015. http://www.meti.go.jp/meti_lib/report/2015fy/001102.pdf, [最終アクセス2016年3月16日].
- [Ohsawa 13] Ohsawa, Y., Kido, H., Hayashi, T., and Liu, C.: Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, Vol.22, pp.709-716, 2013.
- [Hayashi 13] Hayashi, T., and Ohsawa, Y.: Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game Plus Action Planning, International Journal of Knowledge and Systems Science, Vol.4, No.3, pp.14-38, 2013.
- [Berners-Lee 06] Berners-Lee, T.: Linked Data – Design Issues, 2006. <https://www.w3.org/DesignIssues/LinkedData.html>, [最終アクセス2016年3月16日].
- [岡嶋 15] 岡嶋成司, 山根昇平, 糸照宣: Linked Data活用を促進するプラットフォーム, 人工知能, Vol.30, No.5, pp.568-573, 2015.

¹ <http://taku910.github.io/mecab/>