セマンティックWebと多言語処理

立命館大学 理工学部情報学科
前田 亮

発表の概要

- セマンティックWebとは?
- セマンティックWebを支える技術
- Webにおける多言語処理の現状
- セマンティックWebにおける多言語処理
  - 多言語横断検索
  - 多言語オントロジの構築
- 応用事例
  - 古文書ディジタル図書館

セマンティックWebとは?

- 1998年にWebの創始者Tim-Berners Leeが提唱
  - 「機械（エージェント）が意味的に処理できる次世代Webを目指す」
- Web情報の意味（セマンティクス）をコンピュータが理解できるようにする
  - 人と機械のコミュニケーション
- Web上でのさまざまな問題解決を支援
  - 例：わからないことを調べる，旅行の手配，オンラインショッピング，etc.

現在のWebとの比較

- 「現在のWeb」**は人間が読むためのWeb**
  - ***コンピュータに自然言語の理解は難しい！***
  - ***単純な処理しかできない (e.g.文字列マッチ)***
- 「セマンティックWeb」**はコンピュータが理解できるWeb**
  - ***情報が表す意味をコンピュータが理解できる形で記述***
  - ***コンピュータによる知的処理が可能に！***

現在のWeb検索エンジン

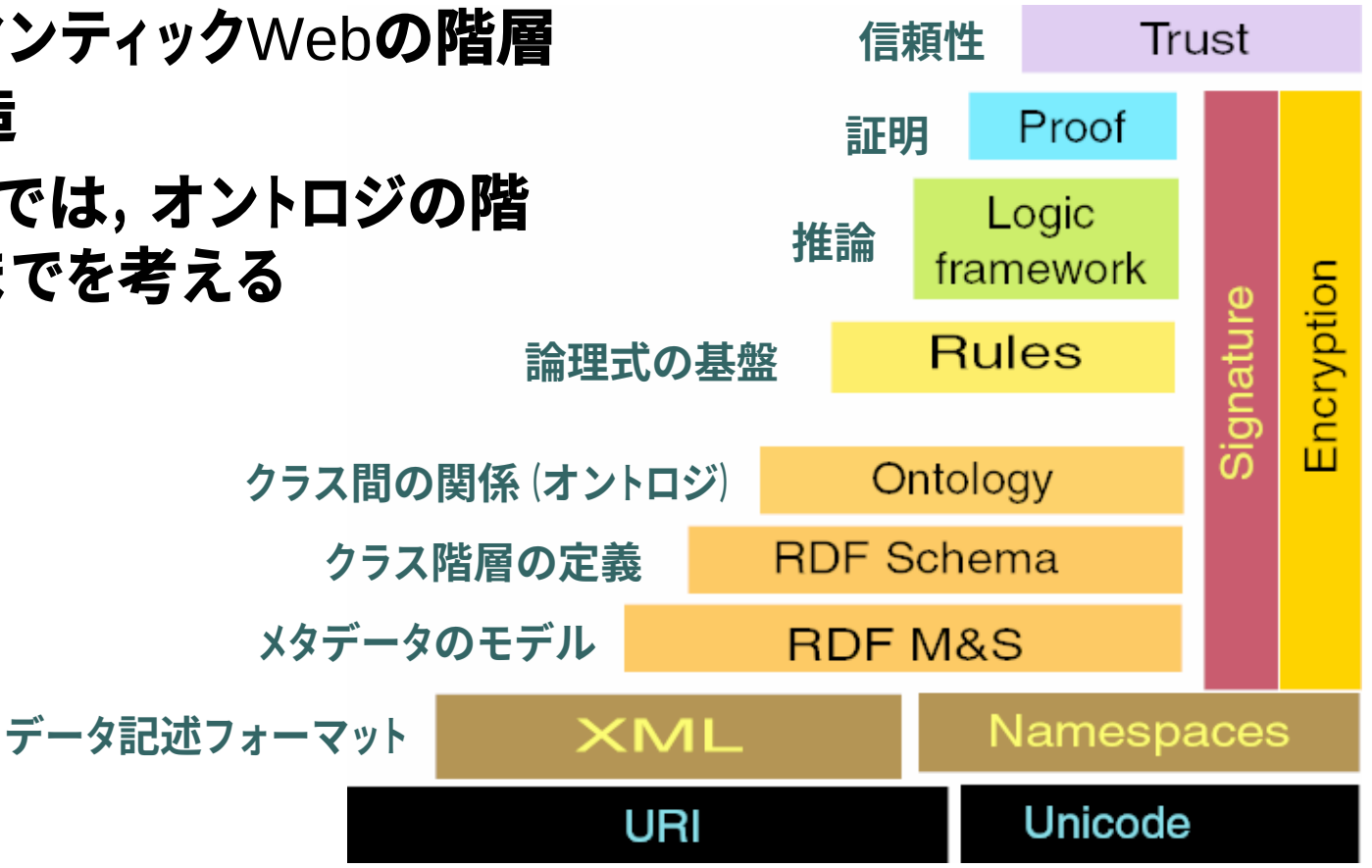
- ***「今日開いている藤沢の歯医者は?」***
- Web**検索エンジンで検索すると…**
  - **藤沢さんがやっている歯医者も検索されてしまう(人名と地名の区別がつかない)**
  - **「歯科」「デンタルクリニック」などは検索されない**
  - **ページに診療日が載っていても，今日開いているかどうかは検索できない**

