1B5-4

上位オントロジーに基づく生物表現型データ記述の考察

A inquiry of the methodology to describe biological measurements with top-level ontologies

桝屋啓志[*1] 田中信彦[*1] 脇和規[*1] 櫛田達矢[*2] 古崎浩司[*3] 溝口理一郎[*3]
Hiroshi MASUYA Nobuhiko TANAKA Kazunori WAKI Tatsuya KUSHIDA Kouji KOZAKI Riichiro MIZOGUCHI

[*1] 理研バイオリソースセンター RIKEN Bioresource Center
[*2] ナラプロテクノロジーズ NalaPro Technologies, Inc.
[*3] 大阪大学産業科学研究所 The Institute of Scientific and Industrial Research (ISIR) Osaka University

Toward the understanding of the whole life systems, it is revealed one of the new challenges for the bioinformatics studies that the development of the methodology to integrate broad range of biological measurements such as dynamic state of molecules, phenotypes, experimental conditions and environments. In this context, it is desired the sophisticated ontological framework to define quality-related concepts such as attributes and quality values. In this study, we tried expansion of one of OBO ontologies, PATO, to ensure more advanced framework to integrate broad quality descriptions in experimental data in biomedical studies, as well as the facilitation of interoperability with non-biological community through the ontology mapping PATO terms into YAMATO's logical framework.

1. はじめに

近年、ゲノム科学の進展に伴って、生物学分野では大規模かつ網羅的に収集されたデータを解析するシステム生物学的アプローチが注目されている。従来、生物学では、あらかじめ立てられた仮説に従って立証を行う小規模な解析が主体であったが、DNA の転写や物質の代謝などの分子動態を網羅的に解析する技術が発達したことで、一実験で大量のデータを取得する事が可能になり、データに基づいて推論を行う仮説発見型の研究が成功をおさめるようになりつつある。そこで、さらに大量かつ多種多様なデータを統合して生物の網羅的な特性プロファイルを作成することができれば、これを総合的に解析する事で、新たな生物機能の発見や、最終的には生命全体をシステムとして理解することにつながると期待されている。本稿では、すでに生物学分野で測定データを生物種横断的に分類整理する目的で利用されている Phenotypic Quality オントロジー(PATO)について、最新の上位オントロジーである Yet Another More Advanced Top-level Ontology (YAMATO)に基づいた考察と検討を行った。

2. PATO オントロジー問題点の検討

膨大な解析データの統合的な利用を目指し、Open Biomedical Ontology コンソーシアムが中心となって、オントロジーの整備が進められている。PATO は、表現型記述に焦点を当てながらも、長さ、重さ等、実験科学で用いられる一般的な定性的特性を分類したオントロジーであり、測定データを生物種横断的に分類整理する目的で用いられている [Gkoutos 05][Mungall 10]。また、上位オントロジーである Basic Formal Ontology (BFO) [Grenon 04]に基づいており、定性的特性を1つの分類ツリーでモデル化している。ツリーの上位では、属性（length, weight など）が分類されており、これがある程度分類されたところで、その下位に定性値である（increased length／decreased length）が配置されていて、これは属性と値を根本的には区別しないことを意味している。上位の属性の分類では、生物学における測定や観察項目を広くカバーしており、ある程度成功していると考えられるが、一方で PATO は、下記に示すようないくつかの問題を抱えている。

2.1 定性値と定量値の統合における問題

第 1 の問題点は、PATO に「値」の分類が存在しないことである。多くの自然科学同様、生物学研究においても、計測結果として多様な値の記述が存在する。例えば、統計学的には、値を記述する尺度として、順序も加減などの演算もできない単なるカテゴリである「名義尺度」、順序（大きいか小さいか）比較ができるが加減などの演算ができない「順序尺度」、加減演算のみ可能な「間隔尺度」、加減に加え乗除演算も可能な「比例尺度」といった値の分類がある。これらの値は長さ、重さ‥といった属性の分類とは独立であり、これらの分類は、データの統合には不可欠と考えられる。BFO および PATO における 1 つのツリーのモデルでは、値を分類することは困難であり、結果的に定性値と定量値を効率的に統合することができない。これに対して、別の上位オントロジーである Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)[LOA]では、*quality-space*（属性）と *quale*（値）を区別するモデルを採用している。このモデルでは属性の分類と値の分類の2つのツリーに分かれており、値の分類が可能である。また、属性と値の組み合わせによって、例えば、「長さ」という性質に対して、「10cm」という定量値と「長い」という定性値の双方を関連づけること、つまり、定性値と定量値を統合的に扱うことができるので、自然科学分野の性質記述には、こちらの方が実用的であると考えられる。

