

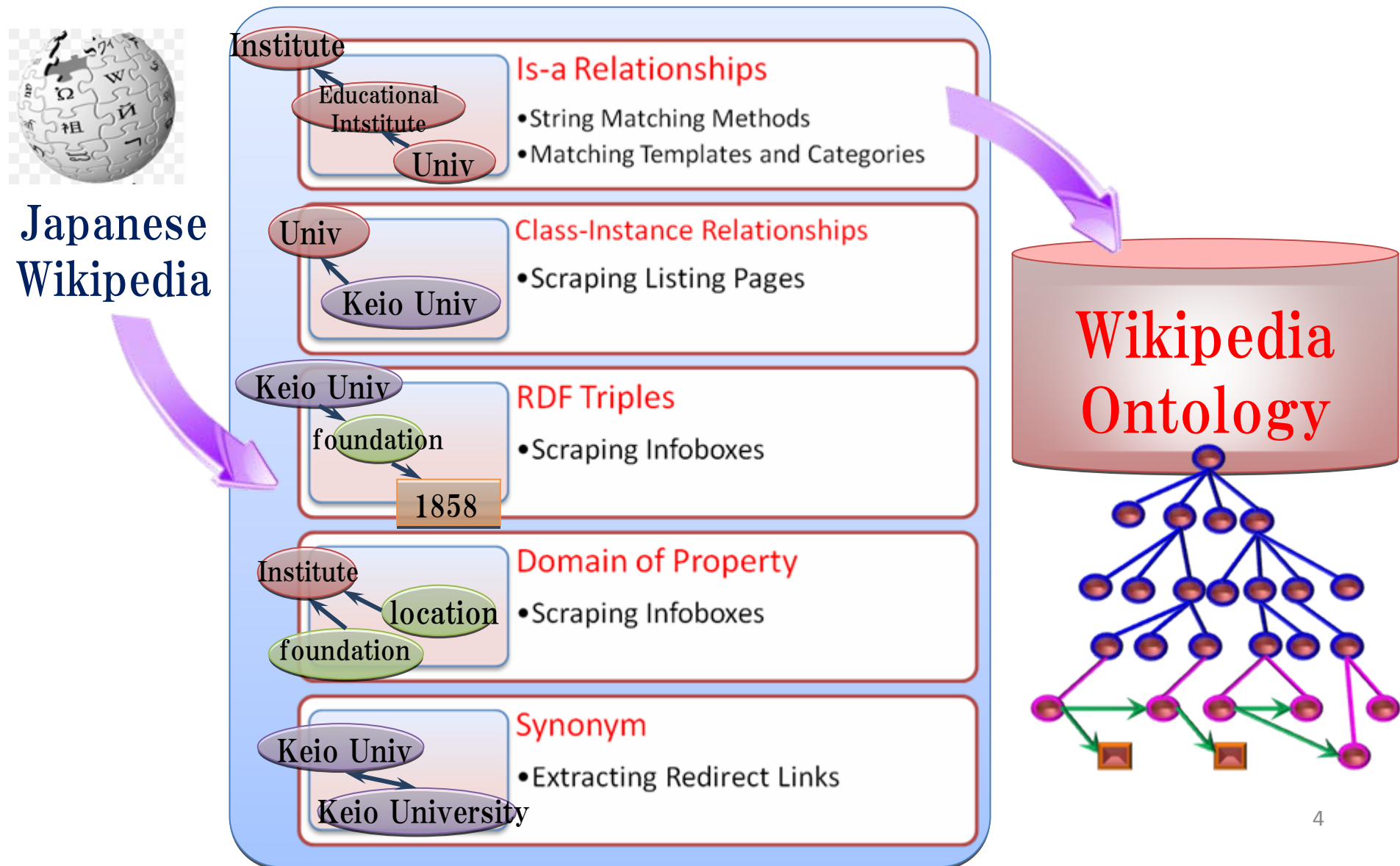
Building up a Large Scale of Ontology from Japanese Wikipedia

Takahira Yamaguchi
Keio University, Japan

Today's Talk

- Background
- Proposed Methods
 - IS-A Hierarchy, Class – Instance, RDF triple, Property Domains, Synonyms
- Application 1 (Knowledge Creation Support) Demo
- Application 2 (Human Robot Interaction) Demo
- Evaluation (Results, Related Work)
- Conclusions and Future Work