セマンティックWebでは. . .

- Webページにコンピュータが理解できるメタデータを付与
  - 住所：神奈川県藤沢市・・・
  - 業種：歯科医
  - 診療日：月～金 (第3水曜休診)
- 背景となる知識がある
  - 「歯科医」と同じ概念を表すものして「歯医者」「デンタルクリニック」
  - 1週間は月火水木金土日の順，「休診」は診療しない日

セマンティックWebを支える技術

- **セマンティックWebの階層構造**
- **ここでは，オントロジの階層までを考える**

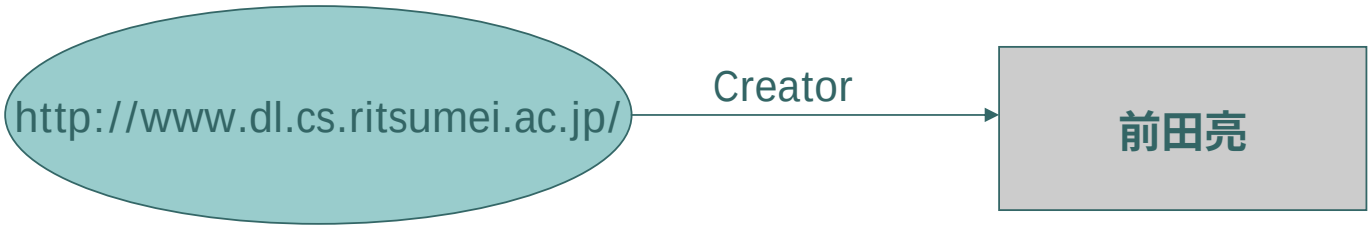


メタデータ

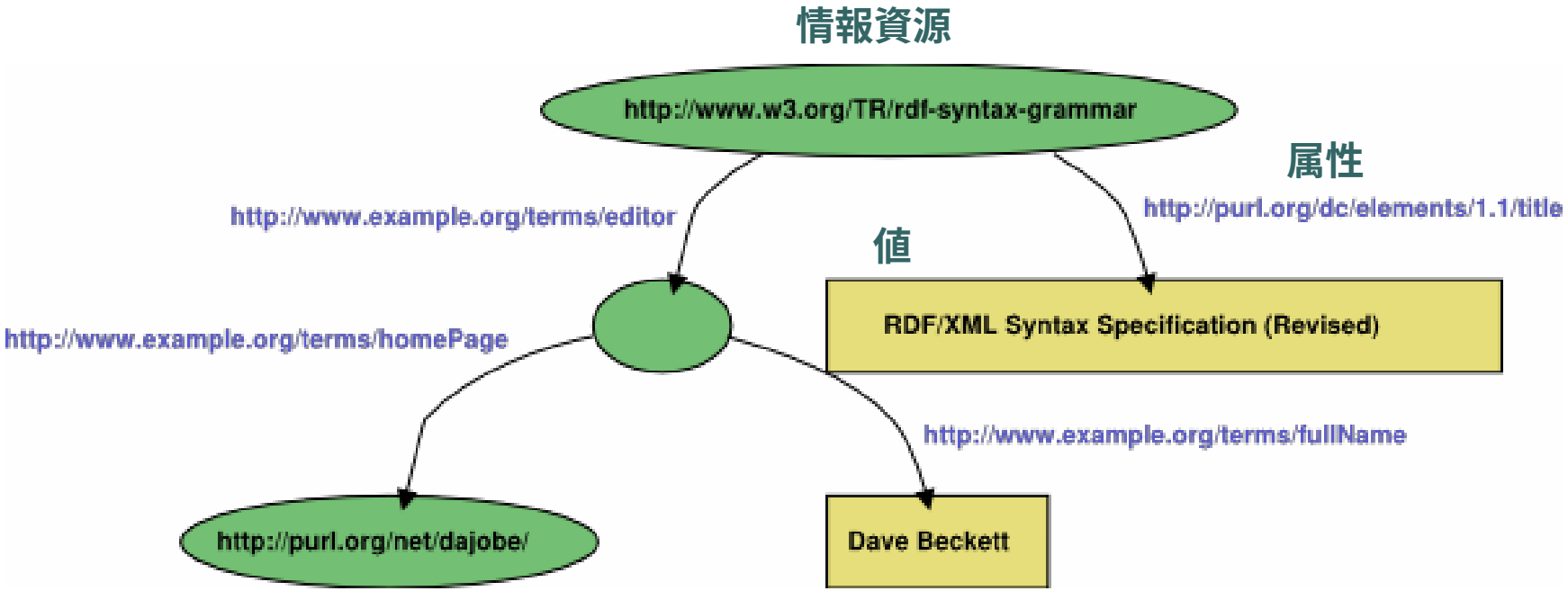
- **「データに関するデータ」**
  - **本の書誌情報など**
- DublinCore
  - **情報資源の基本的なメタデータ要素を定義**
  - **特定の表現形式は持たない**
- RDF (Resource Description Framework)
  - **メタデータを記述する表現形式**
  - **3つ組モデル**

RDFのモデル（1）

- 「情報資源」，「属性」，「値」の3つ組



RDFのモデル (2)

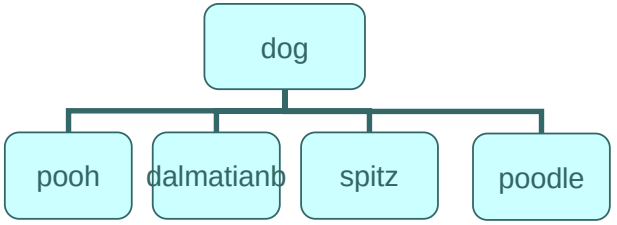

情報資源
http://www.w3.org/TR/rdf-syntax-grammar
属性
http://www.example.org/terms/editor
http://purl.org/dc/elements/1.1/title
値
