Vol.2011-DD-82 No.5
2011/10/8

医薬品情報の Linked Data 化と活用への取り組み

細見 格[†]

本稿では、「データの Web」を具現化した Linked Data の構築と活用に対して積極的に取り組んでいる医薬品業界の動向を概説する。同業界では、後発医薬品の普及による収益低下を回避するため、また、難病に対して有効な治療薬を早期に開発するため、医療・製薬に関する様々な企業や大学、研究機関が幅広く情報を共有し協力する動きが活発になっている。その例として、W3C による Linking Open Drug Data、ontotext 社による Linked Life Data などを紹介し、医薬品業界の課題にどのように応えているかを述べる。

An Impact of Linked Data to Pharmaceutical Industry

Itaru Hosomi[†]

In this paper, we give an overview of some activities in pharmaceutical industry which tackle effective use of Linked Data. In the industry, many companies and universities share their information and work together to maintain sales of drugs and to develop efficient drugs for intractable diseases. We present Linked Open Drug Data, Linked Life Data and other some projects to contribute to solution of issues of the industry.

1. はじめに

従来の World Wide Web が「文書の Web」（Web of Documents）であったことに対し、「データの Web」（Web of Data）を具現化した Linked Data が注目されてきている。従来の Web は、基本的に URL で特定され HTTP プロトコルでアクセスされる HTML 文書の集合だが、これと同様に、Linked Data は URI で特定され HTTP でアクセスできる RDF データの集合である。また、従来の Web 文書間のリンクは殆どが HTML の A リンクのみであったが、Linked Data では RDF のプロパティを用いた任意のタイプ付きリンクでデータ間を結びつけることができる。世界中のどこからでも参照可能な公開された Linked Data は、特に Linked Open Data（LOD）と呼ばれている。

文書の Web は、人間が直接読むことを前提とした情報のグローバルな公開と共有を劇的に容易にし、急速に発展してきた。その結果、あまりに多くの情報が公開され、人間が読んだりリンクを辿って関連する情報を探していくだけでは、その広大な情報空間のポテンシャルを十分活用できなくなっている。データの Web は、情報を計算機が容易に理解し操作できる形で公開することにより、膨大な情報を人間より遙かに幅広く且つ素早く活用できることを狙いとしたコンセプトである。その標準となりつつある Linked Data は、今や行政や出版・メディア、ライフサイエンス、地理など様々な分野の情報公開・共有に活用されている[1]。

本稿では、その 1 つであるライフサイエンス分野の情報について、特に医薬品関連情報の Linked Open Data（医薬品情報 LOD）とその動向について述べる。医薬品情報 LOD については[2]でも紹介しているが、ここではさらに[2]で取り上げていなかった情報や考察を加える。

本稿は以下の構成で進める。2 章では医薬品業界の課題を述べ、3 章では医薬品関連情報の LOD 化と関連ツールによる LOD の活用例を紹介する。4 章では医薬品関連企業による取り組みや今後の展開について述べ、5 章でまとめる。

2. 医薬品業界の課題

2.1 2010 年問題

医薬品業界には、2010 年問題と呼ばれるものがある。これは、製薬企業大手の主力製品に関する特許が 2010 年前後に相次いで期限切れとなり、低価格な後発医薬品（ジェネリック）による置き換えが脅威となる問題である。全く新しい医薬品の開発には百億円単位の投資が必要だが、その結果取得した特許は時に年間 1000 億円以上の売上を独占できるという価値をもたらす[3]。世界最大手の製薬企業である米ファイザーは、有名なバイアグラで年間 1000 億円以上、高脂血症治療薬リピトールでは年間 1 兆円以上の売上を誇るが、両製品の特許が 2010 年から 2012 年までに世界の主要市場国で次々と期限切れを迎えるため、最近はこれらの製品の売上減が続いている。

新薬の開発には、莫大な投資に加えて各国政府の認可を得るまでに長い年数がかかるという問題もある。しかし、さらに深刻な問題は、有効な分子化合物を発見し易くポピュラーな（＝市場が大きな）病気に対してはジェネリックが既に普及していて、多少効果が高くとも高価格な新薬では開発投資が回収できず、有望な対象となる病気