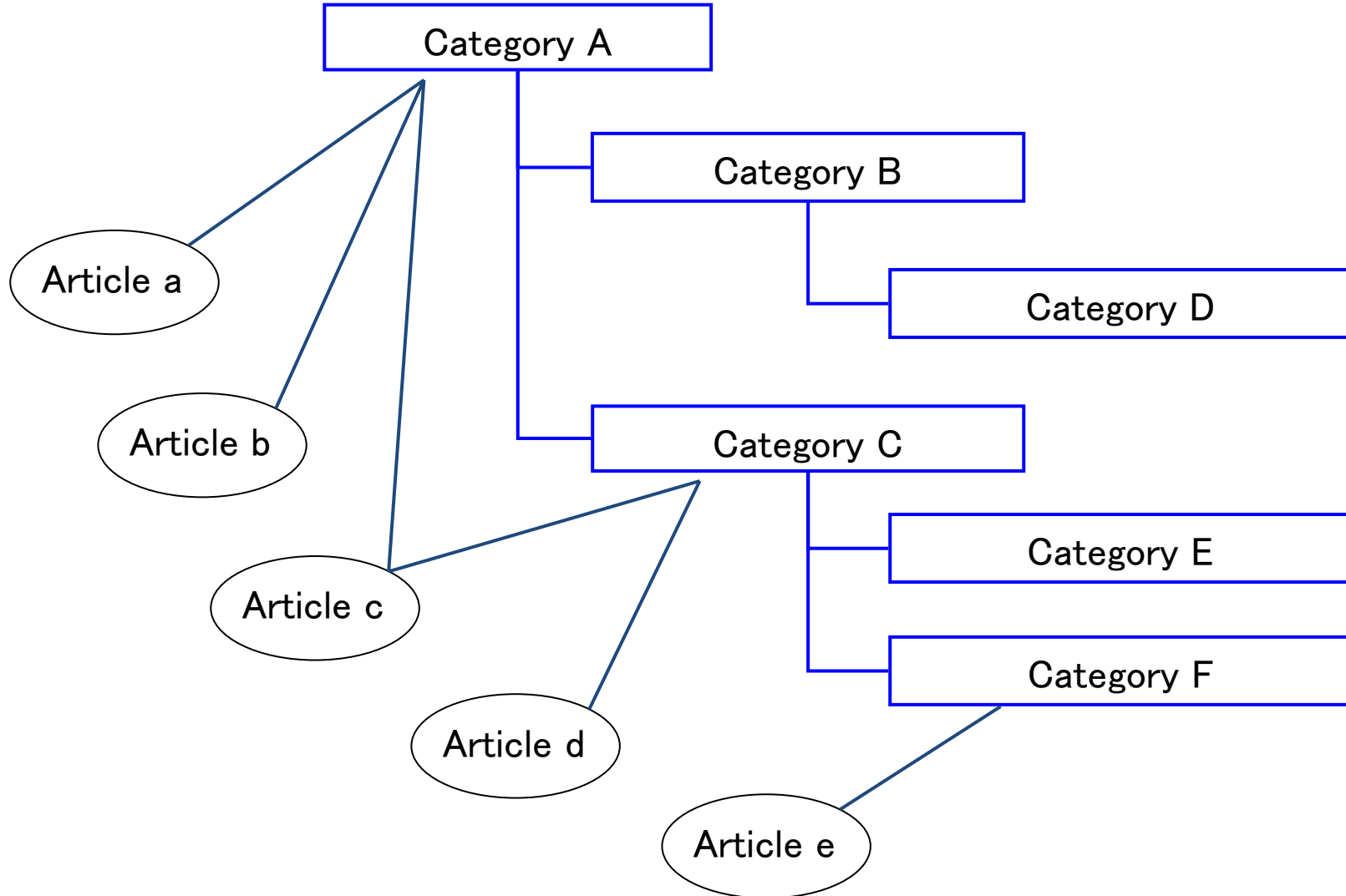
Proposed Methods



Extracting Is-a Relationships

- String Matching Methods on the category names
- Matching Infobox Templates and Categories

Wikipedia Category Tree



Category and Category Tree

<http://ja.wikipedia.org/wiki/Category:プログラミング言語>

Category「Programming Language」

主要カテゴリ > 技術 > 情報 > 情報技術 > コンピュータ > プログラミング > プログラミング言語

主要カテゴリ > 文化 > 言語 > 人工言語 > コンピュータ言語 > プログラミング言語

プログラミング言語に関するカテゴリ。

関連項目 [編集]

