様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　１５　日現在

機関番号：　８２６５７
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０３４３

研究課題名（和文）文献全文からの網羅的な生物メタ情報抽出技術の開発

研究課題名（英文）Extracting Biologically Interesting Metadata from Full-Text Papers

研究代表者

山本　泰智（Yamamoto, Yasunori）

大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任准教授

研究者番号：５０４７００７６

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：生命科学の基盤となる遺伝情報が収められたゲノムに関する情報を、その機能に着目して整理したり俯瞰したりするためにセマンティックウェブ(SW)と呼ばれる技術が注目されている。しかし、このゲノム情報を解釈する上で必須の知見の多くが、いまだ学術文献という自然言語で記載された媒体に収められているままであり、SW技術を利用したゲノム情報解析がしにくい状況にある。本研究では学術論文全文から生物学的に重要なデータを抽出する技術を開発し、その結果をResource Description Framework (RDF)で構造化した。

研究成果の概要（英文）：Accompanied with the sheer increase of genome papers, we need an automatic acquisition system of biological knowledge from full papers. Moreover, several relevant datasets are scattered throughout multiple institutions, and biologically interesting analyses need to use them in an integrated manner. As for the microbe research, habitat environments and sampling locations are among the relevant data to be extracted. In addition, the Semantic Web technologies have been adopted by major biological institutions such as National Center for Biotechnology Information (NCBI) or European Bioinformatics Institute (EBI). In this situation, we built a manually annotated corpus and an automatic extraction system using text mining technologies. We use Resource Description Framework (RDF) to express the extracted knowledge, and are publishing it as Linked Open Data (LOD) to be efficiently and effectively used with other relevant datasets in an integrated manner.

研究分野：　テキスト処理

キーワード：　テキストマイニング　微生物ゲノム　セマンティックウェブ

１．研究開始当初の背景

1980 年代後半に始まったヒトゲノム計画を嚆矢として、多数の生物種に対するゲノムプロジェクトが産出するゲノム配列データ(ゲノム情報)の量は増加する一方であり、2000 年代から始まった相次ぐ新型シーケンサーの市場投入がその増加速度に拍車をかけている。得られたゲノム情報は、いわば遺伝情報の白地図であり、配列上のどこにどのような遺伝子が存在するかといった、構造化されたアノテーション情報を収める多くのデータベースが構築・公開されてきた。

一方で、ゲノム情報を解釈するためには生物学的特徴(以下、生物メタ情報と呼ぶ)の活用が重要であるが、これらは主として文献に記載されたままの状況にある。具体的には、どのような環境に生息しているのか、至適温度は何度なのか、どのような運動性を持つのか、必須とする栄養は何か、等である。Genomes OnLine Database (GOLD)は、生物メタ情報について、最も網羅的に収集しているデータベースであるが、それでも項目で平均すると、収録されている 13,786 生物種の 34.5%程度にしかメタ情報が記載されていない。

文献に記載されている生物メタ情報の多くは本文中に自然言語で記述され、また表現方法もバラバラで構造化されていないため、必要とする情報を効率的に取り出すことが困難である。しかも、生命科学分野では毎年約 70 万本もの文献が発表され続けており、その総数はすでに 2 千万件を超えているため、もはや人手では網羅的に情報抽出することができない。今後ますます増加するゲノム情報に対して大量の文献全文(文献のアブストラクトだけでなく、本文も含むもの)から生物メタ情報を網羅的かつ体系的に抽出して構造化する機械的な処理システムが必須である。

生命科学分野の研究成果を報告する学術文献は米国 National Library of Medicine（NLM）の提供する書誌情報データベース PubMed や、再利用のしやすいライセンスとファイル形式で文献の全文が提供されている PMC がある。そのため、これらを利用して情報を抽出することとする。