[†] NEC 情報・メディアプロセッシング研究所
Information and Media Processing Laboratories, NEC Corporation

は癌やアルツハイマー病のような難病ばかりになっていることである。病気は個人によって症状の程度や発症部位、薬の効き方などが異なるため、非常に多くの臨床試験を経てようやく効果を確認できる。人体の細胞の異常である癌や、動物実験では症状や薬の効果が確認し難いアルツハイマー病などに対し、1 社が集めうる情報や分析力、臨床試験の数では、多くの患者に有効な新薬を開発することは非常に困難である。

2.2 医薬関連情報の公開と共有

病気の原因解明や困難な新薬開発を促進するため、医薬品に直接関係するタンパク質や遺伝子の情報に加え、臨床試験や患者特性などに関する様々な情報が、既に数多く公開されている。例えば、ClinicalTrials.gov[1]では臨床試験データが、DrugBank[2]やDailyMed[3]では薬に関する情報が、The human disease network[4]では病気や遺伝子に関する情報が公開されており、定期的に更新されている。しかし、これらは各々異なる組織が独自の表現形式で公開しているため、互いに関連するデータ同士のリンクがなく、同じものでも表現が異なるために機械的な対応付けも容易に行なえなかった(図 2-1)。

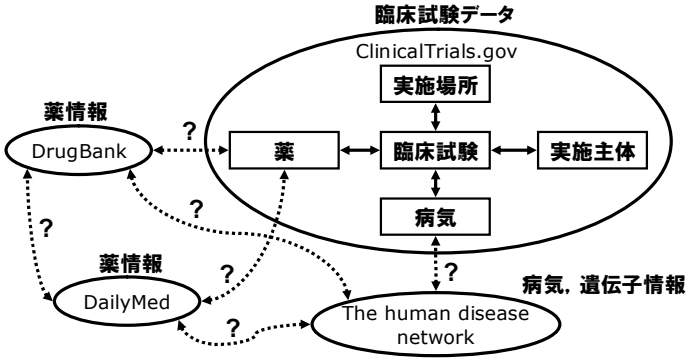

図 2-1 相互に独立している各種の公開情報

3. 医薬品情報の Linked Data 化とツール

3.1 Linked Data の基本

1 節で述べたように、Linked Data は W3C 標準の RDF 形式で記述されたデータの集合である。RDF は、主語・述語・目的語（またはリソース・プロパティ・オブジェクト）の 3 つ組を 1 つのステートメント（またはトリプル）と呼ぶ単位としてデータを記述する[4]。主語（リソース）は、ステートメントを一意に識別する URI で表される。例えば、アルツハイマー病は病気の一種であること、すなわち “Alzheimer’s disease is a disease.” を RDF で書く場合、“<http://～/alzheimer_disease> a <http://～/disease> .”のようになる（Notation3 という RDF 構文を用いた場合。 “a” は “is-a”（～の一種）を表すプロパティ表記）。

3.2 Linking Open Drug Data

RDF や XML、HTTP などの Web 標準を策定している W3C では、2005 年 9 月にヘルスケアやライフサイエンスに関する活動グループ “Semantic Web Health Care and Life Sciences Interest Group” が発足し、その下で医薬品関係の LOD を構築するためのタスクフォース “Linking Open Drug Data”（LODD）が 2008 年 10 月に設立された[5]。LODD は、様々な形で公開されている医薬品関連情報を RDF 形式に変換し、相互に対応するリソースをリンクすることで、「LODD クラウド」を構築している（図 3-1）。

