ビッグデータ時代の
オントロジー技術

山口　高平
（慶應義塾大学理工学部）
（人工知能学会　会長）

データマイニングから
ビッグデータへ

第1世代(1995-2000年前半）
データマイニング開発手順

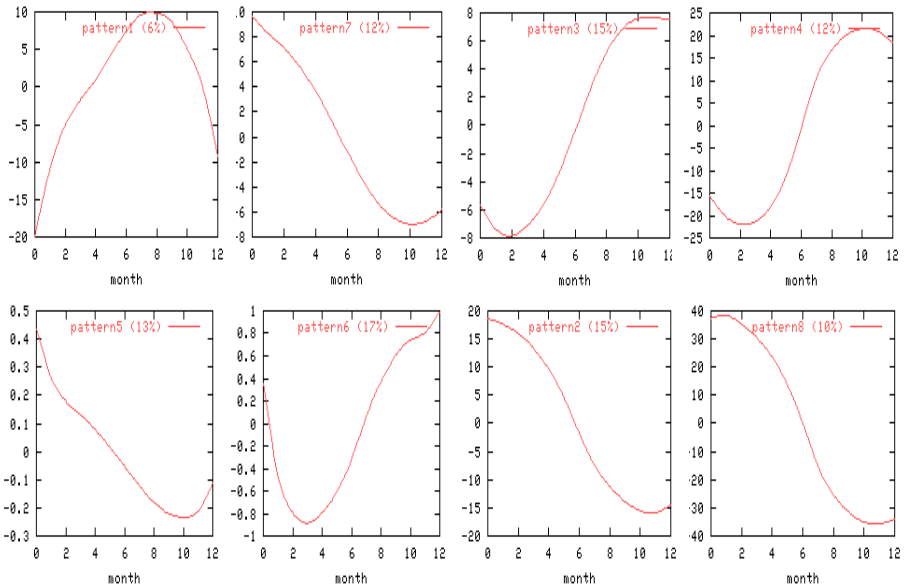
＋？

非常識な陳列？

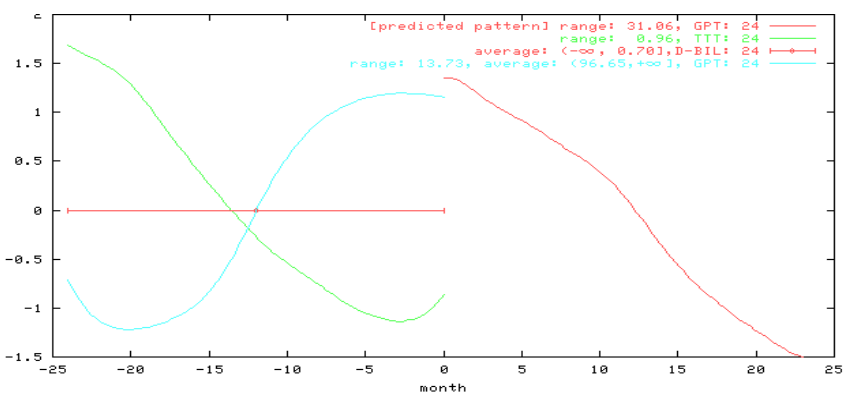
肝炎データマイニング

- 提供データセット
  - 患者基本情報
    - 患者のプロフィール
  - 検体検査結果情報
    - 検体検査（血液＆尿）の結果情報 → 院内＋外注データ
  - 肝生検情報
    - 肝生検情報（肝炎の進行具合）
  - インタフェロン投与情報
    - インタフェロンの投与時期
- データの特徴
  - 大規模な未整備時系列データ
    - 最大 160 万レコード
    - 膨大な数の表記揺れが存在
  - 検査項目数が非常に多い
    - 最大 950 項目
  - 時期により検査項目の再現性が変化＆欠損値が多い
    - 観測機器＆医学の進歩
  - 医者によるバイアスが存在
    - 重病患者には特殊な検査

データ前処理：GPTの８変化パターン



IF 直前24ヶ月のビリルビンの平均値が高く、TTT（チモール混濁試験）が減少する
THEN GPTが減少に転じる



- 予測正答率: 60.90% (21/34), 再現率: 1.43% (21/1470)

GPTは周期的な多少の上下動があるもののほぼ一定と理解してきた。
このルールは、GPTの上下動の転移を説明する可能性があり興味深い。
ウィルス活動・バクテリア増殖の周期性とも関連するのか？

データマイニングの課題

- データ整備はコストがかかる
- 他のデータの連携も調べたくなる
- マイニング結果の意味を説明しろと言われても
- マイニング結果も大量になり絞り込みたい。
- 専門家の壁（主観vs.客観）
- 組織の壁

