

ウェブで使える「常識」の辞書

NII

武田 英明(情報学プリンシプル研究系)・大向 一輝(コンテンツ科学研究系)

どんな研究?

ウェブ上でコンピュータが自律的な情報処理を行うセマンティックウェブ環境では、事物や概念に関する辞書が必要となります。

辞書の構築には多大なコストがかかりますが、近年ではWikipediaや各種データベースのデータを構造化し再利用する「リンクトデータ」が注目されています。

本研究では、200万語規模の日本語リンクトデータの構築・提供を行うとともに、これらのデータを利用したアプリケーションの開発を行っています。

セマンティックウェブと日本語リンクトデータ

<http://lod.ac>

ウェブの情報流通

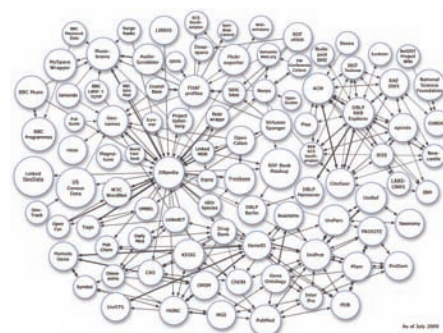
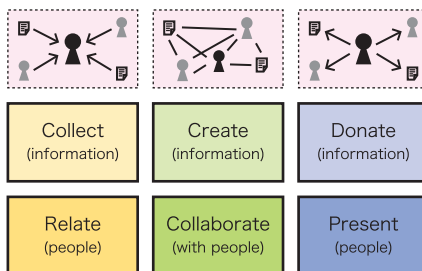
情報のネットワークと人のネットワークの相互接続
ユーザ参加・コミュニティが必須
同じ事物・同じ概念を指すためのしくみが必要: URI

メタデータの創造的流通

メタデータを自由・柔軟に生成し、流通させる
データ指向・コミュニティ指向
URIを用いて「ゆるく」つなぐ: リンクトデータ

LODACプロジェクト

図書館・博物館・美術館などのアカデミックな情報をリンクトデータに
サービス・アプリケーションが自由に情報を活用できる環境を提供
現在の活動
リンクトデータ対応の横断検索エンジンの開発
スプレッドシートからのリンクトデータ生成



lod.ac

日本語リンクトデータリポジトリ「ことはふ」

共同研究者: 小出 誠二(総合研究大学院大学) 丹 英之(株式会社アルファシステムズ)

<http://wordnet.jp/kotohub>

WordNet日本語版・RDF/OWL表現

プリンストン大学で構築されている概念辞書WordNetを
NICTが日本語化したもの
セマンティックウェブ対応・OWL-Full推論が可能に

ことはふ

日本語の「ことば」を網羅的に集める
各種辞書の見出し語のみを取得: 295万語
パーマリンク・構造化データ(RDF)を提供
ライトウェイト・オントロジー
意味によるつながり・読みによるつながり
異なるサービスをつなぐための共通語彙

→ 日本語リンクトデータのハブを目指す



辞書名	取得語数	その辞書のみ	%
日本語WordNet	231,865	152,304	65.9
国立国会図書館件名標目表(2008)	51,093	20,155	39.4
IPAdic legacy	232,002	113,312	48.8
Wikipedia(日本語)	627,411	411,582	65.6
はてなキーワード	268,539	109,477	40.8
ニコニコ大百科(仮)	48,939	26,484	54.1
Yahoo!百科事典	105,056	22,707	21.6
コトバンク	385,548	198,238	51.4
Weblio	953,941	810,494	85.0

メタデータに基づくインタラクティブ検索「DashSearch」

共同研究者: 後藤 孝行(総合研究大学院大学)

<http://www.ahirulab.com/DashSearch/>

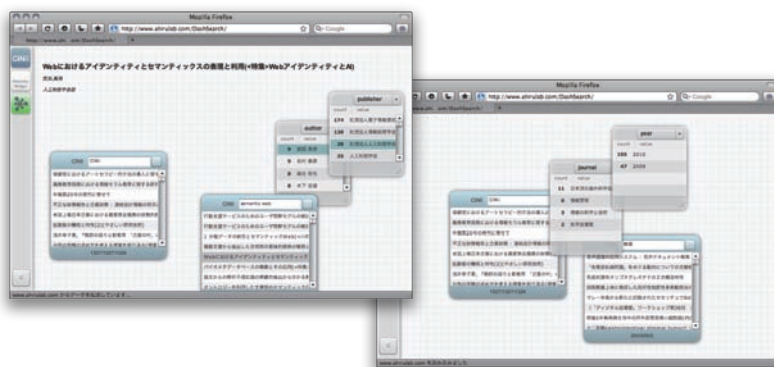
メタデータ検索の難しさ

情報要求を具体化させる必要がある
情報に付与されているメタデータ構造を理解する必要がある

→ 試行錯誤の過程でこれらを理解させるようなシステムを提案
繰り返ししやすいインターフェイス
検索結果を直感的に把握できる可視化

DashSearch

デスクトップウィジェットを用いたメタデータ検索
ウィジェット同士を接触させることで検索結果が変化
メタデータの提示とその選択行為による検索
CiNiiの文献データを対象としたプロトタイプ構築
1200万論文のファセット検索・統合検索



連絡先: 武田 英明(たけだ ひであき) 〒101-8430 東京都千代田区一ツ橋2-1-2 国立情報学研究所 Tel 03-4212-2543 Fax 03-3556-1916 takeda@nii.ac.jp <http://www.kasm.nii.ac.jp/>