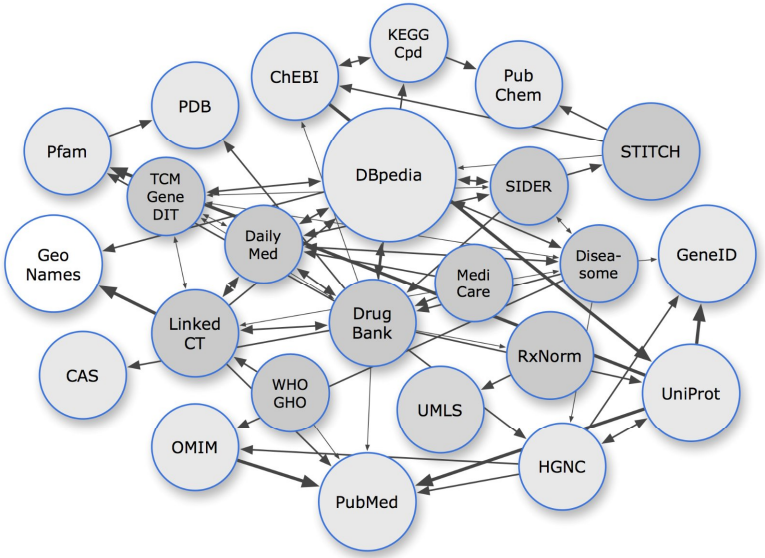

図 3-1 2010 年 12 月 4 日時点の LODD クラウド[6]

1 http://clinicaltrials.gov/
2 http://www.drugbank.ca/
3 http://dailymed.nlm.nih.gov/
4 http://www.nd.edu/~aib/Publication06/145-HumanDisease_PNAS-14My07-Proc/Suppl/
5 http://www.w3.org/wiki/HCLSIG/LODD
6 http://www.w3.org/wiki/File:2010-12-04_lodd_cloud.png

例えば、前述した The human disease network が公開している病気に関する情報は、LODD によって Diseasome という Linked Data に変換され、SPARQL エンドポイント（RDF 検索言語 SPARQL でアクセス可能なサーバ）が公開されている。LODD によって、従来独立していた情報同士が相互に関連づけられ、過去の膨大な知の蓄積を効率的に活用できるようになってきている（図 3-2）。

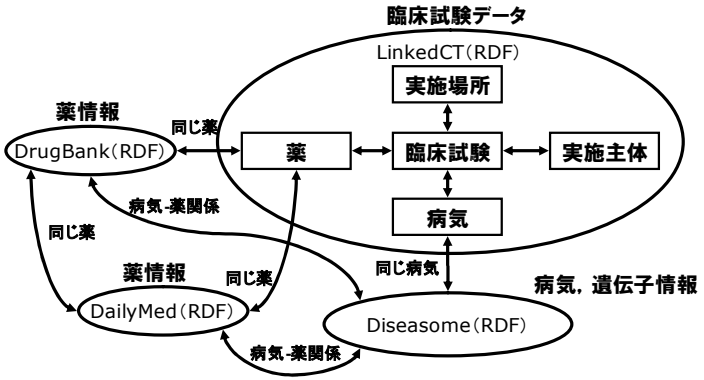

図 3-2 LODD によって関連づけられた公開情報

LODD で共有されている RDF データは既に 1600 万トリプル以上に達しており、その有効活用に幾つもの大学や企業が取り組んでいる。LODD および LODD クラウド上の各データセットについての詳細は[2]を参照されたい。

3.3 Linked Life Data

Linked Life Data（LDD）は、2009 年頃から ontotext 社が構築しているライフサイエンス情報の統合・検索プラットフォームである[7]。W3C の LODD と同様に、医薬品業界やバイオ産業界が直面している多種多様な情報の統合利用に関する問題を解決するため、各種のデータを RDF 化し横断的に検索可能にしている。2011 年 9 月現在、27 種類の情報源から 50 万トリプル以上の RDF データを蓄積し、30 種類以上（LDD v.0.8 では 33 種類）の RDF データが自由にダウンロード可能になっている。LODD に組み込まれている Diseasome は LDD にもあり、その RDF データは、3.1 節で紹介した例に近い形式で病気の名称や分類、関連する医薬品などについて 7 万トリプル以上が登録されている。

こうした医薬品関連情報の Linked Data 化については、他にも Bio2RDF や Chem2Bio2RDF といったプロジェクトがあり、世界的なムーブメントとなっている[5]。