→2000年前半「データマイナーの憂鬱」
→2011年以降「ビッグデータ」 多くの関心

（背景1）データの量、様式、更新頻度の劇的変化

http://www.soumu.go.jp/main_content/000160628.pdf

（背景2）ビッグデータ基盤技術の進展

- Hadoop（オープンソース分散並列処理技術）

http://hadoop.apache.org/images/hadoop-logo.jpg

- NoSQLデータベース：非構造の大量データ処理可能
（スキーマフリー、スケールアウト）

- ML/DMの進展:
カーネル関数によるSVM
CRF(Conditional Random Filed, 条件付確率場)
ベイズモデル
Deep Learning（多層ニューラルネット）

http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all

（背景3）国策としてのビッグデータ

Office of Science and Technology Policy
Executive Office of the President
March 29, 2012

OBAMA ADMINISTRATION UNVEILS
“BIG DATA” INITIATIVE:
ANNOUNCES $200 MILLION
IN NEW R&D INVESTMENTS

ビッグデータの現状

- ソーシャルメディア、位置データの情報発信環境、HADOOP, NOSQLといった情報管理環境が整い、多種多様大規模データが扱える環境が整う
- ユーザ行動履歴を中心に、見える化が進み、ビジネスチャンスが広がる
- でも、高度な分析にはセマンティクスが必要となり、オントロジーのような意味処理技術との連携が必要とされるであろう。

オントロジー技術

知識工学とセマンティックWebにおけるオントロジーの研究開発

知識工学

1991-現在

- 概念化の明示的仕様
  （Tom Gruber オントロジーの定義）
- オントロジー記述言語(Ontolingua)
- 知識交換言語(KIF)
- PSM
- Task Ontology
- Generic Ontology
- CYC, WordNet, EDR...
- オントロジー構築方法論

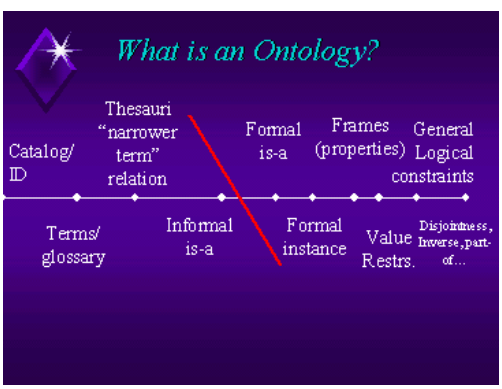
セマンティックWeb

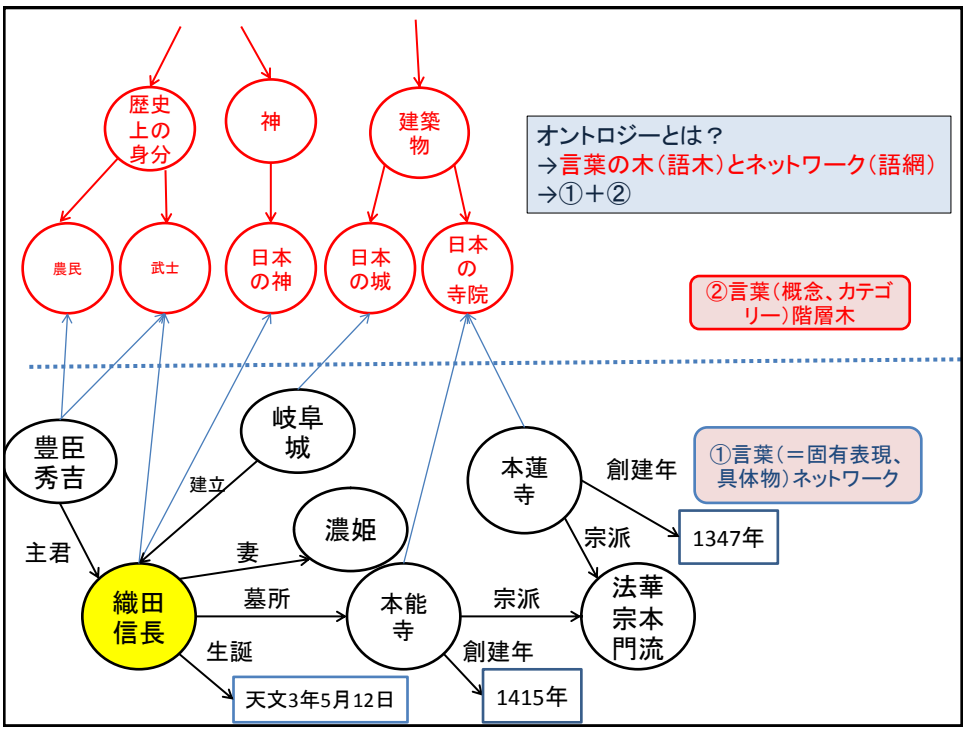
1997-現在

- 95-97: XML as arbitrary structures
- 97-98: RDF
- 98-99: RDFS
- 00-01: DAML+OIL
- *2004.2.10: OWL*
- *2009.10.27: OWL2*
- *2010.6.22: RIF*

オントロジーとは？

- 哲学のオントロジーvs.情報系オントロジー（上位オントロジー）
- 存在論vs.存在観：モデリングプリミティブ（領域オントロジー）
- UMLダイアグラムvs.オントロジー（コンピュータが理解・処理可能）
- プロセスvs.プロダクト：体系化vs.概念仕様
- 概念（化）の明示的仕様：クラス，プロパティ，公理，インスタンス