2.2 定性値のコンテキストの区別と分類における問題

もうひとつは順序尺度値（定性値）の目的依存性の問題である。例えば、アリにおいて身体が大きくなる異常と、ゾウにおいて身体が大きくなる異常とは、進化的に相同な成長因子の変異など、共通した原因が推論される一方で、たとえいかに大きなアリであっても、小さなゾウに比べれば、極めて小さいため、大きい／小さいという定性値を扱うコンテキストの整理が必要である。このような「大きい」「小さい」という概念の目的依存性は PATO では扱われていない。

生物学研究における定性値コンテキストを考えた場合、大きく下記の2種類に分けられる。ひとつは、偏差あるいは変位に依存する定性値である。例えば、「正常と異常」のように、特定の生物種内での deviation に基づいて閾値を定め、それとの大小関係により定められる定性値などで、表現型の記述の大部分はこ

連絡先：桝屋啓志，理化学研究所バイオリソースセンター・マウス知識化研究開発ユニット，〒305-0074 茨城県つくば市高野台 3-1-1，Tel: 029-836-9013，FAX: 029-836-9017，E-mail: hmasuya@brc.riken.jp

の範疇に入る。PATO も主としてこのようなコンテキストにおける定性値を想定して作成されており、「重い」を意味する increased weight (PATO:0000582)は、「異常に重い」に相当する。もうひとつは、このような偏差に依存しない単純な相対比較である。例えばアリとゾウを比較するとゾウが大きいというような場合で、生物進化を論じる場合に用いられる。

これらのコンテキストをさらに注意深く検討すると、コンテキストをさらに詳細に分類できることが分かる。例えば、「生物種内での偏差に準じて」というコンテキストの下位には、「アリにおける偏差に準じて」、「ゾウにおける偏差に準じて」というコンテキストが定義可能である。これらのコンテキストは明確に区別されると同時に、類縁関係がある。例えば、ヒトでの人差し指が長い異常と、マウスでのヒトでの人差し指が長い異常との間では、生物学的な意味が類似していると推測されるし、実際にこのような表現型比較から、同じ遺伝子の機能異常をつきとめた例も少なくない。このように、コンテキストの分類関係とそれぞれのコンテキストにおける定性値の対応関係は重要な意味を持っており、これを一貫した方法論で記述できるフレームワークを開発できれば、生物分野での情報統合に大きく貢献できると考えられる。

2.3 生物の性質を記述する「測定データ」定義の必要性

実験科学では、我々の知りうるのは常に誤差付きデータであり、「真の値」は推測するしかないものとされている。生物の測定データをオントロジーの世界に存在させる際に、直接的に「生物の性質＝真の値」として存在させるべきか、あるいは、性質を記述した「データ」として存在させるべきかについては、PATO では明確な指針が示されていない。しかし、「生物の性質」と「データ」は明確に区別すべき全く異なる概念である。我々は、実験科学の測定結果記述において、「データ」の定義は下記の理由で必須であると考えている。

もしも、「データ」を定義せず、「生物の性質」のみを存在させるとすると、一見シンプルに情報を記述できるように見えるかもしれない。しかしながら、実験によって得られた結果としての値を、そのまま物が持つ性質、すなわち実験科学で言うところの「真の値」に代入することとなる。この場合、同じ測定対象に対する食い違う結果を同時に代入する事は、性質の定義に反し、推論に本質的な矛盾を生じると考えられるので不可能である。このような方法では、例えば、測定後の計算処理等によって異なる結果が導き出された場合に記述ができない。

さらには、以下のような場合にも不都合がある。異なる測定方法 A と B で、マウス体重 W を交互に計って体重増加を記録したと仮定する。時間軸を Tn とし、常に A は B より値が大きく出るとする。これらの測定の結果、w1(a,t1), w2(b, t2), w3(a, t3), w4(b. t4)….を考えた場合、上記の方法では、全てが真の値となるため、体重が w1(a,t1), w2(b, t2), w3(a, t3), w4(b. t4)….と順番に変化したと解釈される。常識的には、A による測定結果 wi(a, ti)と B による措定結果 wj(b, tj)を区別し、真の値を推定すると考えられるが、これを記述する事ができない(図1)。