RDF/XML Syntax Specification (Revised)
http://www.example.org/terms/homePage
http://www.example.org/terms/fullName
http://purl.org/net/dajobe/
Dave Beckett

オントロジ

- 「対象とする世界に存在するものごとを体系的に分類し，その関係を記述したもの」
  - 語彙の定義
  - 構造の定義
  - 語彙と構造の関係の定義
- コンピュータがメタデータを理解するための背景知識
- シソーラスは，ある言語の語彙の関係を階層構造で定義したもの（オントロジの一種）

WordNet

- **英語の語彙のシソーラス（オントロジ）**
  - **約10万語の単語を概念によって分類し，階層構造で記述したもの**



- WordNet.OWL
  - WordNetを，セマンティックWebで用いるオントロジの形式に変換したもの

Web初期の多言語処理

- Webに対する主な処理は「表示」「入力」「検索」
- 「表示」
  - **自国語と英語以外表示できない！**
- 「入力」
  - **自国語と英語以外入力できない！**
- 「検索」
  - **自国語と英語以外検索できない！**

多言語処理の現状

- 「表示」は解決されつつある
  - パソコンではそうだが，携帯・PDAでは?
- 「入力」はかなり進歩している
  - Windows XPでは，数10言語の入力メソッドをインストール可能
- 「検索」もかなり進歩している
  - Googleは30ヵ国語以上のWeb文書を検索可能
  - 日本語だけではWebの1割程度しか検索できない！