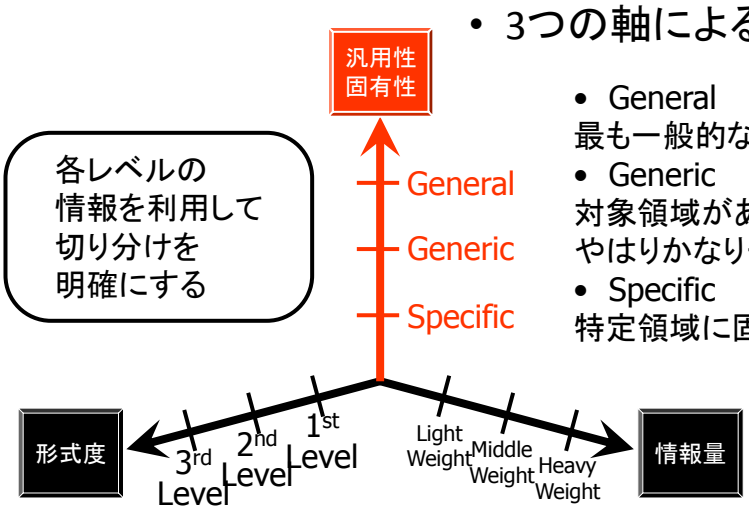




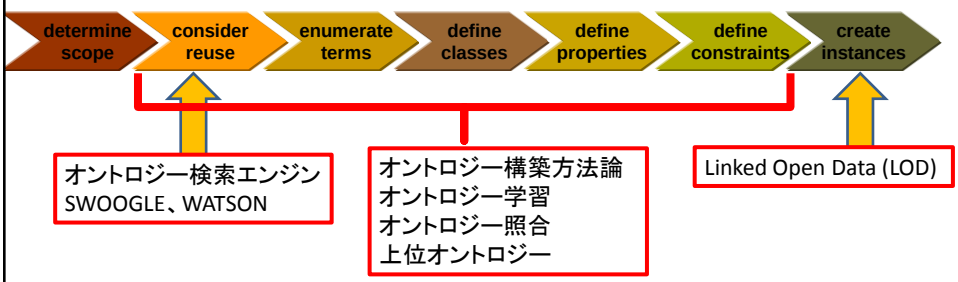
オントロジーの分類

- 3つの軸による分類
  - General
    最も一般的なレベル
  - Generic
    対象領域があるが
    やはりかなり一般的
  - Specific
    特定領域に固有

各レベルの
情報を利用して
切り分けを
明確にする



オントロジー開発手順



WordNet

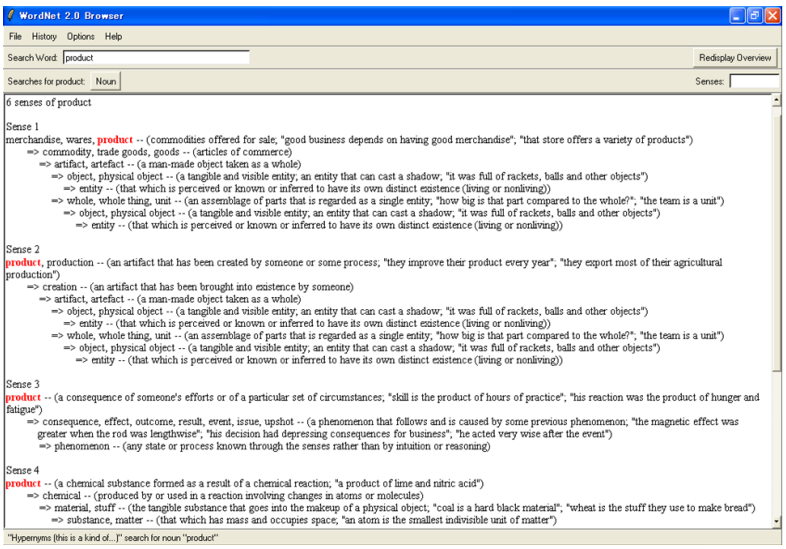
- http://wordnet.princeton.edu/
- 最新版：ver.3.0 for Unix-like system
  - Windows版はver.2.1
- 約11万7千のsynset(同義語の集合)
- 約15万語（名詞，動詞，形容詞，副詞）
- synset間には，品詞ごとにいくつかの関係が定義されている