これに対して、測定結果を、「データ」として存在させる場合は、測定値が真とは限らないことを示すことができ、さらに対象物に複数の測定値を関連づけること、異なる計測方法によって異なる結果がもたらされたことなどを正しく記述することが可能である。よって、この方法は実験科学の哲学を正しく反映するだけでなく、測定データ記述に現実に即した拡張性と柔軟性を与えると考えられる。

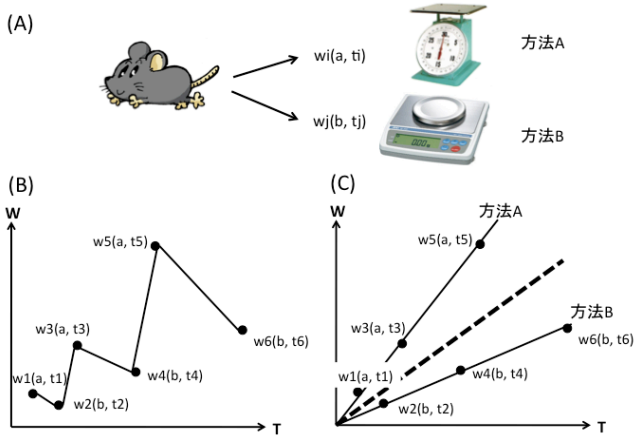

図1 マウス体重を異なる方法で交互に測定したとき(A)、データを定義しない場合(B)とした場合(C)での解釈の違い

3. 上位オントロジーYAMATO による PATO の拡張

PATO は生物学コミュニティで広く受け入れられており、表現型の記述を生物横断的に標準化することに大きく貢献しているが、今後生物学分野でシステム生物学的なアプローチを拡張するにあたって、上に述べた問題に対応していないことが障壁となって、測定データの統合に問題を生じることが予想される。そこで我々は、最新の上位オントロジーである Yet Another More Advanced Top-level Ontology (YAMATO)に基づいて、PATO を拡張することを試みている。YAMATO は、現存する上位オントロジーを詳細に考察し、各オントロジーが含む概念を1つのクラスツリー上で分類し、相互関係を記述することにより、異なる上位オントロジー間の相互運用性をもたらすことを目指して開発されており[Mizoguchi 09]、性質関連概念に関しては下記のような特徴を持っている。

- YAMATO では、性質関連概念(*質_量*)は*特性*、*属性*、*性質値*などに分類されるが、これらと、BFO, DOLCE を含む複数の上位オントロジーの性質関連概念の関係が明示されている(図2)。例えば、BFO の Quality は、*性質値*を内含する*質*である、*特性*に分類される[垂見 08][Mizoguchi 09]。
- YAMATO では性質値の分類が行われており、これが、上述の尺度の分類と一致している(名義尺度値＝*カテゴリカル*、順序尺度値＝*定性値*、間隔尺度値＋比例尺度＝*定量値*)。

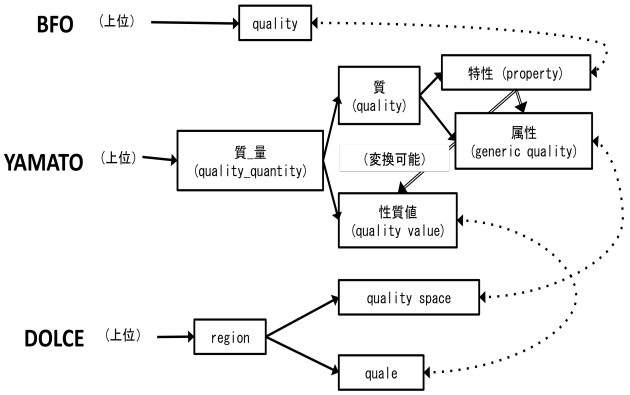

図2 BFO, YAMATO, DOLCE における性質関連概念の相互関係。矢印は is_a(特殊化)リンク、点線矢印は概念同士が相同であることを示す.

