3Q1-4in1

生命科学系画像メタデータの標準化に向けた顕微鏡オントロジーとLODデータベースの開発

Development of Microscopy Ontology and Metadatabase for Standardization of Imaging Metadata in Life Science

久米 慧嗣[*1*2] Satoshi Kume　桝屋 啓志[*3*4] Hiroshi Masuya　片岡 洋祐[*1*2] Yosky Kataoka　小林 紀郎[*2*3*4] Norio Kobayashi

[*1] 理研 CLST RIKEN CLST　[*2] 理研 CLST-JEOL 連携センター RIKEN CLST-JEOL Collaboration Center　[*3] 理研 BRC RIKEN BRC　[*4] 理研 ACCC RIKEN ACCC

Imaging data is one of the most important fundamentals in the current life sciences. Here, we constructed a microscopy ontology in Resource Description Framework (RDF) / Web Ontology Language (OWL) based on the XML-based Open Microscopy Environment (OME) data model. Our ontology is an extension that can describe not only imaging metadata for optical and electron microscopy, including biosamples and experimental conditions. Then, we generated the metadata for electron microscopy images of the tissues such as the liver and kidney in RDF through the RIKEN Metadatabase platform and globally published linked open data.

1. はじめに

細胞等を撮像した顕微鏡画像データは、生命現象を細胞の微細構造や形態から理解するための最も基本的かつ重要なデータの1つである。顕微鏡画像の中でも、電子顕微鏡によって得られる画像データは、生体組織や細胞の形態学的な情報をナノスケールで含むことから、細胞核やミトコンドリアなどの細胞小器官の情報、神経細胞のネットワークやシナプス構造など、様々な組織の形態学的な微細構造を詳細に観察することが可能である [Kasthuri 2015]。

著者らの研究チームは、日本電子株式会社(JEOL Ltd.)と連携センターをつくり、走査型電子顕微鏡(SEM)の技術開発(自動撮像、オートフォーカス技術など)を進め、生体組織試料において、その微細構造の大量撮像を可能にしてきた。現在、SEMを利用して、哺乳類(マウスやラット)の脳組織や肝臓などの生体組織画像を広範囲(1 mm^2 スケール)かつ大量に取得し、それら生体画像データのビッグデータ化を進めている。

このような取り組みは、「組織微細構造情報の総体(モルフォーム、Morphome)」を通して、新たな生物学的な現象を理解・発見することを目指すものである。このような網羅的な組織科学的方法論は、生命科学分野において「モルフォミクス(Morphomics)」と言われ、新たなオミックス研究領域の1つとなりつつある。しかしながら、大量の画像データを取得したとしても、画像データだけではデータの運用が不十分であり、画像として撮像された被写体、撮像条件、表現型データ、及び画像解析結果などとの情報統合が不可欠である。また、遺伝子発現やメタボロームデータ等の生命科学データは RNA や代謝物等といった生体分子を定量したデータであり、空間的な情報が欠如していることがしばしば問題となっている。そこで、空間的情報を有する画像データを従来の生命科学データと総合して解析するかが課題になっている。この課題解決のためには、撮像された哺乳類等の個体や臓器・組織、その遺伝学的背景、試料作製法、撮像機器、及び設定パラメータ、観測されたものの表現型などのメタデータを統合して、総合的で大規模な高度知能型のデータ解析を実現することが望まれる。この結果として、生命の総合的な理解、さらには細胞レベルでの疾患、あるいは病気の前段階である健康脆弱化状態、及び未病状態の高精度な検出等の実現につながることが期待される。

しかしながら、この課題解決はほぼ未着手であるといってよい。例えば、撮像機器の性能向上や新たな技術・手法の確立による画像データのビックデータ化が急速に進んでいるものの、得られた画像データの共有、統合、及び解析を共通のプラットホームで可能な方法論は未だ構築されていない。また、生命科学系の画像データのみに限っても、種々の光学顕微鏡(実体顕微鏡、共焦点レーザー顕微鏡、超解像顕微鏡など)、透過型や走査型の電子顕微鏡、二光子顕微鏡や光ファイバー顕微鏡などの in vivo 光イメージング、磁気共鳴画像(MRI)、X 線コンピュータ断層撮影(CT)など、異なる測定系で撮像された画像データは、撮像方法を超えて、包括的に取り扱えていない現状がある。さらには、画像データには、実験条件等の相互運用可能なメタデータの記述を含めることが不可欠である。