■ Category:ソフトウェア開発ツール - 開発ツールの記事についてはこちらを参照

(前の200件) (次の200件)

サブカテゴリ

以下にこのカテゴリ

B

■ [+] BASIC (0)

C

■ [+] C言語 (1)

■ [+] C++ (0)

J

■ [+] Java (1)

L

■ [+] LISP (0)

P

■ [+] PHP (1)

■ [+] PostScript (0)

P の続き

■ [+] Python (0)

■ [+] Pascal (0)

R

■ [+] Ruby (1)

あ

■ [+] アセンブリ言語 (1)

か

■ [+] プログラミング言語 (0)

か

■ [+] 関数型言語 (0)

こ

■ [+] コンパイラ (1)

す

■ [+] 数値解析ソフトウェア (0)

■ [+] スクリプト言語 (5)

な

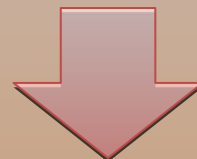
■ [+] 難解プログラミング言語 (0)

ふ

■ [+] プログラミング言語設計者 (0)

■ [+] プログラミング言語の歴史 (2)

Mixing is-a, has-a, class-instance, and other relationships in the category tree



Extracting is-a relationships from category tree

The number of categories: 91,316

[−] 主要カテゴリ

[+] 総記

[+] 学問

[−] 技術

[+] 各国の科学技術

[+] 技術と産業のポータル

[+] 安全技術

[+] 医療

[+] 印刷

[+] インタフェース

[+] 技術のウェブサイト

[+] 宇宙開発

[+] 快適技術

[+] 科学技術雑誌

[+] 科学技術の賞

[+] 科学技術博物館

[+] 架空の技術

[+] 加工

[+] 環境技術

[+] 技術革新

[+] 技術系教育機関

[+] 技術史

[+] 技術者

[+] 技能者

[+] 教育の手法

[+] 軍事技術

[+] 原子力

[+] 建設

[+] 工学

[+] 工業規格

[+] 構造物

[+] 交通

[+] 誤動作

[−] コンピュータ

[+] コンピュータの一覧

[+] コンピュータ関連のスター項目

[+] 計算機科学

[+] 計算機科学界における議論と対立

[+] コンピュータキデクチャ

[+] コンピュータ関連人物

[+] コンピュータ企業

[−] コンピュータ言語

[+] 数式処理システム

[+] スタイルシート言語

[+] 問い合わせ言語

[+] 統一モデリング言語

[+] ハードウェア記述言語

[+] プログラミング言語

[+] マークアップ言語

コンピュータ言語

<http://ja.wikipedia.org/wiki/Wikipedia:カテゴリ>

String Matching Methods on the category names

■ Backward String Matching

Super Class	Sub Class
Motorway	Japanese Motorway
High-speed rail	Taiwan High-speed rail
Seafood	Japanese Seafood
Sumo	Amateur Sumo
Junior college	Japanese Junior college
Short film	Disney Short film

Air

rt

s-a

Japane

Airport

7,971 Is-a relationships

Category Tree

Is-a relationship

Total: 12,558

■ Forward Matched String Eliminating

Super Class	Sub Class
Noodle	Yakisoba
Bird	Domestic duck
Bird	Penguin
Musician	Lyricist
Musician	Composer
Media	Music
Media	Newspaper
Author	Poetry

Athlete

Is-a

Golfer

Is-a relationship

4,587 Is-a relationships

Matching Infobox Template Name to Category Name

Super Class	Sub Class	Super Class	Sub Class
Organic compound	Amine	Software	Free Software
Organic compound	Ester	Software	Image Processing Software
Organic compound	Carboxylic acid	Software	Text Editor
Organic compound	Terpenoid	Software	Web Browser
Organic compound	Organonitrogen compounds	Software	E-mail Software
Organic compound	Aromaticity	Software	Security Software
Organic compound	Heterocyclic compound	Software	Word processor
Organic compound	Amide	Software	CAD Software
Organic compound	Amino acid	Software	System Software
Organic compound	Alkaloid	Software	Windows Game Software
Organic compound	Alcohol	Software	Application Software
Organic compound	Aldehyde	Software	TeX

Categories that the piano article belongs

Keyboard instrument | Piano

Is-a relationships:
3,782

Extracting Class-Instance Relationships

List of People from Tokyo

文化人 [編集]