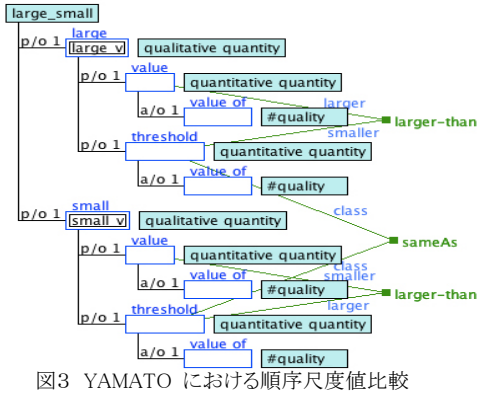

図3 YAMATO における順序尺度値比較コンテキストのモデル。

- 基本概念がコンテキストに依存した役割(ロール)を担った状態にあることを示すロールホルダーという概念を用いて、例えば、*定性値*コンテキスト(*large_small*)に依存して、大きい値(*large_v*)、小さい値(*small_v*)などをとることがモデル化されている(図3)。
- BFO や DOLCE では、モノが持つ性質は定義されているが、オントロジーの世界の中で、モノが持つ性質を記述することで生じた測定データとの分離が明示されていない。YAMATO では、準抽象物である表現の下位概念として性質表現が定義されていて、モノの持つ性質(真の値)とデータとが明確に分離されて定義されている。

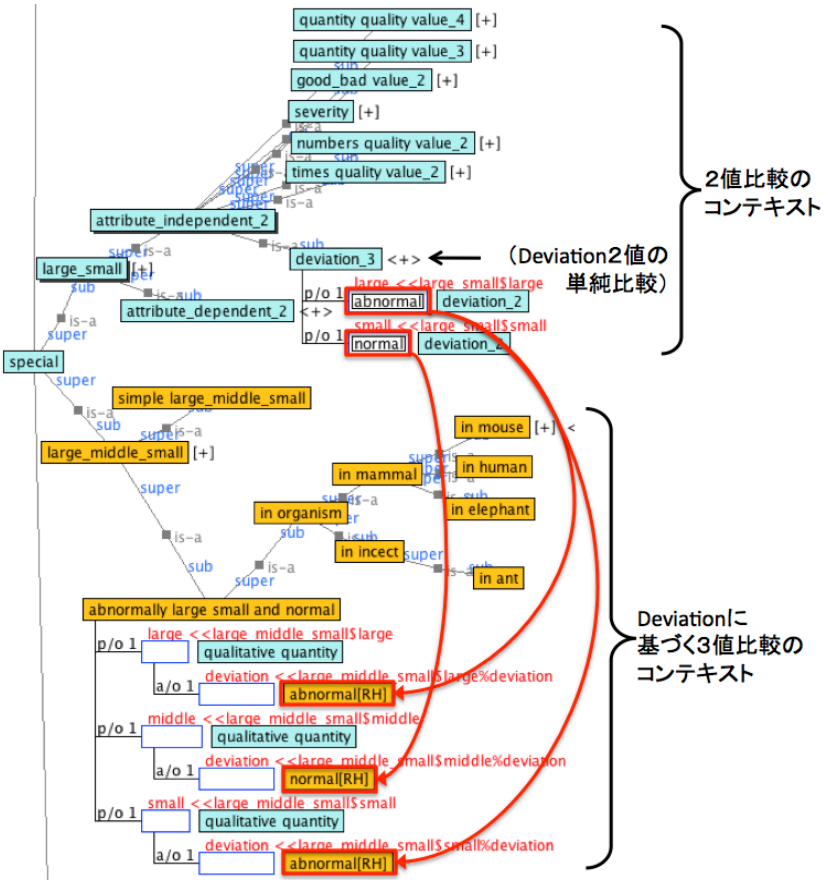

図4 PATO2YAMATO における順序尺度値比較のコンテキスト。deviation2値の単純比較における normal/abnormal を参照して、deviation に基づく3値比較のコンテキストツリーを定義した。

我々は、このような特長を利用して、PATO の抱える問題点を解決する最も有効な手段として、PATO の各概念を YAMATO フレームワークにマッピングした参照オントロジー作成を試みた。手順の概要は以下の通りである。

1) PATO の OWL 形式ファイルより、全概念を*特性*の下位にインポートする。
2) インポートした特性から、対応する*属性*と*性質値*を生成する。生成にあたっては、PATO は以前 DOLCE モデルを採用していた経緯があり、内部に DOLCE の quality-space に相当する"attribute_slim"、quale に相当する"value_slim"のフラグが付加されているので、これを利用する。
3) *定性値*に関しては、2.2 節で論じた通り、コンテキストのツリーを作成し(図4)、各コンテキストの中でロールホルダーとして定義する。今回は特に、マウスにおける偏差に基づいた3値比較である、*in mouse* を定義し(図4)、その下位に、PATO の各定性値のコンテキストを定義する。