日本語ワードネット1.1 by NICT

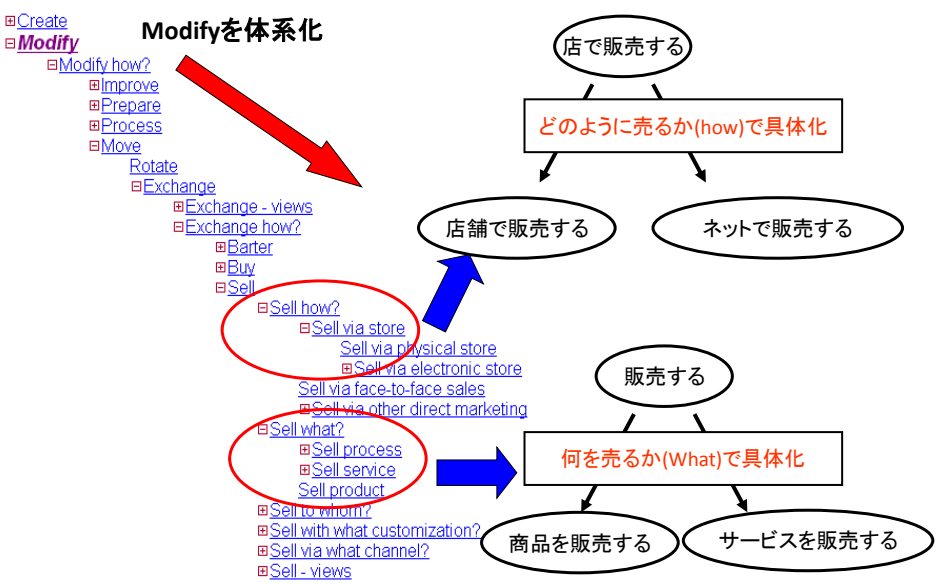
57,238 概念 (synset数),93,834 words 語

http://nlpwww.nict.go.jp/wn-ja/index.ja.html

WordNet 実行例

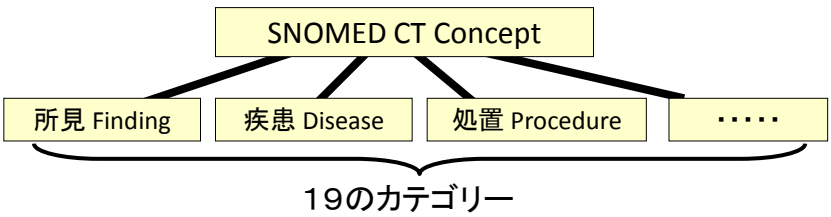


ビジネスプロセスオントロジー
Process Handbook (MIT)



医療オントロジー：SNOMED-CT

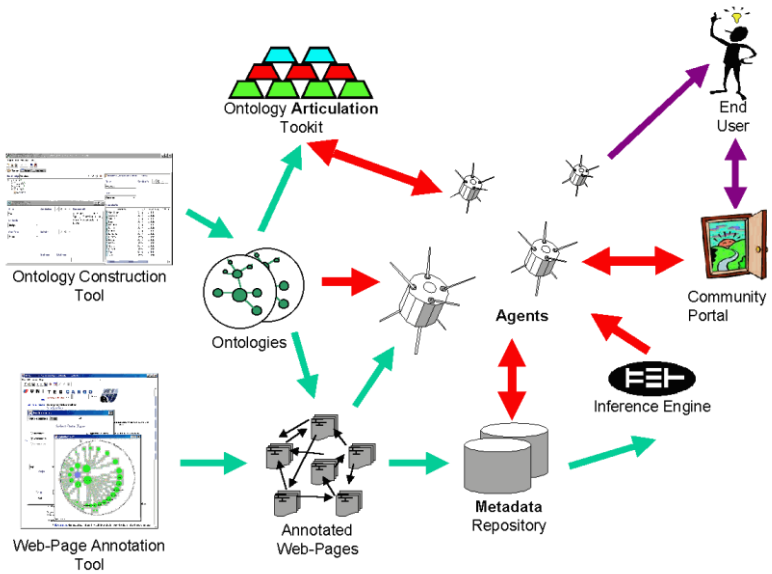
- すべての概念は「SNOMED CT Concept」という最上位概念の下位概念になっている



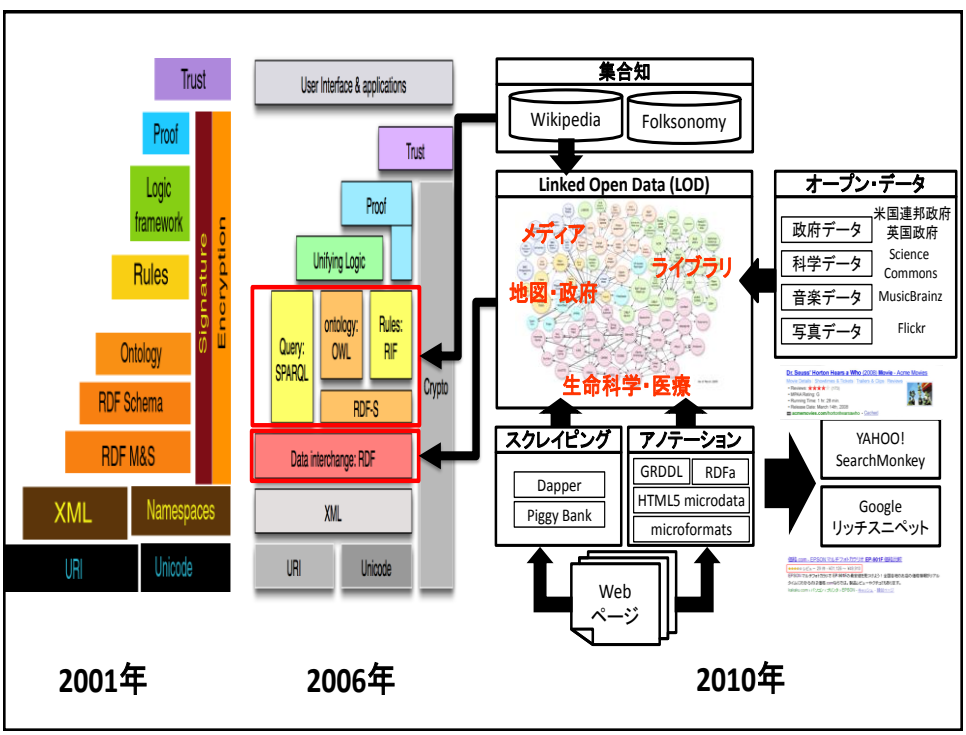
- 「所見」「疾患」「処置」などの最上位概念のすぐ下の概念を、「カテゴリー」と呼び、全部で１９種類ある。
  すべての概念(34万）はいずれかのカテゴリーに属する