本稿では、このようなメタデータ記述のために著者らが構築を進めている、「顕微鏡オントロジー(Microscopy Ontology)」と呼ぶオントロジーについて議論する。このオントロジーは、光学顕微鏡によって得られる多次元やヘテロな画像データの管理に特化し、同分野で標準規格になりつつある Open Microscopy Environment (OME)[Allan 2012]の XML(Extensible Markup Language)準拠の語彙に基づいている。著者らは、OME の語彙に電子顕微鏡撮像にも対応する装置や操作に関する語彙や、バイオ試料や実験条件などの実験データを説明する語彙を追加して、Web Ontology Language (OWL)準拠のオントロジーとして構築している。以前、そのプロトタイプとなるオントロジーを

連絡先: 久米 慧嗣, 理化学研究所 ライフサイエンス技術基盤研究センター, 理研 CLST-JEOL 連携センター, 神戸市中央区港島南町 6-7-3, 電話番号: 078-304-7160, Fax 番号: 078-304-7161, 電子メイルアドレス: satoshi.kume@riken.jp

報告し、画像データのメタデータを記述できることを示した[Kume 2016]。今回、SEM で得られる大量の画像データを、機械が理解可能な要素間の関係性を記述していく枠組みである Resource Description Framework (RDF)上で記述するために、さらなる画像データ語彙の拡張を試み、実際に RDF 基盤のメタデータベースでメタデータ間の繋がりを可視化することに成功した。本稿では、それらの結果を報告する。

2. OME の RDF 記述、及び画像ビッグデータ対応に向けた語彙拡張

筆者らは、以前、概念の階層関係や概念間の推論などの高度知識処理を行えるように、XML 準拠の OME データスキーマ(version: January 2015)を OWL/RDF に変換した[Kume 2016]。これは、OME データスキーマから概念やプロパティを抽出し、OWL/RDF 準拠のオントロジーとして再構成したプロトタイプである。

本稿では、OME が仕様化した概念のみならず、実験条件、撮像条件、バイオ試料などを記述できるように、新たなクラスやプロパティを追加した拡張版について述べる。この拡張は、生体組織の画像を撮像する際の実験条件(撮像条件や試料作製条件など)、バイオ試料などのメタデータを記述でき、かつ表現型データといった他の実験データとの統合を実現するために行ったものである。具体的には、顕微鏡オントロジーの上位概念を、「IMAGE(画像)」、「BIOSAMPLE(バイオ試料)」、「INSTRUMENT(装置)」、「SCREENING(実験処理)」、「EXPERIMENTER(実験者)」と定義したオントロジーを作成した(図1)。著者らの最新の顕微鏡オントロジーは OWL に準拠して記述され、324 クラスと 225 プロパティを備えている。

2.1. 実験メタデータ記述のための語彙拡張

画像データあるいはバイオ試料に対する表現型データ、及びバイオ試料に加え、顕微鏡の撮像条件、及びサンプル調整法などの実験条件などに関する実験メタデータを記述するために必要な新たな語彙を追加した。この拡張によって、実験毎で得られる画像データを撮像ごとや個体ごとに管理することができ、撮像条件のクラスによって、電子顕微鏡などの撮像条件を装置ごとに記述することができる。著者らのオントロジーの特徴として、画像データやバイオ試料は生物学的なアノテーションが異なる場合が多く、それを埋める工夫として、画像に映った細胞などの表現型データ、ある個体から取得されたバイオ試料が有する表現型データを分けて記述できるようにした。こうしたことによって、