多言語情報検索

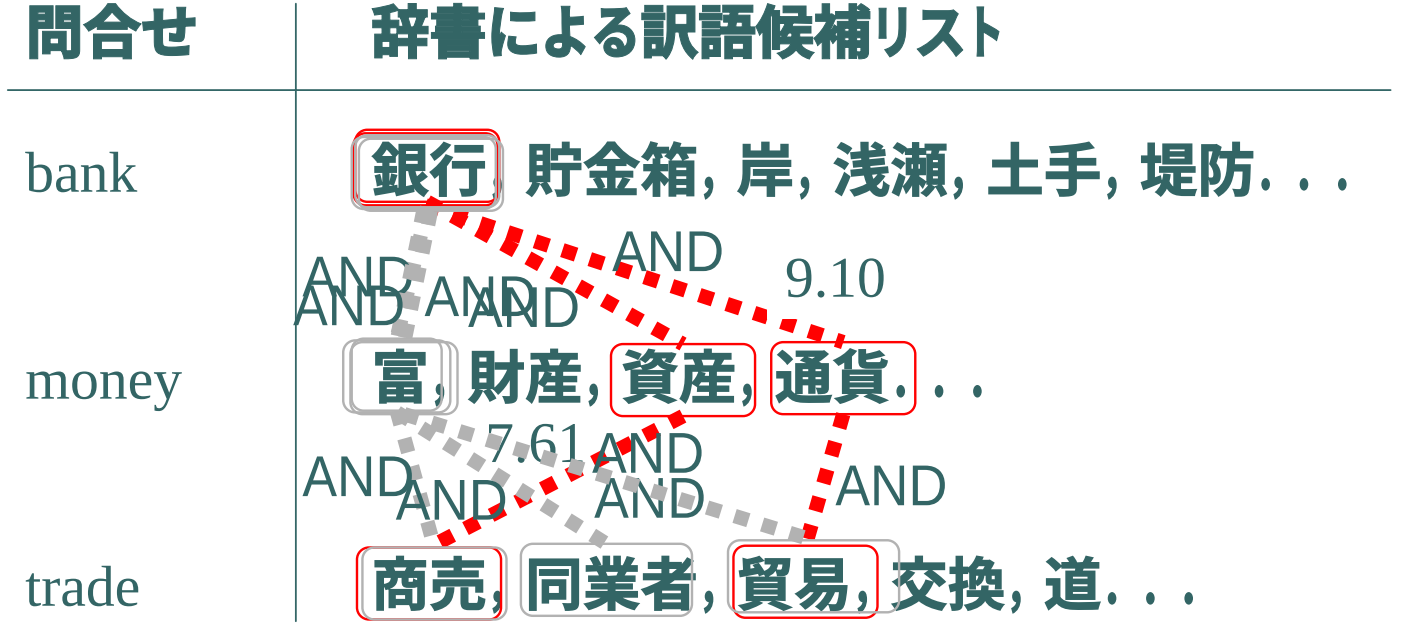
- level1：一つの検索システムで複数の言語に対応
  - 現状のWeb検索エンジンはこのレベル
- level2：ある言語のキーワードで，別の言語で書かれた文書を検索（言語横断検索）
  - 研究レベルでは実現されている
- level3：ある言語のキーワードで，Web全体を検索（多言語横断検索）
  - 辞書などの言語資源を用意できない！

言語横断情報検索

- **日本語でキーワードを入れれば，関連する英語の文書も検索してくれる**
  - **キーワードを翻訳**
  - **辞書を引いただけでは，異なる意味の訳語も（訳語の曖昧性）**
  - **コーパス中の単語間の関連性などを用いて，訳語の曖昧性を解消**
    - **Webのような，あらゆる分野を網羅するコーパスは入手困難**