学者 [編集]

数学者 [編集]

- 小平邦彦
- 秋葉忠利
- 秋山仁
- 浦太郎
- 安島直門:新庄藩士
- 伊原康隆
- 弥永昌吉

物理学者 [編集]

- 三田一郎(素粒子理論)
- 石原純(理論物理学)
- 菊池正士

天文学者 [編集]

- 石田五郎:台東区
- 大沢清輝(観測天体物理学)
- 岡崎彰
- 小川清彦(古天文学・暦)
- 小尾信彌(天体物理学)
- 加藤正二(理論天体物理学)

地球惑星科学者 [編集]

- 井田茂(惑星科学)
- 上田誠也(地球物理学)
- 坪井誠太郎(地質学・鉱山学)
- 坪井忠二(地球物理学)

生物学者 [編集]

- 石塚喜明(土壌肥科学)
- 江上不二夫(生化学)
- 木原均(細胞遺伝学)
- 小宮輝之(恩賜上野動物園園長):千代田区
- 坪井正五郎(自然人類学)
- 日高敏隆(動物行動学):渋谷区
- 箕作佳吉(動物学)
- 茂木健一郎(脳科学)

People from Tokyo

= **Mathmatician** =

Class	Instance
Keio University People	Yukichi Fukuzawa
Keio University People	Atsushi Seike
Keio University People	Yuichiro Anzai
Japanese Tourist attraction	Kyoto
Japanese Tourist attraction	Hiroshima
Japanese Tourist attraction	Akihabara
Japanese Tourist attraction	Yokohama
Nobel laureates	Hideki Yukawa
Nobel laureates	Masatoshi Koshiba
Nobel laureates	Yoichiro Nambu

Title → **Class**

Item → **Instance**

Item

Class-Instance Relationships: 421,989

Procedure

explanatory text of the list.
per listing pages.

(4) Eliminate '*' lines which belong to unconcerned content index like "recital".

(5) Eliminate '*' lines which aren't correct as instances like "*REDIRECT".

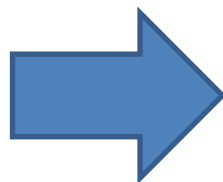
(6) Eliminate '*' lines which are used to describe year like "*19th".

(7) Scrape the string of instance out of each '*' lines by using symbol of link "[[]]" to identify it.

Extracting RDF Triples

東京都 とうきょうと	
地方	関東地方
団体コード	13000-1
ISO 3166-2:JP	JP-13
面積	2,187.65km ² ※境界未定部分あり
総人口	12,988,797人 (推計人口、2009年10月1日)
人口密度	5,940人/km ²
隣接都道府県	神奈川県、山梨県、埼玉県、千葉県
都の木	イチョウ
都の花	ソメイヨシノ
都の鳥	ユリカモメ
他のシンボル	都の歌:東京市歌・東京都歌 都民の日:10月1日
知事	石原慎太郎
所在地	東京都庁 〒163-8001 東京都 新宿区西新宿二丁目8番1号
電話番号	03-5321-1111
Predicate	Object

Subject



Subject	Predicate	Object
Tokyo	region	Kanto
Tokyo	area	2,187.65
Tokyo	population	12,988,797
Tokyo	density	5,940
Tokyo	tree	Ginkgo tree
Tokyo	flower	Somei-Yoshino
Tokyo	bird	Black-headed Gull
Tokyo	governor	Shintaro Ishikawa

.....