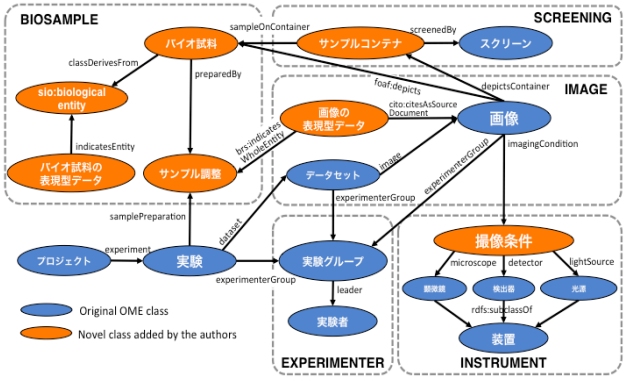

図1. 顕微鏡オントロジーの上位概念
オレンジのクラスは、今回著者らは提案した上位概念であり、青のクラスは OME に存在する上位概念である。

同じ疾患モデルのバイオ試料であっても、画像ごとに異なる生物的な形態やイベント(例えば、細胞の種類、細胞の機能や構造の変調など)を記述することができるようになった。

さらに、顕微鏡オントロジーの概念を検討・吟味することで、顕微鏡の記述語彙の共通化と各種顕微鏡の差異の記述を同時に実現するように、概念の上位概念を定義した。具体的な例として、顕微鏡の撮像条件では、光学顕微鏡と電子顕微鏡では装置の違いや設定パラメータが異なる部分があるが、それらはサブクラスとして個別化するものの、共通化できるものは上位概念である「ImagingCondition」として統合できた。また、撮像するもの(被写体)と映し出されたもの(画像データ)の表現型の記述を「PhenotypeData」として共通化することもできた。それらの結果、概念としての統合が、15 件(rdfs:subClassOf の上位概念)、選択肢カテゴリとしての統合が、27 件(rdf:type の上位概念)というように整理できた。このように、単にフラットなデータ構造である XML に準拠した OME データスキーマでは整理されていなかった、同種の概念を大きなくくりで扱うことでデータの再利用化が進んだ。

2.2. タイリング画像記述のための語彙拡張

網羅的に生体組織の微細構造を調べるモルフォミクス解析では、数千枚単位で連続撮像された画像データ、あるいはタイリング処理されて1枚当たり数 GB レベルの巨大な画像データを管理する必要がある。最近、このような巨大な画像データは、Deep Zoom Image(DZI)フォーマットという特殊な画像形式に変換して可視化が行われる。そこで、著者らは、それら大量の画像データ、タイリング画像データ、及び DZI フォーマットをメタデータで記述するのに、必要な新たな語彙(画像の重なり、タイリング方法、タイリングマップなど)を追加した。また、Deep Zoom Image Dataset のクラスは、データセットと rdfs:subClassOf 関係にあり、個別の画像データともリンクを張れる構造にした。

3. 生体組織の微細構造画像データを対象にしたメタデータベースの構築

実用的には、様々な実験結果やその条件と同時に、それらに対応する画像データを参照すること、または、画像データから検索して実験条件や生体組織の病態、あるいは表現型データなどを調査できることが重要である。そこで、著者らは、生体組織試料のサンプル調整から、SEM による画像撮像までの一連の

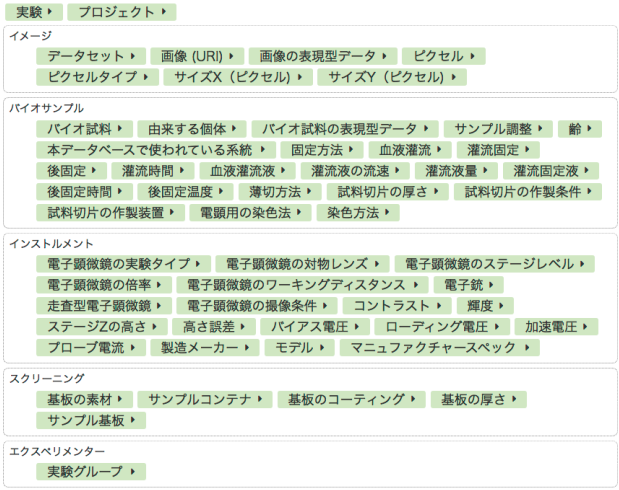

図 2.電子顕微鏡で取得した画像データを記述するために使用した顕微鏡オントロジーのクラス一覧

組織学実験を行い、これらの実験過程、及び画像データを開発したオントロジーを用いて RDF 準拠のメタデータを作成し、画像、実験条件、表現型データが繋がり、実用的なメタデータベースとして実装できることを検証した。実際に、画像データは RDF 準拠のメタデータ公開基盤である、理研メタデータベースにて可視化と検索機能を実現させた[Kobayashi 2016]。