一方で、ライフサイエンス分野においては、これまで様々な組織により多くの有用なデータベースが構築、公開され、インターネットを経由して誰でも簡単に関連データにアクセスできるようになってきた。しかしながら、特定の研究目的に合わせて複数のデータベースから関連するデータを取得しようとすると、特にそれが相当数のデータ量が想定される場合、機械的に網羅的な取得が困難となる。そこで、近年、セマンティックウェブ技術を用いた機械的な統合処理のしやすい環境の構築が世界各地の研究機関で進められている。そこでは、すべてのデータを Resource Description Framework（RDF）というモデルで表現する。すなわち、データの ID はインターネット上の住所にあたる URI をもちいてグローバルに一意に指し示すようにし、それに対する様々な属性および属性値を、同じく URI や具体的な数値を用いて表現する。このようにすることで、データの識別子がグローバルに一意に表現できることから、分散的に構築されたデータベースでも、同じ概念には同じ識別子が用いられているので、容易にジョインでき、機械的に統合的なアクセスが可能になる。RDF を用いた再利用が容易でインターネットを介して機械的なアクセスが可能なデータベースは、Linked Open Data（LOD）と呼ばれる。ライフサイエンス統合データベースセンター(DBCLS)では平成 25 年度までのプロジェクトにおいて生物学研究に必要なオントロジーと、Linked Open Data (LOD)を構築してきた。海外でも UniProt や European Bioinformatics Institute (EBI)の RDF プラットホームなど、生命科学分野における RDF データの公開が進んでいる。

２．研究の目的

本研究では、構造化された生物メタ情報の充実を目的として、自然言語処理技術（NLP）や情報抽出技術により、文献全文を対象として生物種とオントロジーの統制語彙を対応付けたうえで、生物メタ情報を抽出する。

背景で記述した通り、抽出したメタ情報はより広く容易に機械的な処理がしやすいように、RDF データとして公開することとする。

また、情報抽出技術を開発するにあたり、領域の専門家によるアノテーション作業が必要になる。これは、予め特定量の文献全文に対して、機械的に抽出すべき表現に印をつけておくもので、この情報を基に機械的な抽出ができるように学習させたり、ルールセットを構築したりする。

３．研究の方法

最初に、表現型オントロジーに対する生物メタ情報抽出システムのプロトタイプを開発し、実現可能性を確認する。続いて、培地、環境、病原性など他の利用可能なオントロジーについて順次アノテーション作業を行い、それぞれに適した抽出技術を検討したうえで、情報抽出システムを開発する。情報抽出技術としては、広く有用性が報告されている Conditional Random Field（CRF）を利用することを想定するが、適宜パターンマッチに基づくルールベースの手法も検討する。生物メタ情報の抽出システムの概略を図 1 に示す。

この中で、PMC OA Subset とは、PMC に収められている文献全文のセットのうち、再利用のしやすいライセンス、例えば Creative Commons Attribution（CC BY）および機械的な処理のしやすい XML 形式で提供され、