Webとオントロジー
Semantic Web

Big Picture for Semantic Web (2001)

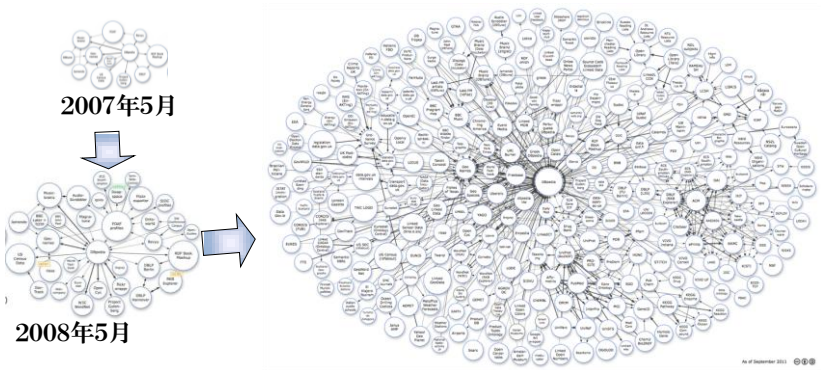


http://www.semanticweb.org/about.html#bigpicture



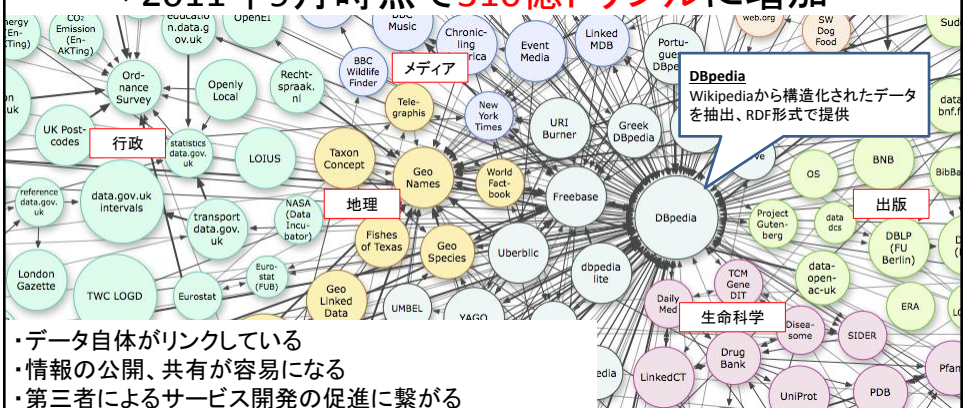
Linked Open Dataの普及

- Web上で公開され，相互に連結し合っているRDFデータ
  - これまで多く研究されてきた抽象的な概念構造が現実的な有用性を生むには依然高いハードルがある
  - 具体物であるインスタンスの記述をしたRDF（Linked Open Data）のデータベースを公開・共有し合うべきという風潮が高まっている



RDFモデルによるLinked Open Data（LOD）

- LOD規模：5億トリプル（2007）
  ⇒2011年9月時点で310億トリプルに増加

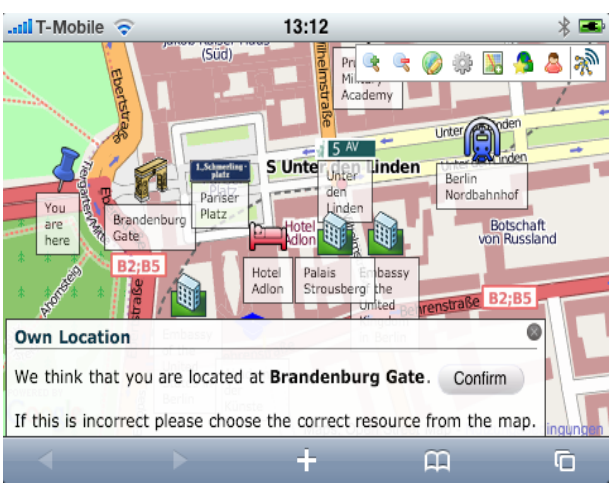


・データ自体がリンクしている
・情報の公開、共有が容易になる
・第三者によるサービス開発の促進に繋がる
➡ **情報流通基盤として期待が集まる**

LODの例

- DBPedia (2007年~)
  - 英語版Wikipediaから構築された，LODのハブ的存在
- BBC (2009年~)
  - 英国放送協会が提供しているニュースとテレビ番組の情報
- News York Times (2009年~)
  - 蓄積された新聞記事に現れる人名，組織．団体名，地名，主題のキーワード約1万字にURIを与えてLODとして公開