著者らは、ラットから灌流固定後に生体試料(肝臓、及び腎臓)を取得して、固定、脱水、樹脂包埋した後に、各組織の超薄切片(約 70 nm)を作製した。さらに、電子染色を行った後に、SEM の反射電子検出法を用いて組織試料を観察し、3,000～5,000 枚の画像データを取得した。次に、このような実験条件やバイオ試料の情報を、顕微鏡オントロジーを用いて画像データのメタデータとして、実際の記述を試みた。その結果、画像データに対して、各種の実験条件、撮像条件、画像形式、バイオ試料の情報、表現型データなどのメタデータを付加できることが分かった(図2)。さらに、理研メタデータベース上でメタデータ連結の可視化を試みたところ、例えば、ラット肝臓の画像データでは、画像データの URI(左側)に対して、画像の説明、バイオ試料、被写体、サンプルコンテナ、電子顕微鏡の撮像条件などのメタデータを表示し、各メタデータの繋がりを検索することもできた(図3)。

以上のように、著者らは、生体組織の画像データを記述するための顕微鏡オントロジーの開発から始めて、このオントロジーに準じた実際の画像メタデータ作成、及び RDF 基盤でのデータベース構築と可視化というワークフローを構築した。

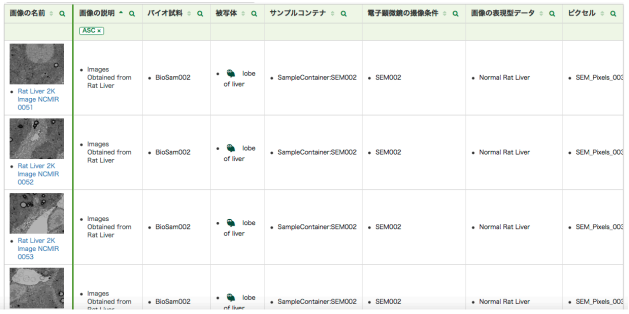

図 3. ラット肝臓を撮像した画像データに対するメタデータ記述の具体例

4. 今後の課題

これまでに述べた顕微鏡オントロジーの概念を基礎にして、モルフォーム解析基盤として発展させるための今後の課題について述べる。

4.1. 顕微鏡オントロジー概念の整理

顕微鏡オントロジーの概念をより詳細に検討・整理して、概念間の関係をより精密に記述する能力を向上させることが課題の1つである。具体的な検討項目は、以下の通りである。

(1) (選択肢として)共通語彙を整備すべき概念を明確化すること。

(2) 多様な目的に対しての当オントロジーの共通利用性を高めること。

(3) 当オントロジーが、多様なコミュニティで合意できるように洗練されたデータモデルとすること。

(4) 生命科学分野における他のスキーマ、及び同分野以外の画像メタデータスキーマ (Friend of a friend (FOAF)等)との相互運用性を検証すること 。

さらに、筆者らは、光学顕微鏡や電子顕微鏡以外の装置で得られる画像データ(MRI 画像など)も、同一のオントロジーによって RDF 化する試みを進めている。この取り組みを通じて、顕微鏡オントロジーの上位オントロジーとして、生命科学イメージングオントロジー(Life-science Imaging Ontology)が、生命科学系画像データの包括的なプラットホームとして提案できるものと期待される。

4.2. 表現型データとの連携など、幅広いユースケースを集める

顕微鏡オントロジーの利活用向上のためには、幅広いユースケースへの対応も課題である。この課題解決に向けて、以下の各項について検討を行う。

(1) 実験方法、画像形式、バイオ試料などのバリエーションを増やすこと。

(2) IMPC や J-Phenome といった哺乳類(マウスなど)等の表現型データとの連携を行うこと。

(3) OME コミュニティとの連携を行うこと。

特に、(2)の表現型データとの連携は、画像データの検索や解析には必須である。以下に、生命科学分野における表現型データを中心とする RDF データが、画像データと統合される意義について考察する。