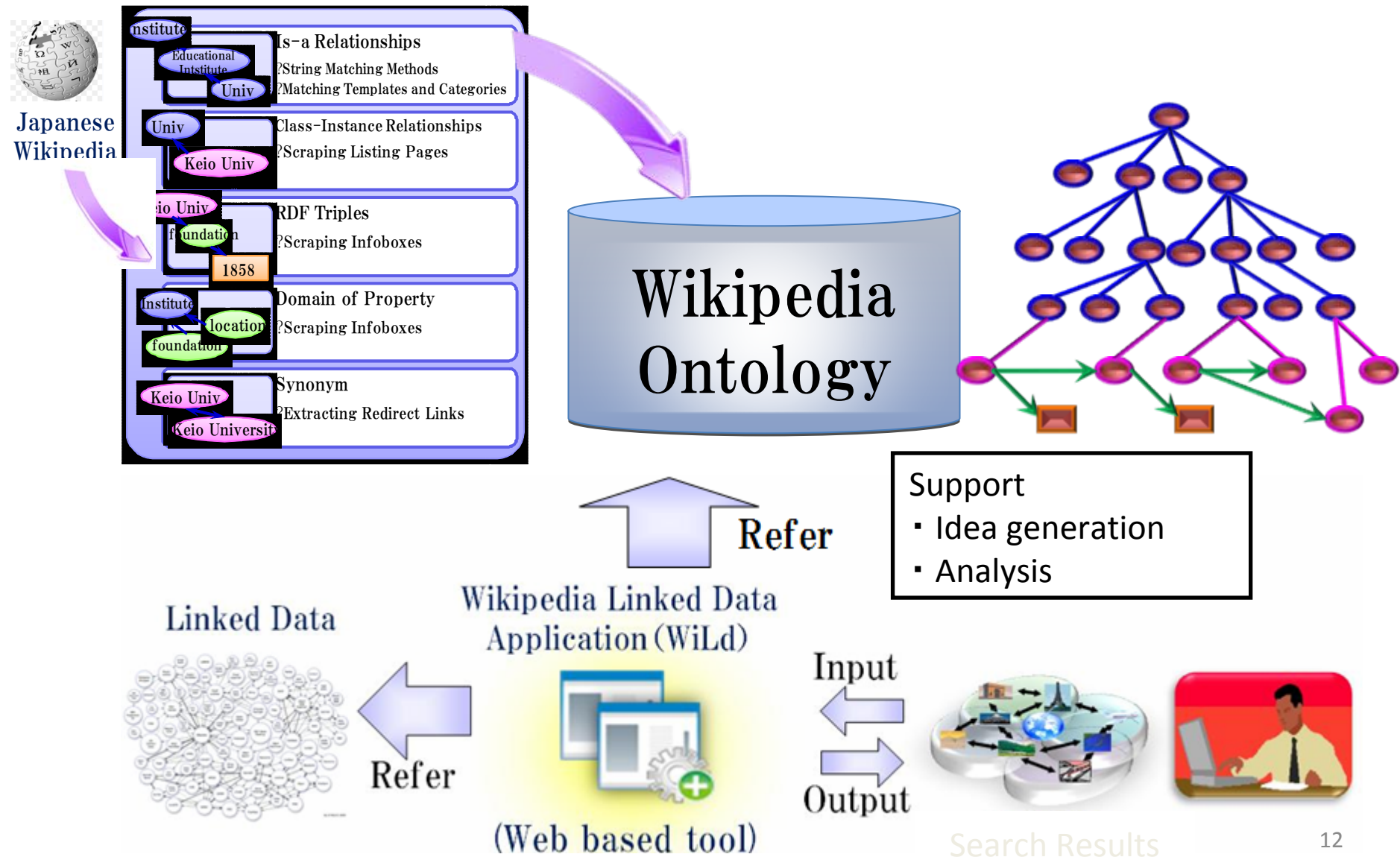
1,485,751 Triples



ウオッチャー Google Map Live Search geohack
mapfan yahoo navitime ゼンリン ちず丸

ウィキポータル [日本の都道府県/東京都](#)

Implementation of Wikipedia Ontology Search Application



Demo1 WiLD (Wikipedia Linked Data Application) (think about Japanese famous novelist) 1min.

WILD - Wikipedia Linked Data Application - (version 0.5.0.build20090201)

多摩川

検索

東京都のご当地ソング

社会

歴史

施設

地域別の川

日本の一級河川

リスト(多摩川)

地域別の川

東京都のご当地ソング

日本の一級河川

河川

地理

環境

河川

土木施設

日本の一級河川

多摩川

地域別の川

東京都のご当地ソング

インスタンス表示

max degrees of separation

max number visible

item spacing

Auto-fit

インスタンスリスト

大松川 (徳島県)

ドナウ川

ドゥー川

新利根川

七瀬川 (大分県)

奥入瀬川

高橋川 (兵庫県)

江川 (徳島県)

アングラナ川

内川 (石川県)

