特許文書から抽出した化学物質情報の知識化
―オープンデータとの連携―

Knowledge Based Information Processing in Chemicals from Patent Documents

株式会社富士通研究所 **池田　紀子**

R＆Dマネジメント本部　企画部、技術士（応用理学／総合技術監理部門）、電子・光学デバイス材料の設計および分析、並列処理および分子モデリングの研究、特許読解支援システムの開発に従事。

株式会社富士通研究所 **田中　一成**

知識情報処理研究所　ナレッジシステム PJ、テキストマイニング技術の研究、特許読解支援システムの開発、LOD の研究に従事。

1 はじめに

化学分野の特許文書には、化学物質に関する多様・膨大な情報が蓄積されている。しかし、特許文書には独特の読み難さがある。そして、化学分野の特許文書には、他分野とは異なる請求項の書き方や、化学物質名の膨大な羅列がある。これらのことが原因で、人が化学分野の特許文書を読んで知識を得るには、予想以上に高度なスキルと多大な労力が必要である。そこで、化学分野の特許読解を支援する目的で、化学物質名と化学式の対応関係を抽出し、可視化する手法について考案して、実証実験を行っている[1]。さらに、化学分野に特化した特許読解支援システムの応用を通して、様々な情報を統合し利活用する可能性が広がると考えている。本稿では、化学物質情報の知識化について、特許読解支援システムの実用化とオープンデータとの連携事例について述べる。

2 特許読解支援システムの実用化

化学分野の特許文書を読み解くには、化学物質名とその化学構造を認識することが重要である。そのための化学物質情報ツールを開発し、特許読解支援システム上で実用化した。

2.1 化学物質情報ツール

我々が開発した化学物質情報ツールは、以下のような特徴があり、テキストの化学物質名から化学構造理解を支援できる[1]。

①特許文書をコーパスとして用いることで、有機化合物（炭素原子の骨格構造が基本）の化学物質名と化学式の対応関係を抽出する。

②化学物質の命名規則を元に作成した部品化ルールによって、化学物質名と化学式の対応関係を増やす。

③部品のデータベースを利用することで化学物質名を意味ある単位に分割する。

④化学物質の部分構造の配置関係を表構造で可視化する。

2.2 特許読解支援システム

当社では、テキストマイニング技術を応用した特許検索システムを構築している。その特許文書表示機能として特許読解支援システムを位置付け、図 1 に示すように、検索から読解までをシームレスに連携したシステムを構築している[2]。特許読解支援システムのオプション機能として、2.1 の化学物質情報ツールを実装した。本ツールを用いて、化学物質に関する特許出願の権利範囲の把握に役立てている。

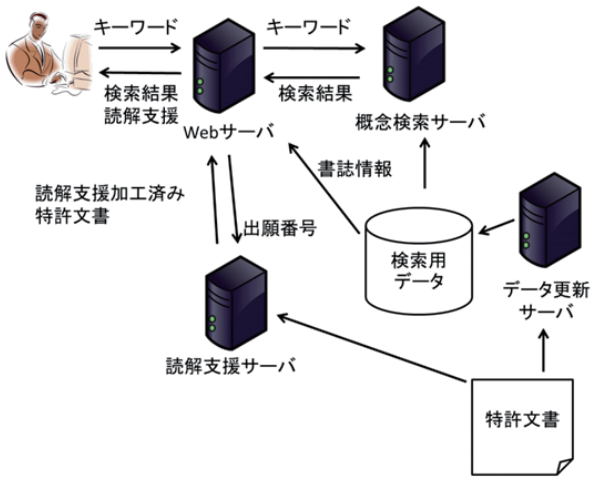


図１　特許読解支援システムの構成図

3 オープンデータとの連携

人工知能関連技術を仲介として、特許情報と様々なオープンデータをつなげることにより、単独のデータからは得られない有用な知識が得られるものと期待されている。ここで、「オープンデータ」と言えるためには、①機械判読に適したデータ形式で、②二次利用が可能な利用ルールで公開されたデータである必要がある[3]。日本政府のオープンデータセット数は 17,000 を超え[4]、オープンデータ化を進める地方自治体数も増加中で、日本にも本格的なオープンデータの時代がやってきた。こうした状況の中、LOD（Linked Open Data）は機械処理が可能なデータによる Web の実現を目指して誕生し、急速に普及発展しようとしている。データを相補的に使うことで、より多くの化学物質情報を入手でき、LOD が、特許からの化学情報（製造方法、化学反応、パラメータ、用途など）抽出のブレイクスルーを生み出す可能性があると考えている。