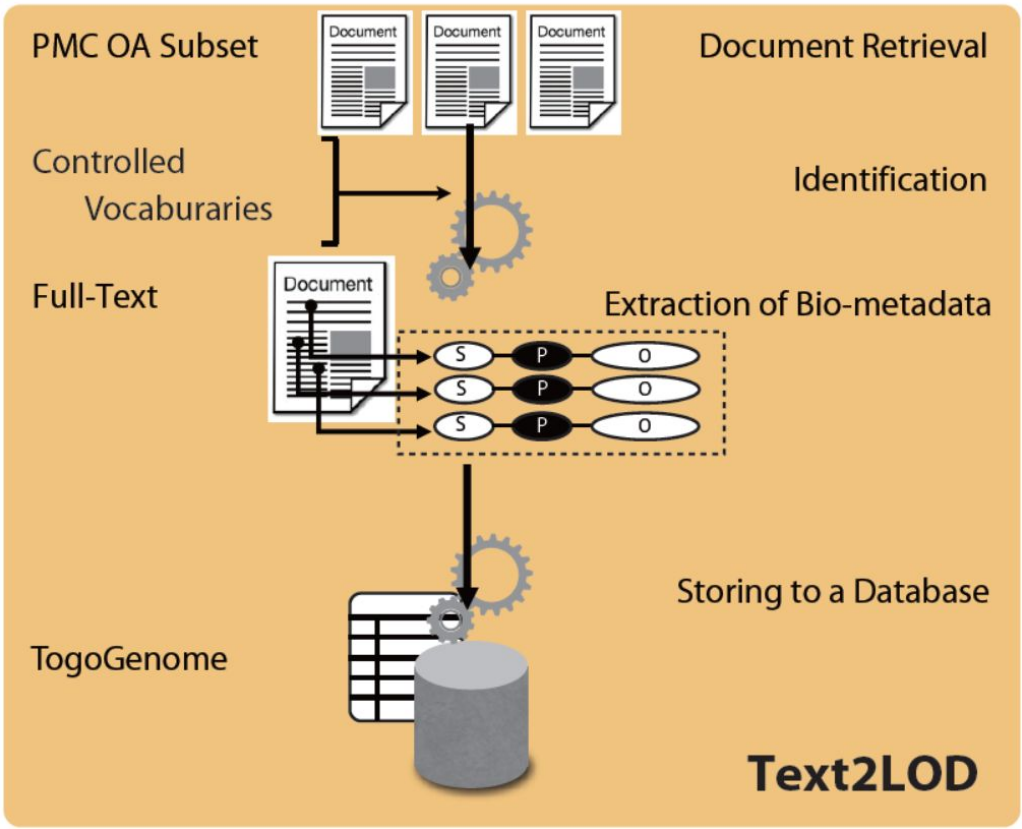

図 1: 文献からの生物メタデータ抽出システム概略図

さらに、一括ダウンロードが可能とされているものである。これらを対象として、領域の専門用語が収められている辞書やオントロジーである統制語彙（Controlled Vocabulary）を用いて構築した情報抽出技術を適用し、必要な生物メタ情報を取得する。アノテーション作業は、出現する記述パターンの収集を行い、ルールセットやコーパスを構築することが含まれる。これらの作業については領域の専門家に依頼し、PMC から取得した微生物のゲノムに関する文献全文に対して必要なアノテーション作業を実施する。
また、オントロジーごとに適切な生物メタ情報を定め、開発された情報抽出技術を用いて文献全文からの抽出、RDF での蓄積を行う。得られた生物メタ情報は LOD として一般公開し、外部アプリケーションからの利用を可能にするほか、構築した言語資源についても公開する。

４．研究成果

まず、微生物の生育環境や単利場所、生育至適温度、細胞サイズなど、ゲノム情報を解釈するために重要な各事項に対するデータ、すなわち、生物メタ情報を抽出するシステムを開発した。
機械的な抽出を行うためには、あらかじめ領域の専門家により、抽出すべきデータを実際の学術論文中に特定する作業（アノテーション）が必要となる。このため、論文全文を機械的な処理がしやすいライセンスと形式で公開されている論文データベース PMC から 1000 件程度取得して作業した。これに対して生物メタ情報を抽出するプログラムを開発し、結果を評価した。
その結果、生育環境と単利場所については、固有表現抽出（Named Entity Recognition）タスクで良い抽出性能を持つことが知られている機械学習手法、Conditional Random Fields（CRF）を用いると良い結果に繋がることが判明した。また、生育至適温度及び細胞サイズについては、正規表現を用いたパターンマッチと、それで得られた結果に対する文脈判定フィルタを用いると良い結果が得られた。
得られた生物メタ情報は RDF 形式で、再利用しやすいように一般公開するため、その表現方法を規定するオントロジーを構築した。本オントロジーを用いたデータ表現例を図 2 に示す。