DBpedia mobile

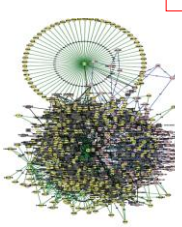


Linked Dataと
日本語Wikipediaオントロジー

日本語Wikipediaオントロジー

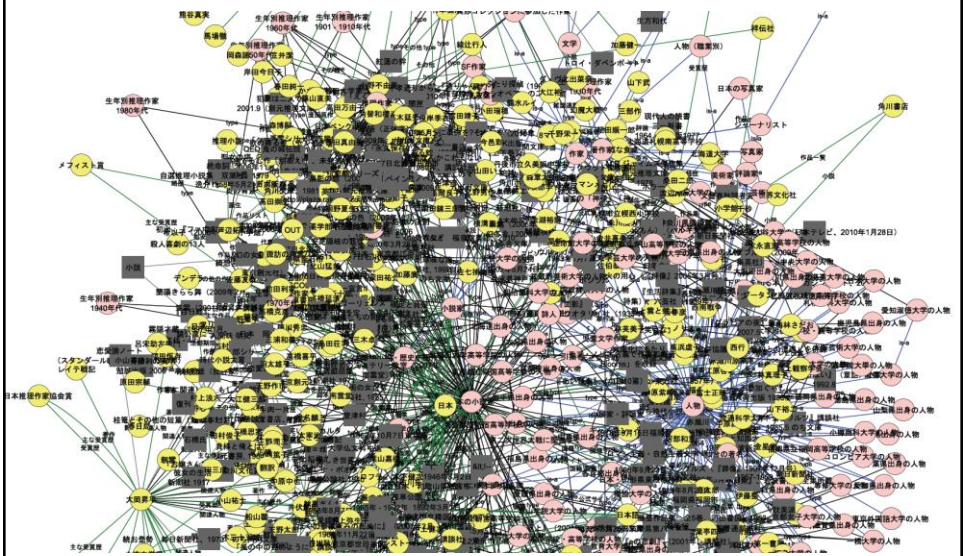
人間には，ウィキペディアの内容（意味）が判るけど人工物（コンピュータ，スマフォ，ロボット．．．）には判らない

Wikipediaからオントロジー（言葉階層木，言葉のネットワーク）に自動変換して，
人工物に言葉の意味(Sense)を理解させる
→日本語Wikipediaオントロジー

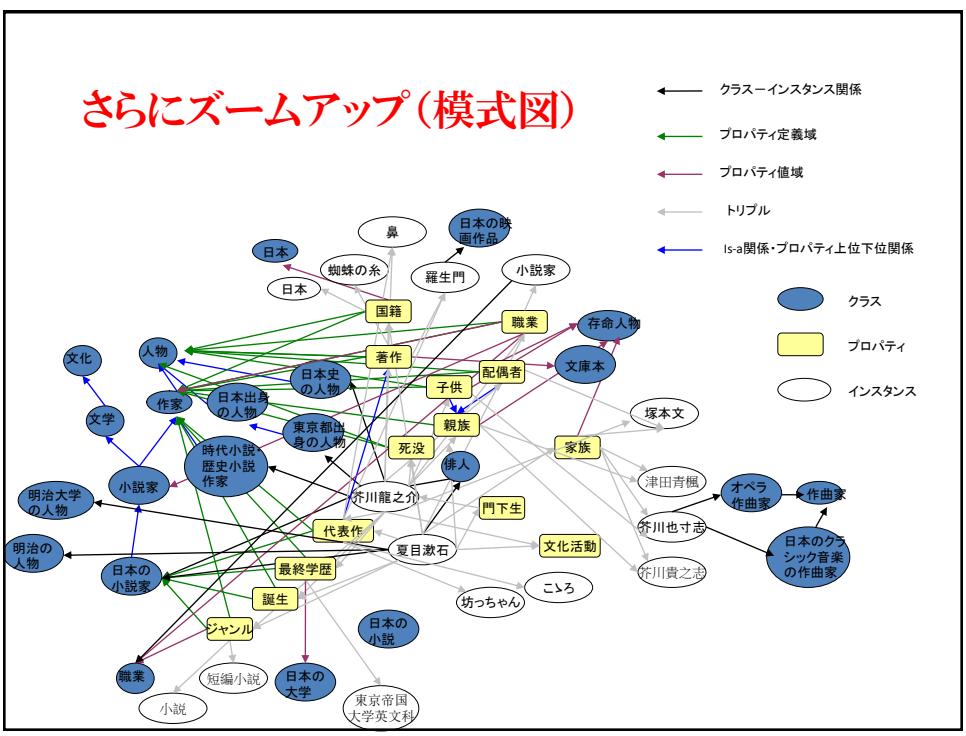


クラス数	51322
インスタンス数	1373953
プロパティ数	22639
is-a関係数	37745
タイプ数(rdf:type)	505487
定義域関係数(rdfs:domain)	13391
値域関係数(rdfs:range)	15034
プロパティトリプル	6236495
Infoboxトリプル数	3964027