3.4 TripleMap

医薬品情報 LOD に関する最近の研究開発動向を概観した文献[5]では、LODD で公開されているデータをビジュアルに参照し分析できるツールとして、TripleMap を紹介している[8]。LODD 用の TripleMap は、図 3-3 のように病気や患者や薬などを表すアイコンが用意されており、リンクのプロパティによって相互関係も容易に確認できる。

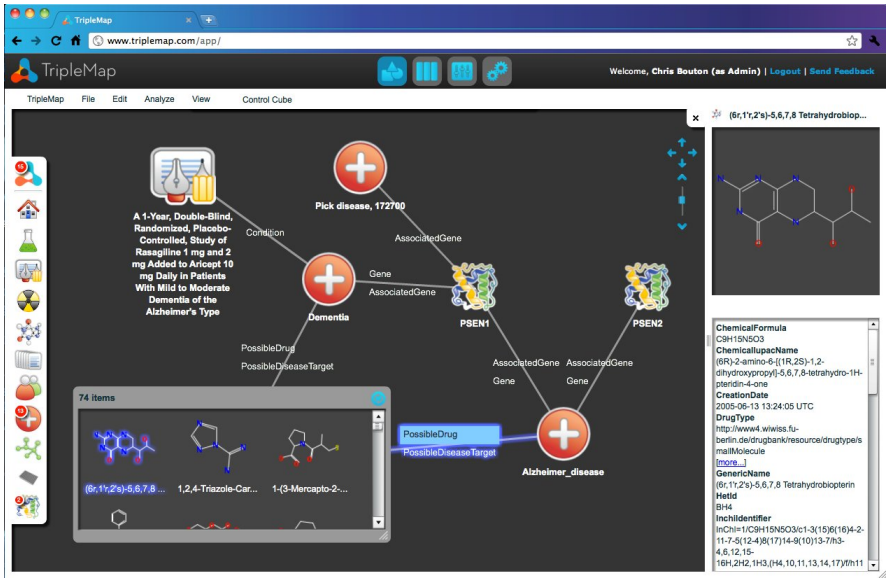

図 3-3 TripleMap の画面例（[5]から引用）

[5]にある LODD と TripleMap の利用例を紹介する。ある研究者が、LODD からアルツハイマー病に関する情報を集めたいと考え、TripleMap で “Alzheimer’s” とキーワードを入力して検索を行なう。TripleMap は、自動補完機能で入力文字列を含む LODD 上のキーワードをリストアップし、研究者はそこから “Alzheimer’s Disease” を選択して TripleMap のワークスペース（図 3-3 の主領域）にドラッグ＆ドロップする。すると、画面右下のプロパティパネルに表示された様々な情報源と関係に基づく Linked

7 http://linkedlifedata.com/

8 http://triplemap.com/

Data がワークスペース内で視覚的に表示される（図 3-3 のワークスペース右下の＋印の丸いノードが Alzheimer's_disease アイコン）。研究者が Alzheimer's_disease アイコンを選択すると、システムは、LODD 上の Diseasome からその病気に関連する遺伝子を、DrugBank と DailyMed からは関連する化合物を、LinkedCT からは関連する臨床試験データを自動的に表示する。すると、研究者は表示された遺伝子間のリンクや臨床試験データ間のリンクを参照し、自分が知らなかった（他の誰かによって発見された）関係を見つけられる可能性がある。このように、誰かによる発見を世界中の研究者などが容易に素早く共有できるようになることが、LODD や TripleMap のようなツールの主要な価値だと言える。

4. 業界による取り組みと今後の展開

LODD のような取り組みには医薬品業界の企業も参加しており、米国の Eli Lilly や Johnson & Johnson などが代表的である。また、Eli Lilly とファイザー、独メルクの 3 社は、アジアに患者が多い肺癌と胃癌の研究をより効果的に進めるため、2010 年 2 月に Asian Cancer Research Group を共同設立した[2]。このようにオープンではないが企業の枠を超えた情報共有も、2 節で述べたような医薬品開発の困難さを克服する手段として行なわれている。より国家的な取り組みとしては、欧州の Khresmoi プロジェクトがある[9]。これは、2010 年に開始された 4 年間のプロジェクトであり、前述の ontotext 社を含む 9 ヶ国 12 組織が多種多様な生体医学情報を統合・分析し、医学研究者だけでなく非専門家に対しても幅広く高信頼の情報を提供することを目指している。