堤川

中の沢川

万之瀬川

タグダ川

アルダン川

作業リスト

多摩川

利根川

属性リスト

流域

平均流量

水源の標高

流域面積

延長

データリンク

チャート表示

DB Pedia

Tama_River

Wikipedia

多摩川

InfoBox

Data

平均流量

36.6_m³/s

延長

138_km

流域

山梨県・東京都・神奈川県

種別

一級河川

水源の標高

1,953_m

水系

一級水系_多摩川

流域面積

1,240_km²

河口

東京湾

水源

笠取山

BOOK

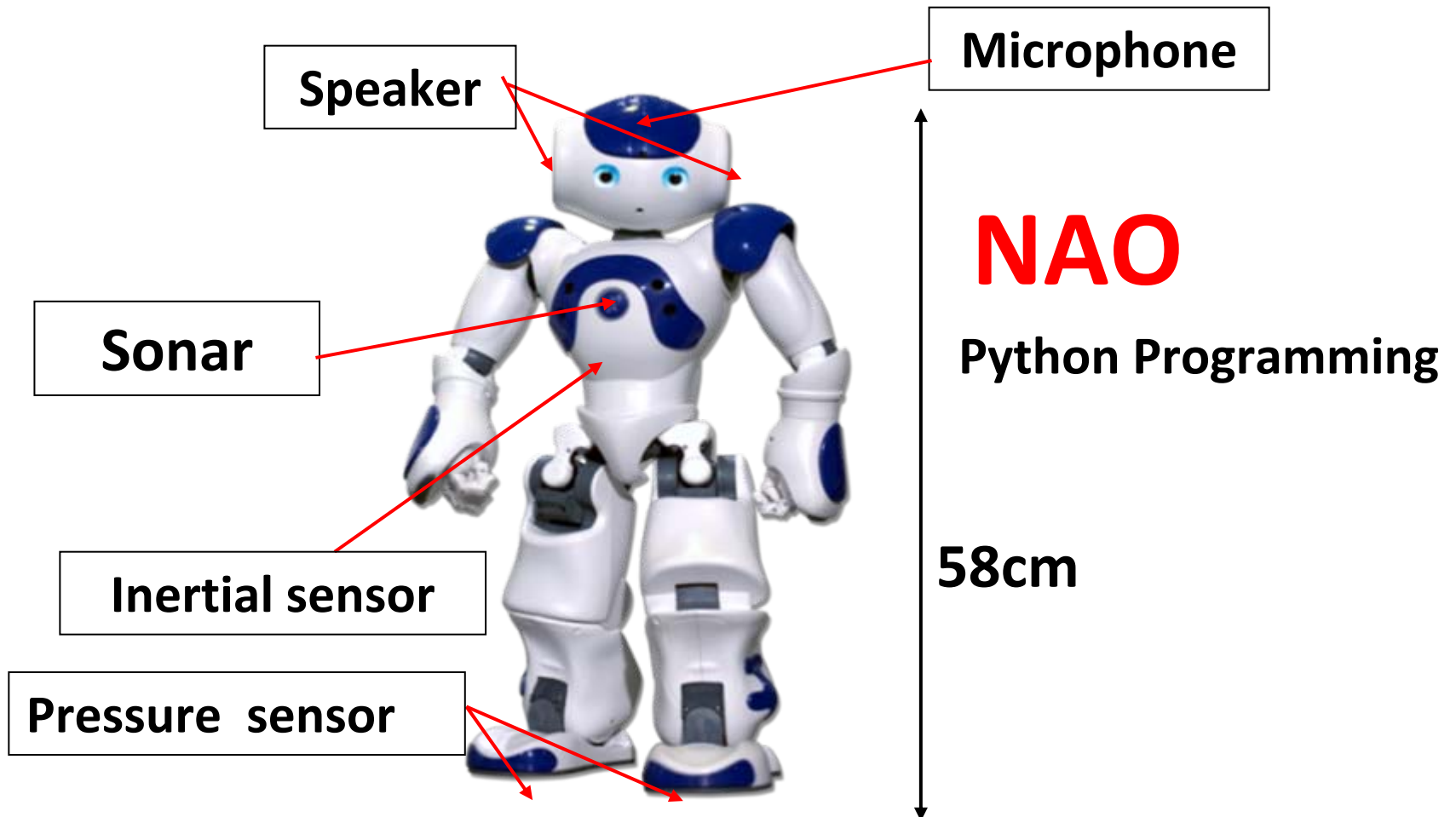
Name

絞り込み

Linked Data

- Book
- Restaurant

Demo2 HRI (Human Robot Interaction) (1)