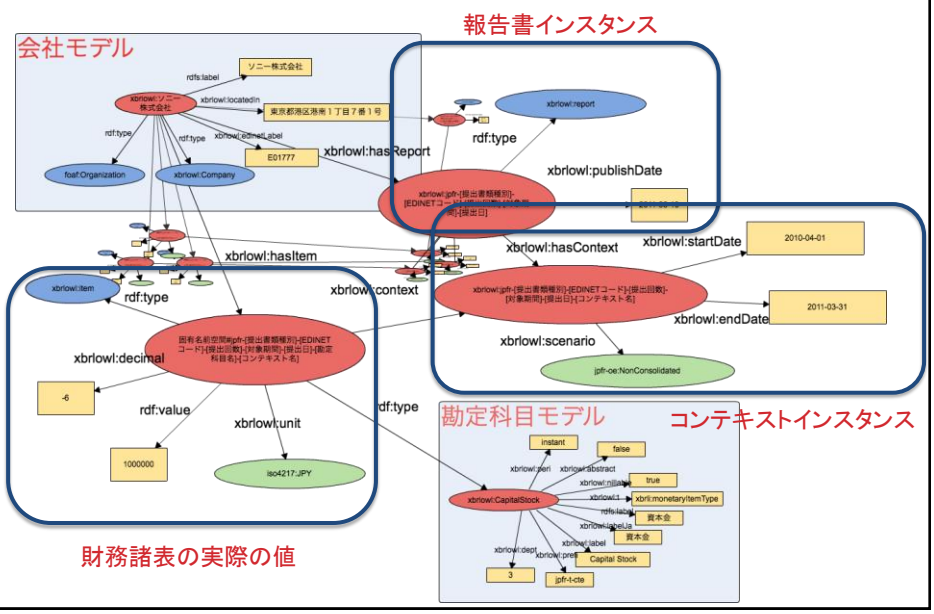
日本語Wikipediaオントロジー（文学）



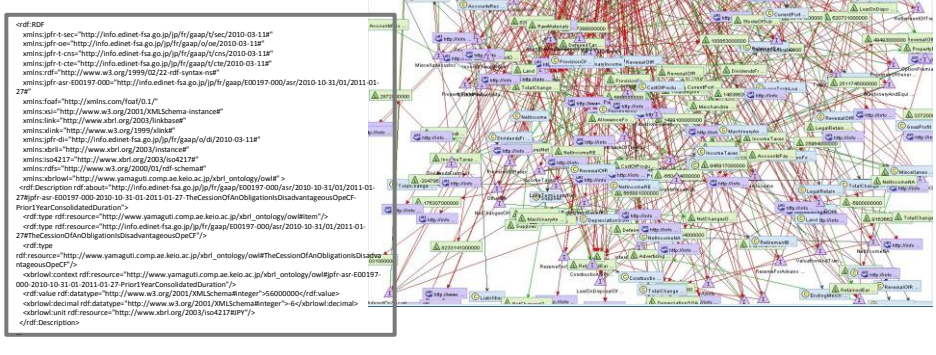
さらにズームアップ（模式図）



XBRL LOD

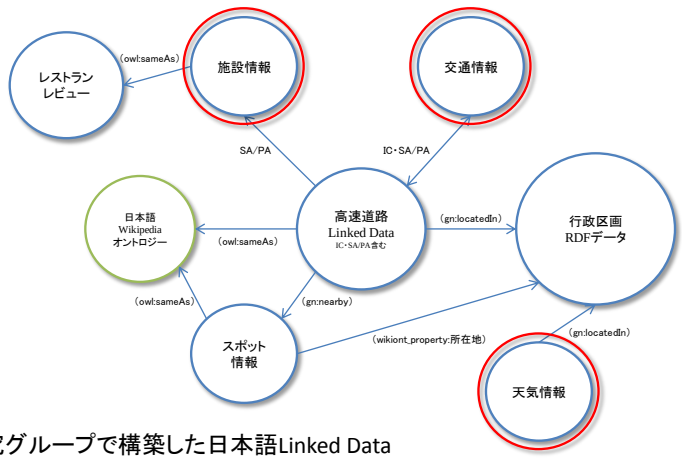


一報告書に含まれるXBRL LOD（平均約1200トリプル）



```
<rdf:RDF
  xmlns:jpfr-t-sec="http://info.edinet-fsa.go.jp/jp/fr/gaap/t/sec/2010-03-11#"
  xmlns:jpfr-oe="http://info.edinet-fsa.go.jp/jp/fr/gaap/o/oe/2010-03-11#"
  xmlns:jpfr-t-cns="http://info.edinet-fsa.go.jp/jp/fr/gaap/t/cns/2010-03-11#"
  xmlns:jpfr-t-cte="http://info.edinet-fsa.go.jp/jp/fr/gaap/t/cte/2010-03-11#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:jpfr-asr-E00197-000="http://info.edinet-fsa.go.jp/jp/fr/gaap/E00197-000/asr/2010-10-31/01/2011-01-27#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance#"
  xmlns:link="http://www.xbrl.org/2003/linkbase#"
  xmlns:xlink="http://www.w3.org/1999/xlink#"
  xmlns:jpfr-di="http://info.edinet-fsa.go.jp/jp/fr/gaap/o/di/2010-03-11#"
  xmlns:xbrli="http://www.xbrl.org/2003/instance#"
  xmlns:iso4217="http://www.xbrl.org/2003/iso4217#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xbrlowl="http://www.yamaguti.comp.ae.keio.ac.jp/xbrl_ontology/owl#" >
  <rdf:Description rdf:about="http://info.edinet-fsa.go.jp/jp/fr/gaap/E00197-000/asr/2010-10-31/01/2011-01-27#jpfr-asr-E00197-000-2010-10-31-01-2011-01-27-TheCessionOfAnObligationIsDisadvantageousOpeCF-Prior1YearConsolidatedDuration">