3.1　LOD4ALL

富士通研究所では、LOD の活用基盤である LOD4ALL を公開中である[5]。世界中で公開されている LOD を収集して一括検索することを可能にする。利用可能な Linked Data の横断的なブラウジング、検索、アクセスを提供している。また、LOD を利用したアプリケーションの開発プラットフォームを提供することも目指しており、今後は開発プラットフォームとしての機能を拡充していく予定である。LOD4ALL は利用可能なデータセットのカタログを提供することではなく、オープンデータ利用の容易化･促進に焦点を置いている。LOD4ALL は富士通研究所、Fujitsu Laboratories of Europe、および、アイルランド国立大学ゴールウェイ校の研究機関 INSIGHT Centre for Data Analytics（旧名称 DERI）の共同プロジェクトである KI2NA プロジェクトの成果である。

3.2　化学物質データベース

主な化学物質データベースについて、2016 年 8 月現在の収録化学物質数、公開年、ライセンスを表 1 に示す。

表１　主な化学物質データベース

データベース	化学物質数	公開年	ライセンス
日化辞	3,490,000	2005	無料(CC BY 2.1 JP)
ChEMBL	1,900,000	1994	無料(CC BY-SA 3.0)
CAS	118,330,000	1907	有料
PubChem	91,670,000	2004	無料
ChemSpider	57,000,000	2007	無料

3.2.1　日化辞

日本化学物質辞書（日化辞）は、科学技術振興機構 JST（The Japan Science and Technology Agency）が管理運営する国内最大級の有機化合物データベースである[6]。日化辞では化合物（2 種以上の元素原子から成る純物質）の名称（日本語および英語表記）、分子式、分子量、法規制番号、化学構造（MOLfile 形式）、用途語（JST 科学技術用語シソーラス由来）などを収録している。現在、J-GLOBAL にて提供が行われている[7]。セマンティック Web に適合した記述モデルである RDF（Resource Description Framework）で Linked Data として記述し、RDF の問い合わせ言語である SPARQL により検索することができる[8,9]。これらの情報は、CC BY のもと 2015 年 5 月に一般公開された[10]。RDF データを利用することで、EBI（The European Bioinformatics Institute）[11] が提供する UniChem[12] に収録されているデータベースを含め 30 種類以上のデータベースへのリンク情報を一括して取得することができる。UniChem は、化学構造識別子間のポインタのデータベースで、単純、大規模、非冗長である。

3. 2. 2　ChEMBL

日化辞からは、医薬品及び医薬品候補化合物などの生

物活性低分子のデータベース ChEMBL[13] と連携でき、目的とする化学物質の情報が一括して効率的に収集することができる。ChEMBL では化合物プロパティ（化学物質の性質の指標、例：分子量、分配係数等）の計算結果を蓄積している。化合物構造は半無限的に考えられるが、その性質を評価するために指標を実験的に求めていくプロセスは時間とコストがかかる。この化合物プロパティを LOD 連携に加えると、スループットをあげることができる。また、世界の特許情報から抽出した 1,500 万件以上の化学構造式のデータベース SureChEMBL も無料公開中である。

3.2.3 CAS

米国化学会 ACS（American Chemical Society）の下部組織である CAS（Chemical Abstracts Service）が、有料データベースを運営している[14]。現在、CAS は、化学研究に不可欠な巨大データベースとして世界中で利用されている。ACS が発行する Chemical Abstracts 誌で使用される化合物番号（CAS registry number）は、化学物質を特定するための番号である。

3.2.4 PubChem

PubChem は米国国立医学図書館 NLM（National Library of Medicine）が提供するデータベースで米国国立衛生研究所 NIH（National Institutes of Health）ロードマップに基づく化合物（1000 原子以下、1000 結合以下の低分子）とその生体活性を網羅的に収集している[15]。アッセイ（実験）の結果を収録した PubChem BioAssay、化学物質の情報を収録した PubChem Compound および PubChem Substance の 3 つから構成されている。

3.2.5 ChemSpider

ChemSpider は、英国王立化学会が提供する、クラウドソーシング的アプローチの化学データベースである[16]。ユーザが化学構造、スペクトル、情報整理等を提供することができる。文献やウェブページから化合物名を同定して抽出し、化合物名を化学構造に変換するアルゴリズムを用いている。その結果、ChemSpider は 150 以上のデータリソースの統合システムとなっている。このデータリソースは、文献や化学情報である。