4. 参照オントロジー、"PATO2YAMATO"

以下に上記の手順により作成した参照オントロジー"PATO2YAMATO"について解説する。

4.1 定量値とコンテキスト依存的な定性値を統合的に扱うフレームワーク

図5は、increased weight (PATO:0000582)を例にして、PATO からインポートした*特性*と、*属性*、*定性値*、および比較コンテキストがどのような関係になるかを示している。(ただし、PATO では、PATO2YAMATO のように、定性値の比較コンテキストが定義されている訳ではない。この図では、「マウスの異常に重い値」が PATO を使って記述された場合を想定している。) *特性*である PATO:0000582 は、*increased weight* を内含しており、とりうる値の範囲が同じである事を意味している。*increased weight* は、比較コンテキストである *weight value in mouse*(図4参照)において、*定性値 weight quality value* が"演じる"ロールホルダーであり、「正常マウスより重い」という large ロールで定義された要素を持つ。また、上述のようにそれぞれの比較コンテキストは、図4に示されるように相互関係が定義されており、ヒトにおいて身体が大きくなる異常と、マウスにおいて身体が大きくなる異常とは、生物種内での変位に基づく大きさの比較が同じ哺乳類の中で行われているという意味でとても近い意味がある、ということを記述している。また、このような変位に基づく大きさの比較と、単純な大きさの相互比較とはコンテキストが全く異なることも明示されている。