    <rdf:type rdf:resource="http://www.yamaguti.comp.ae.keio.ac.jp/xbrl_ontology/owl#item"/>
    <rdf:type rdf:resource="http://info.edinet-fsa.go.jp/jp/fr/gaap/E00197-000/asr/2010-10-31/01/2011-01-27#TheCessionOfAnObligationIsDisadvantageousOpeCF"/>
    <rdf:type
rdf:resource="http://www.yamaguti.comp.ae.keio.ac.jp/xbrl_ontology/owl#TheCessionOfAnObligationIsDisadvantageousOpeCF"/>
    <xbrlowl:context rdf:resource="http://www.yamaguti.comp.ae.keio.ac.jp/xbrl_ontology/owl#jpfr-asr-E00197-000-2010-10-31-01-2011-01-27-Prior1YearConsolidatedDuration"/>
    <rdf:value rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">56000000</rdf:value>
    <xbrlowl:decimal rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">-6</xbrlowl:decimal>
    <xbrlowl:unit rdf:resource="http://www.xbrl.org/2003/iso4217#JPY"/>
  </rdf:Description>
```

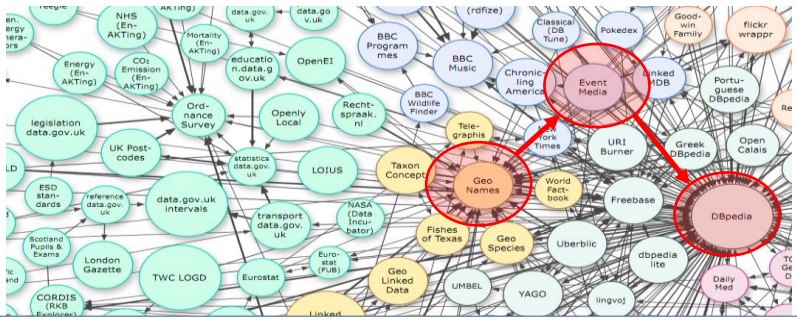
車による移動支援サービス



青:研究グループで構築した日本語Linked Data
赤:仮想日本語Linked Data(企業が所有するデータ)
緑:研究室内に存在する日本語Linked Data

横断検索のイメージ

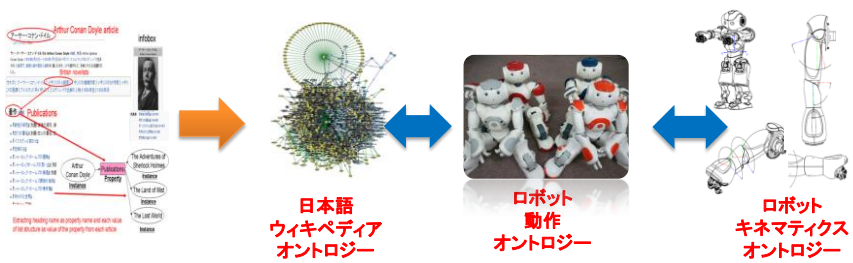
日本語LODが普及すれば



・**DBpedia＋Geonames＋EventMedia**

⇒現在位置＝箱根、周辺の観光スポットを見たい
⇒箱根の情報を放送した番組に「いい旅夢気分」がある
⇒「いい旅夢気分」で紹介された観光スポットに関する情報を
「日本語Winkipediaオントロジー」で確認する

オントロジーロボット



ビッグデータとオントロジー技術

- 現状のビッグデータ→見える化、分析の段階
- 構造データと非構造データの連携が重要
- でもデータ統合・連携にはセマンティクス、オントロジーが必要
- 非構造データ、LOD、オントロジーの連携により、インテリジェントサービスの開発が期待される