表 2　化学物質の機能・用途と代替物質

ポリエチレングリコール		
機能・用途	出現回数	代替物質の候補
可塑剤	687	ジブチルフタレート
		ジオクチルフタレート
界面活性剤	541	ポリビニルアルコール
結合剤	192	ヒドロキシプロピルセルロース
		ポリビニルアルコール
		カルボキシメチルセルロース

表 3　化学物質基本情報

3.3 知識化

特許文書の化学物質名を認識し、オープンデータと連携することで、化学物質情報の知識化を進めている。表 2 に、化学物質「ポリエチレングリコール」について、特許文書から抽出した機能・用途、機能・用途ごとの出現頻度、代替物質の候補を示す。開発中の Web API の利用により、取得した化学物質データを加工して表示したり、新たなアプリケーションを作成したりすることが可能である。関係ありそうなデータ群を仮説検証的に解析することから、想定内の期待を裏切らずに、いろいろなことが発見されるであろうと考えている。そこで、表 2 中の可塑剤、ジブチルフタレートやジオクチルフタレートについて化学物質データベースと連携して化学物質情報を抽出した。表 3 は、日化辞との連携を表示した例である。表 3a は、構造を、表 3b は、別称を抽出した例である。表 4 は、ChEMBL と連携して化合物プロパティを、表 5 は CPCat（Chemical and Product Categories）[17] と連携して製品情報を抽出した例である。CPCat は、米国環境保護庁、EPA (United States Environmental Protection Agency) が提供する、43,000 以上の化合物の用途や機能を分類した化学物質と製品のデータベースである。

表 3a　化学物質の構造式と分子式

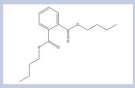
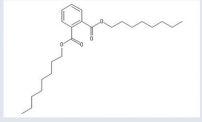
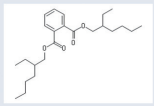
代替物質の候補	構造式	分子式
ジブチルフタレート		$C_{16}H_{22}O_4$
ジオクチルフタレート	 	$C_{24}H_{38}O_4$

表 3b　化学物質のその他の名称

代替物質の候補	その他の名称
ジブチルフタレート	フタル酸ジブチル フタル酸n-ブチル n-ブチルフタラート Dibutyl Phthalate DBP n-Butyl phthalat Phthalic acid dibutyl
ジオクチルフタレート	フタル酸ジオクチル フタル酸ジn-オクチル ジノポールNOP Din-octyl phthalate Phthalic acid dioctyl Dinopol NOP フタル酸ジ-(2-エチルヘキシル) ビス(2-エチルヘキシル)フタラート ビス(2-エチルヘキシル)=フタラート フタル酸ジ(2-エチルヘキシル) 1,2-ベンゼンジカルボン酸ビス(2-エチルヘキシル) フタル酸ビス(2-エチルヘキシル)オクトイル オクトイル 化合物889 ジエチルヘキシルフタラート ジオクチルフタラート ジ(2-エチルヘキシル)フタラート ジエチルヘキシルフタレート フタル酸ジエチルヘキシル フタル酸ジ-2-エチルヘキシル DOP Octoil Compound 889 Dioctyl phthalate Phthalic acid bis(2-ethylhexyl) 1,2-Benzenedicarboxylic acid bis(2-ethylhexyl) ビソフレックスDOP DEHP DAF-68 Compound-889 Bisoflex DOP Bis(2-ethylhexyl) phthalate Phthalic acid di(2-ethylhexyl) ester 1,2-Benzenedicarboxylic acid bis(2-ethylhexyl) ester Phthalic acid bis(2-ethylhexyl) ester Diethylhexylphthalate Di(2-ethylhexyl)phthalate Diethylhexyl phthalate Di-2-ethylhexyl phthalate Phthalic acid di-2-ethylhexyl Phthalic acid diethylhexyl

1）化学物質基本情報

表 3 は、ジオクチルフタレートのラベルと日化辞番号を表示した例である。

2）構造

表 3a は、化学式（構造式と分子式）を連携して表示した例である。ジオクチルフタレートに、2 つの構造異性体を確認できた。

表 4　化合物プロパティ

ChEMBLの化合物プロパティ

件数:20件

name	value
Hydrogen Bond Donors	0
ACD LogD	8.52
ACD LogP	8.52
ALogP	7.57
Aromatic Rings	1
Canonical Smiles	Canonical Smiles
Molecular Formula	C24H38O4
Full Molecular Weight	390.56
Hydrogen Bond Acceptors	4
Heavy Atoms	28