Web検索エンジンを用いた曖昧性解消

- **単語の組をWeb検索エンジンでAND検索し，検索文書数をその単語組の関連の強さとする**



セマンティックWebと多言語処理

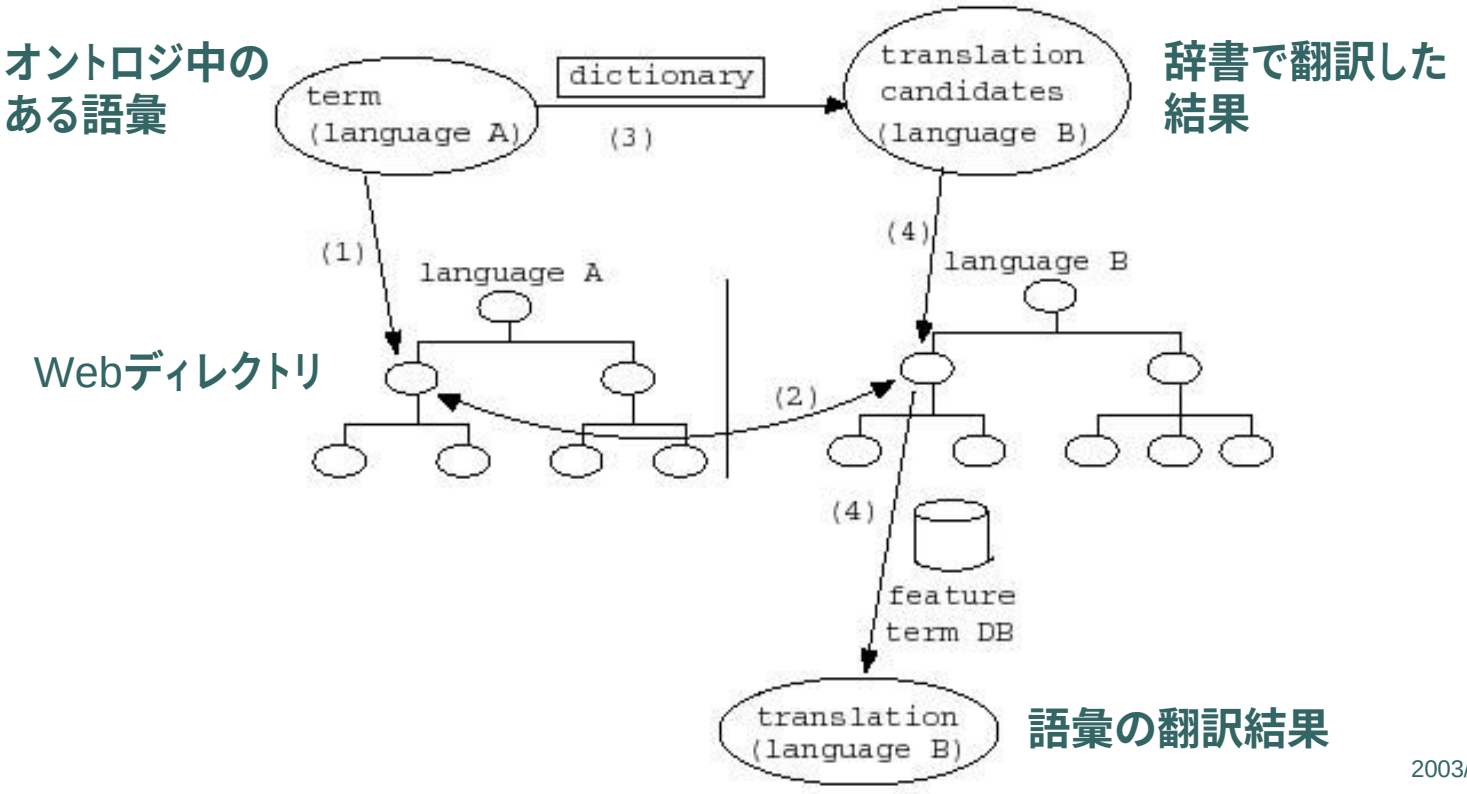
- **セマンティックWebは文書の意味を記述**
- **言語に依存しない部分もある**
  - **日付, 曜日, 地名, 人名,** etc.
- **言語に依存する部分はオントロジで解決?**
  - **多言語オントロジが必要**
- **多言語オントロジ構築の試み**
  - EuroWordNet, GlobalWordNet
  - **Webディレクトリを用いたオントロジの翻訳**

Webディレクトリは貴重な言語資源?