しかしながら、以上のような数々の努力によって医薬品関連情報の共有や統合利用については大きな前進を見たものの、まだその顕著な成果は得られていないようである。[2]で述べたように、副作用の早期発見や創薬における効果を期待するが、医薬品の開発や効用・影響の確認には非常に多くの症例と長い時間が掛かるため、その中で LODD のような取り組みがいつどのような形で貢献するかを見定めるには、まだしばらく様子を見なければならない。

一方、日本の状況について、[2]では欧米に比べて医薬品関連情報の公開・共有が遅れていると述べたが、理化学研究所は、同所が保有する大規模な生命情報統合データベースをクラウド上の仮想ラボで内外の研究者が自由に利用できるようにしており[10]、さらに最近ではそのデータベースを LOD として公開する準備を進めている[6]。国内でこのような大きな進展が見られたことは、今後の励みになるだろう。

9 http://www.khresmoi.eu/
10 http://database.riken.jp/

5. おわりに

本稿では、医薬品業界が抱える問題と医薬品関連情報を Linked Data 化して共有する米欧の取り組みについて述べた。特に、[2]で触れていなかった欧州の Linked Life Data や、LODD を活用できるツールとして TripleMap を紹介した。癌の治療薬開発のように、人類や産業界にとって共通の大きな課題に対し、世界中の組織が協力して取り組むための基盤やツールが求められている。Linked Data は、従来の Web からさらに一歩先へ進むための、シンプルだが有効な一手になりうる。

World Wide Web は、文書の Web からデータの Web へと踏み出そうとしている。文書の Web はこれからも拡大し続けるだろうが、最近のトラフィックは Web のページ単位よりも RSS や Twitter のツイートのような小さな単位のテキストデータが（または映像ストリームが）多くを占めている。データの Web を構成する RDF のような最小限の情報表現も文書として扱うべきだろうか。Wikipedia 日本語版では、文書を「参照されることを前提として記録される情報」と説明し、さらに「コンピュータによって操作される情報も文書の一つである」と述べている。このような解釈によれば、1 つの RDF データ（主語・述語・目的語からなる 1 文で個別に参照可能なデータ）も文書である。ツイートではさらに短く「眠い」のひと言のみといった文書も発信されているが、情報空間で対象を一意に特定でき、且つ事実を明示する最小限の文書としての RDF と、その文書同士を様々な関係で結びつけた Linked Data は、グローバルな情報共有を追及する Web が辿り着く 1 つの究極の姿かも知れない。

多くのステップを踏むストーリーや複雑な構造を持つシステムについて語るには、構造化され適切にレイアウトされた文書が適している。しかし、個々の事実とその間の発見的な関係が重要となり、しかも同時に膨大な量を相手にする必要がある医薬品関連情報などには、Linked Data は必然的な形とも考えられる。データの Web が文書の Web と同様に「デジタルドキュメント」としての研究対象であるかについても、あらためて問うていきたい。

参考文献

[1] Bizer, C., et.al., 萩野(訳), "Linked Data の仕組み", 情報処理, Vol.52, No.3, pp.284-292, 2011.
[2] 細見, 長野, 岡部, "次世代の医薬品開発を支える知識流通", 情報処理, Vol.52, No.3, pp.300-308, 2011.
[3] 佐藤, "医薬品クライシス 78 兆円市場の激震", 新潮社, 2010.
[4] W3C, "RDF Current Status", http://www.w3.org/standards/techs/rdf
[5] Samwald, M., et.al., "Linked open drug data for pharmaceutical research and development", Journal of Cheminformatics, 3:19, 2011.
[6]David, G., "BASE がつなぐ、生命情報と社会", 横浜研究所みんなのコラム, 理化学研究所, http://www.yokohama.riken.jp/column/110817.html, 2011/8/17.