表 5　製品情報

CPCatの製品情報

件数:372件

ProductName	PercentComposition	Manufacturer
COATING EC-1335	<1%	3M_COMPANY__ST__PAUL__MI
COATING EC-1335	<2%	3M_COMPANY__ST__PAUL__MI
COATING EC-1335	28.0%	3M_COMPANY__ST__PAUL__MI
COATING EC-1335	33.0%	3M_COMPANY__ST__PAUL__MI
COATING EC-1335	8.0%	3M_COMPANY__ST__PAUL__MI
GLOSS ENAMEL INK, GE-102	5-10%	INK_DEZYNE_INTERNATIONAL_
GLOSS ENAMEL INK, GE-102	0.5-1.5%	INK_DEZYNE_INTERNATIONAL_
GLOSS ENAMEL INK, GE-102	1-5%	INK_DEZYNE_INTERNATIONAL_

3）化学物質名の別称

表 3b は、化学物質名の別称を連携して表示した例である。ジオクチルフタレートは、フタル酸ジ n- オクチルとフタル酸ジ（2- エチルヘキシル）の上位概念であることや別称数が多大であることを確認できた。

4）化合物プロパティ

表 4 は、ジオクチルフタレートの化合物プロパティを連携して表示した例である。化学製品設計や開発プロセスで化学物性を推定するための基本データとして使うことができる。

5）製品情報

表 5 は、ジオクチルフタレートの製品情報を連携して表示した例である。製品含有率の違いを比較し、購入、設計、使用等で確認できる。

このように、様々な情報ソースを LOD の基盤上で連携させて使うことによって単独のデータからは得られない情報を集約して、知識として活用できるようになる。

4 おわりに

特許データをオープンデータとマッシュアップすることで新たなイノベーションや価値を生み出す可能性は大きい。ビッグデータ解析が一層進み、仮説のないところから新しい因果関係を発見する機会が増えることも期待されている。これまでデータベース作成・提供者は、そのデータベースに化学物質の外部リンク情報を記載する際、個々のリンク情報を地道に収集、整理する必要があり、データの同期がたいへん困難であった。それらの作業が LOD 連携で大幅に軽減することになる。しかし、化学物質のデータから知見を得るには、解析結果を概念的・俯瞰的・動的にいろいろな観点で検討する必要がある。そのための収集したデータをタイムリーかつ効果的に可視化する WebAPI は発展途上である。また、知識の有効活用に向け、パターンが発見できても、様々な要因が絡み、複雑化した場合、そのパターンが有効か否かの判断を下すことが困難となる。そこで、トポロジカル・データ・アナリシス TDA（Topological data analysis）のような人工知能技術をさらに取り込んでいく必要があると考えている。

さらに、現実の課題の地球環境やエネルギーの問題等に取り組もうとすると、化学の知識だけでは不十分で、化学以外の知識収集にも LOD が欠かせない。特許文書から抽出した情報とオープンデータを相補的に使うことで、より網羅的で正確な知識を活用することが可能になると考えている。

今後、化合物製造法の特許を読み解き、複雑な化学反応を呈する化学反応式群を抽出・知識化して可視化する手法に取り組みたい。そのためには、まず、化学物質と化学反応の関係を構造化したオントロジーを構築する必要があると考えている。

参考文献

[1] 池田紀子、田中一成：特許文書からの化学物質情報の抽出、Japio YEAR BOOK 2015、p.274-281（2015）

[2] 田中一成、池田紀子：特許調査業務を改善する特許読解支援システム―特許情報と技術者を近づけるための技術―、情報処理学会デジタルプラクティス Vol.7 No.4（2016）

[3] 総務省 オープンデータ戦略の推進 http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/opendata01.html#p1-1

[4] データカタログサイト http://www.data.go.jp/

[5] LOD4ALL http://lod4all.net/ja/index.html

[6] 日化辞 http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html

[7] J-GLOBAL https://jglobal.jst.go.jp/

[8] NBC RDF Portal http://integbio.jp/rdf/

[9] 日化辞 RDF データの公開と化合物情報の統合、情報管理 vol.58 no.3（2015）

[10] creativecommons https://creativecommons.org/licenses/by/3.0/

[11] EBI https://www.ebi.ac.uk/

[12] UniChem https://www.ebi.ac.uk/unichem/

[13] ChEMBL　https://www.ebi.ac.uk/chembl/

[14] CAS　http://www.cas-japan.jp/

[15] PubChem　https://pubchem.ncbi.nlm.nih.gov/

[16] ChemSpider　http://www.chemspider.com/

[17] CPCat　https://actor.epa.gov/cpcat/faces/home.xhtml