表現型データの RDF 形式での公開はすでに進められている。国際マウス表現型解析コンソーシアム(International Mouse Phenotyping Consortium: IMPC)は、マウスを対象に、その表現型データを世界共通のプロトコールで網羅的に解析し、哺乳類ゲノムの機能を理解する国際的な枠組みである。IMPC では、各遺伝子のノックアウトマウスついて、血液検査、血圧、行動、形態など、幅広い表現型を解析しており、約 2500 のノックアウトマウスについてのデータがすでに公開されている [Dickinson 2016]。著者らは、これらの表現型データを RDF に変換して、RDF 基盤の理研メタデータベース上で IMPC_RDF (http://metadb.riken.jp/metadb/db/IMPC_RDF)として公開している[Kobayashi 2016]。また、日本国内の表現型データを種横断的に収集して公開する「J-phenome プロジェクト」においては、同メタデータベース基盤上で理研 BRC マウスリソース表現型メタデータ(http://metadb.riken.jp/metadb/db/rikenbrc_mouse)やナショナルバイオリソースプロジェクト(NBRP)ラット表現型メタデータ(http://metadb.riken.jp/metadb/db/NBRP_rat)などが公開されている。さらには、細胞の表現型データ(細胞組成、細胞プロセスなど)を記述するオントロジーとして、Cellular Microscopy Phenotype Ontology が提案されている [Jupp 2016]。このように表現型データが RDF データとして充実することで、生命科学データ間のハブとして機能しつつある。

著者らは、これらの表現型データを中心に、遺伝的な要因によって起こる疾患データ、遺伝子発現やメタボロームなど他のオミックス情報、及び生化学や動物行動学的な知見、文献情報など総合的な知識集約に取り組んでいるが、画像データもその1つである。画像データによって得られる生体組織の細胞や形態学的な情報は非常に重要な表現型データの1つであるが、画像データのみでは、その表現型情報を正確に解釈することは困難である。そこで、これら表現型データと画像データとを顕微鏡オントロジーによって統合的に記述することによって、生化学データなどを含む様々なオミックス情報や疾患情報を統合して扱うことができ、画像データの生物学的な解釈が飛躍的に向上するものと考えている。

5. まとめ

OME データスキーマ、及びその拡張によって開発された顕微鏡オントロジーは、バイオ試料の画像データの実験条件や表現型データなどを記述することが可能であった。今回の成果物として、生体組織の画像データ、及びそれらのメタデータを、RDF 基盤である理研メタデータベース（http://metadb.riken.jp）の「理研 CLST マルチモダル微細構造データベース (http://metadb.riken.jp/db/clstMultimodalMicrostruct)」上に公開した。今後、このようなデータベースは、モルフォミクス解析によって得られた画像データを公開・共有・検索する公共のプラットホームとして利用されることが期待される。また、多様な生命科学データに RDF 準拠のメタデータが付与され、その機械可読な情報の利活用が促進されれば、機械学習などの人工知能技術解析技術とメタデータの知識を総合した、より高度で多面的な生命科学データの解析が可能となることが期待される。

参考文献

[Allan 2012] Allan, C., Burel, JM., Moore, J., *et al.*: OMERO: flexible, model-driven data management for experimental biology. Nat. Methods., 9(3): 45–53, 2012.

[Dickinson 2016] Dickinson ME., Flenniken AM., Ji X., *et al.*: High-throughput discovery of novel developmental phenotypes. Nature. 537(7621):508-514, 2016.

[Jupp 2016] Jupp, S., Malone, J., Burdett, T., Heriche, JK., Williams, E., Ellenberg, J., Parkinson, H., Rustici, G.: The cellular microscopy phenotype ontology, J Biomed Semantics., 18;7:28, 2016.

[Kasthuri 2015] Kasthuri, N., Hayworth, KJ., Berger, DR. *et al.*: Saturated Reconstruction of a Volume of Neocortex, Cell, 162(3): 648–661, 2015.

[Kobayashi 2016] Kobayashi, N., Lenz, K., Masuya, H.: RIKEN MetaDatabase: a database platform as a microcosm of linked open data cloud in the life sciences. JIST 2016: Semantic Technology, 99-115, 2016.

[Kume 2016] Kume, S., Masuya, H., Kataoka, Y., Kobayashi, N.: Development of an Ontology for an Integrated Image Analysis Platform to enable Global Sharing of Microscopy Imaging Data, In Proceeding of 15th International Semantic Web Conference, 2016.