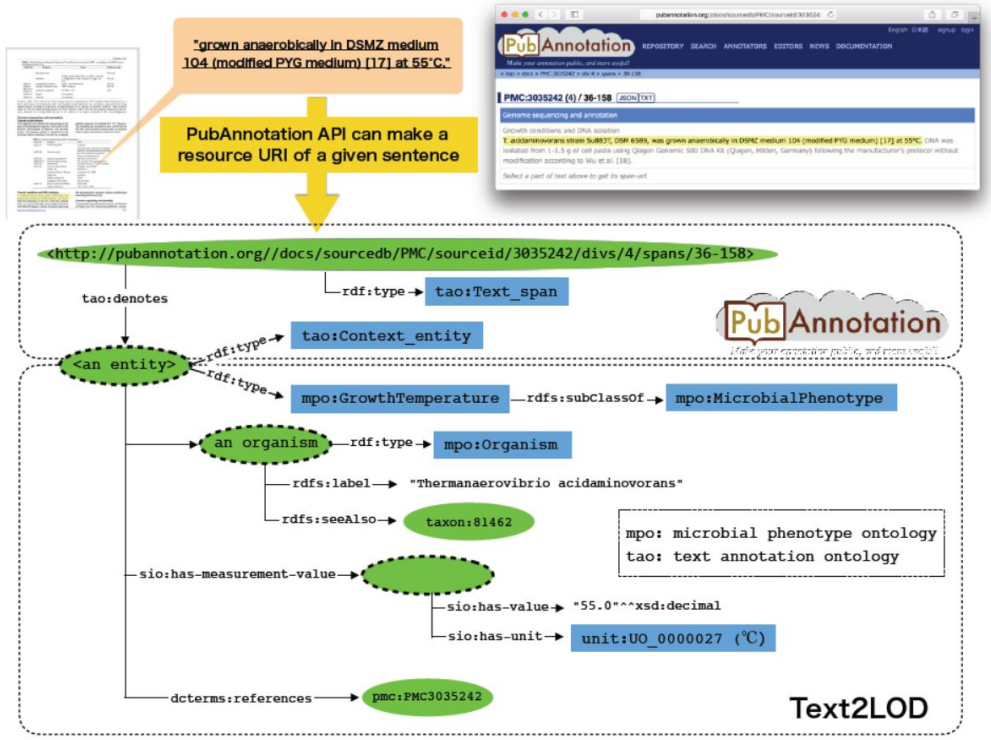

図 2: 提案した生物メタ情報の抽出結果を表現するオントロジー利用例

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　　件）

〔学会発表〕（計 2 件）
Yasunori Yamamoto, Shinobu Okamoto, Shuichi Kawashima, Toshiaki Katayama, Yuka Nakahira-Yanaka, Hiroko Maita and Sumiko Yamamoto. Text2LOD: building high-quality linked open annotation data concerning biological interests. Biocuration 2016. (2016)

Yasunori Yamamoto, Shinobu Okamoto, Shuichi Kawashima, Toshiaki Katayama. Towards Making Knowledge in Literature LOD. Biocuration 2018. (2018)

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://togogenome.org/

６．研究組織
(1)研究代表者
山本 泰智　大学共同利用機関法人情報・システム研究機構（機構本部施設等），データサイエンス共同利用基盤施設，特任准教授

研究者番号：50470076

(2)研究分担者
川島 秀一　大学共同利用機関法人情報・システム研究機構（機構本部施設等），データサイエンス共同利用基盤施設，特任助教

研究者番号： 50314274

片山 俊明　大学共同利用機関法人情報・システム研究機構（機構本部施設等），データサイエンス共同利用基盤施設，特任助教

研究者番号： 60396869

岡本 忍　大学共同利用機関法人情報・システム研究機構（機構本部施設等），データサイエンス共同利用基盤施設，特任准教授

研究者番号： 90623893