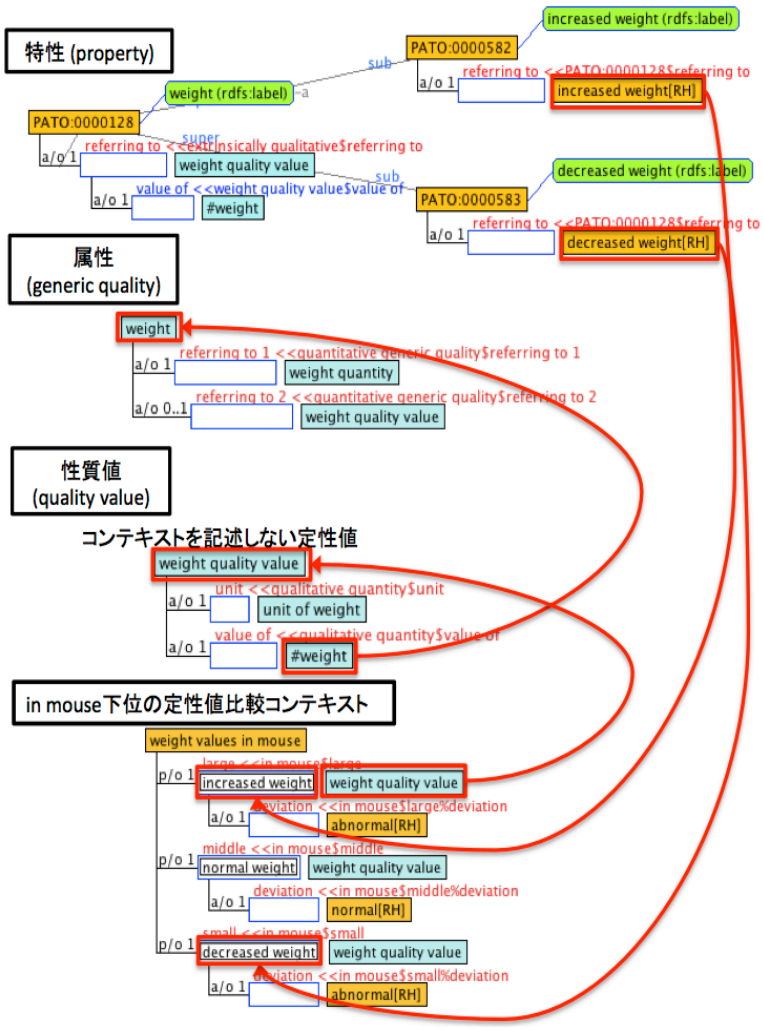

図5 PATO 概念と、新たに生成された DOLCE/YAMATO 方式の性質関連概念との関係

4.2 性質表現概念による測定データの定義

BFO や DOLCE では、それぞれ、モノは、＜実体，特性＞あるいは＜実体，属性，値＞、という構造で性質を持っていると主張している。これに従って、実際の性質記述でもこれらの構造が採用される事が多い。しかしながら、どちらの上位オントロジーも、それぞれの世界の中で、真の値とは異なるデータとしての「性質の記述」がどのように存在するかを定義してはいない。YAMATO では、上述の通り、「データ」を扱う概念として、「*性質表現*」が定義されていおり、表現の形式である*表現形態*と、その意味する内容で構成されている。例えば、文章表現の*表現形態*は、シンボル列である文字列であり、その内容は物語や理論などである。同様に、DOLCE における*性質表現*の*表現形態*は＜実体，属性，値＞という形式、つまり、それぞれの概念を記号化したシンボルの3つ組であり、内容は測定行為の対象である*性質値*(真の値)に対応している。YAMATO は BFO、DOLCE を含む複数の上位オントロジーの性質概念を網羅しているので、*性質表現*においても＜実体，特性＞方式、＜実体，属性，値＞方式の双方が定義され、相互の関係も定義されている。

PATO2YAMATO で定義された性質概念は、この*性質表現*を用いる事で、2.3 節で述べたように「測定データ」として、オントロジーの世界に存在させる事ができる。OBO コミュニティでは PATO を採用する複数のデータベースにおいて、＜実体，特性＞の方式で測定データや表現型のアノテーションがすでに進行しているが、このようなデータも*性質表現*としてインポートして、＜実体，属性，値＞方式との相互の関係性の記述や変換を行う事が可能になると考えられる。

5. おわりに

以上、YAMATO のフレームワークを用いることにより、PATO の抱える問題を解決できること、参照オントロジーである PATO2YAMATO を用いて、PATO の問題点を解決した知識フレームワークを提示できる事を示した。近年、分子動態の測定技術や、画像解析技術の発展により、生物学研究のデータ記述は定量的記述へとシフトするといわれている。しかしながら、ほぼ全ての生物データは、コントロールとの比較により「評価」されるので、コンピュータによる知識処理では、定性値化が必須のプロセスとなると考えられる。また、上に述べた生物の網羅的特性プロファイリングを行う際には、論文から抽出した知識などの定性的な記述を反映することも必要である。この際、比較コンテキストの記述は、極めて重要な事項であり、正しく整理されないと重大な間違いをもたらす恐れがある。これは、定性的な記述のみならず、偏差や変位量などの量的記述においても重要であるので、今後さらに詳細な検討が必要になると考えられる。

現在我々は PATO2YAMATO の作成作業を継続中である。今後さらに、幅広いデータ統合へ向けた性質概念の体系化を進めるとともに、臨床医療オントロジー[国府 08]等、疾患を扱うオントロジーと連携させることによるモデル生物における表現型と疾患との関連性の推論を試みていきたいと考えている。

参考文献

[Grenon 04] Grenon,P. and Smith,B: SNAP and SPAN: towards dynamic spatial ontology. Spat. Cogn. Comput. 4, pp69–103, (2004)

[Gkoutos 05] Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D: Using ontologies to describe mouse phenotypes, Genome Biol, 6, R8. (2005)

[LOA] Laboratory for Applied Ontology, http://www.loa-cnr.it/DOLCE.html

[Mizoguchi 09] Mizoguchi R: Yet Another Top-level Ontology: YATO, Proc. of the Second Interdisciplinary Ontology Meeting, pp.91-101 (2009)

[Mungall 10] Mungall CJ, Gkoutos GV, Smith CL, Haendel MA, Lewis SE, Ashburner M: Integrating phenotype ontologies across multiple species, Genome Biol., 11, R2 (2010)

[国府 08] 国府, 周, 古崎, 今井, 大江, 溝口: 臨床医療オントロジーの構築に関する基礎的な考察 人工知能学会第 22 回全国大会,2E3-1,(2008)

[古崎 02] 古崎, 來村, 池田, 溝口:「ロー ル」および「関係」に関する基礎的考察に基づくオントロジー記述環境の開発, 人工知能学会論文誌, Vol.17, No.3, pp.196-208, (2002)

[垂見 08] 垂見, 古崎, 來村, 溝口: 知識構造化システムにおける機能と性質に関するオントロジー的考察 人工知能学会第 22 回全国大会 ,3F1-1,(2008)