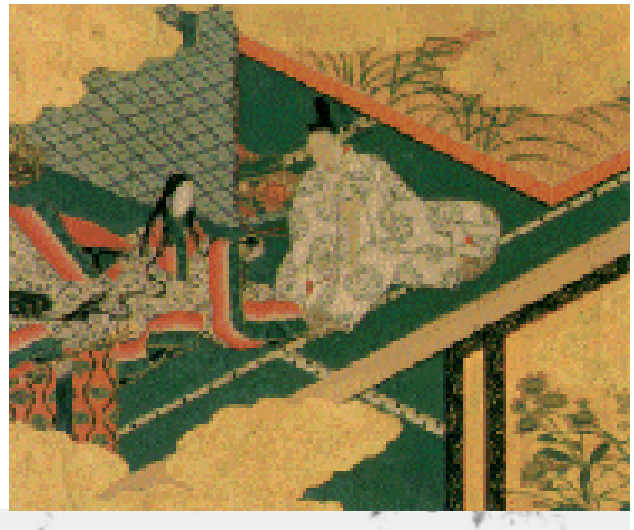
- **一応, 概念 (カテゴリ) の階層構造になっている**
  - **かなり適当な分け方だとしても…**
- **概念ごとに豊富なWeb文書 (コーパス)**
  - **統計的言語処理が可能**
- **言語別に多数の版が存在**
  - Yahoo!**には, 20種類以上の言語版が存在**

Webディレクトリを用いたオントロジの翻訳

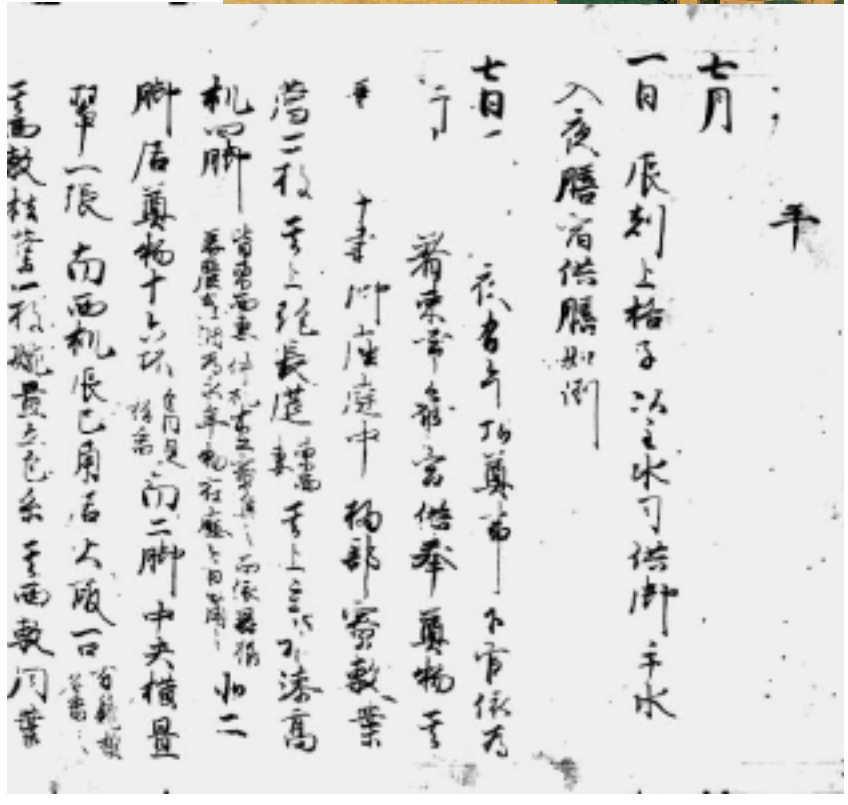
- **昨日のディジタル図書館ワークショップで発表**



古文書ディジタル図書館



- **平安時代の貴族の日記「兵範記」のディジタル化と組織化**
- **日記上で様々な書き方がされている人名・地名・建造物名から，それらに関する情報へのリンクを自動的に生成**
  - 固有表現の抽出
- **現代の言葉を使って古文書を検索**
  - セマンティックWebを用いた概念検索



まとめ

- セマンティックWebも文書を扱うので，結局言語に依存
- であれば，多言語処理の可能性
- 情報検索などの応用であれば，完全に正確でなくても良い
- コンピュータが自然言語を完璧に扱えるのは遠い将来だが，近未来の多言語処理に現実的な解決策の一つ
- 言語資源（オントロジ）の整備が当面の課題