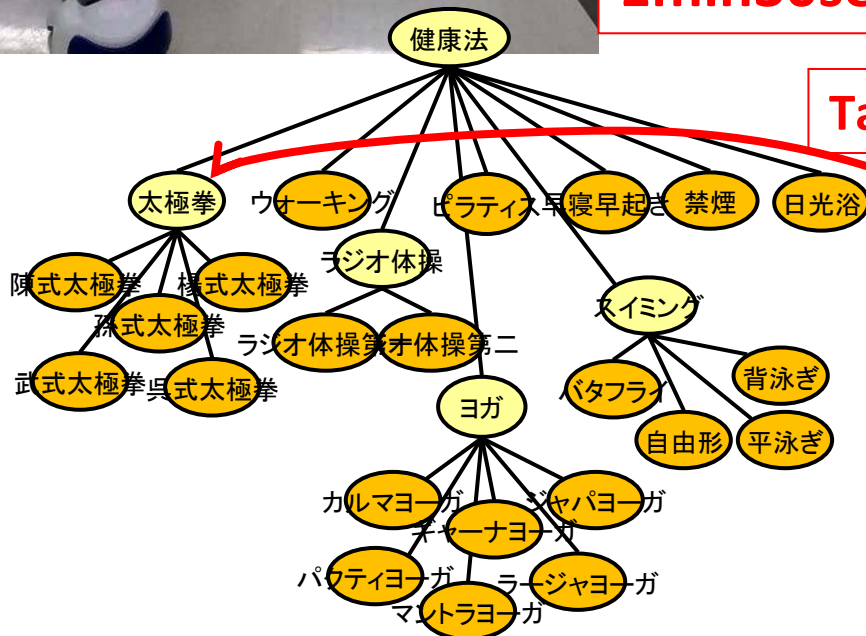
NAO comes from Aldebaran in France.
<http://www.aldebaran-robotics.com/en>

Demo2 HRI (Human Robot Interaction) (2)

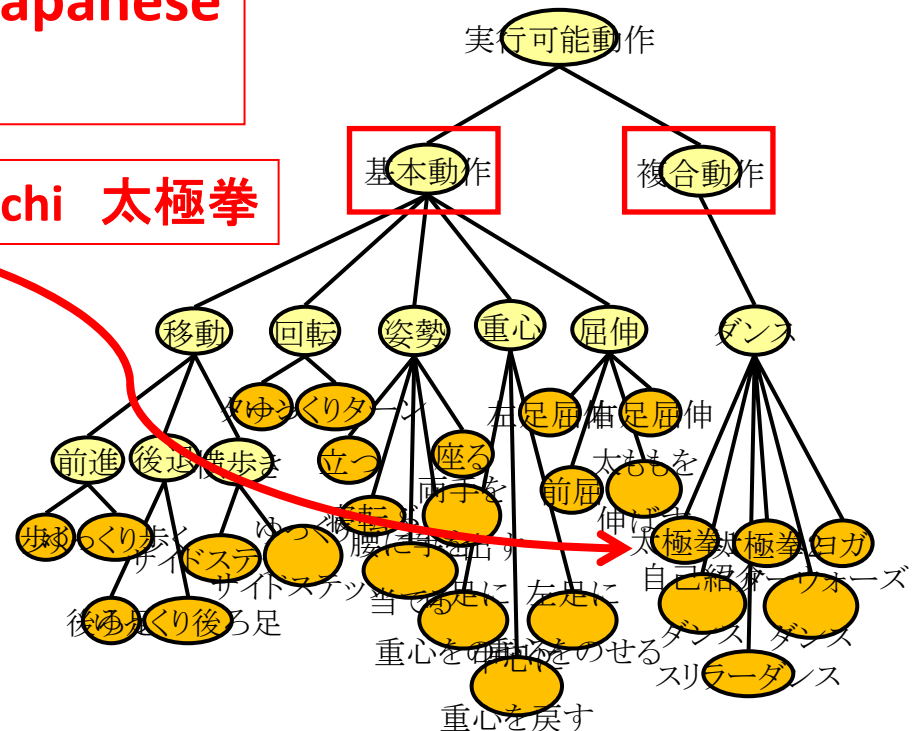


1. An user asks Nao the ways for health-care.
2. Using **WikipediaJapn ontology**, Nao enumerates them.
3. The user selects tai-chi from them.
4. Using **Action ontology** with Nao,
Nao shows the user tai-chi actions that Nao can do.
5. The user selects tai-chi_1 from them.
6. Nao does the action of tai-chi_1.

**Demo: by Japanese
1min30sec**



Tai-chi 太極拳



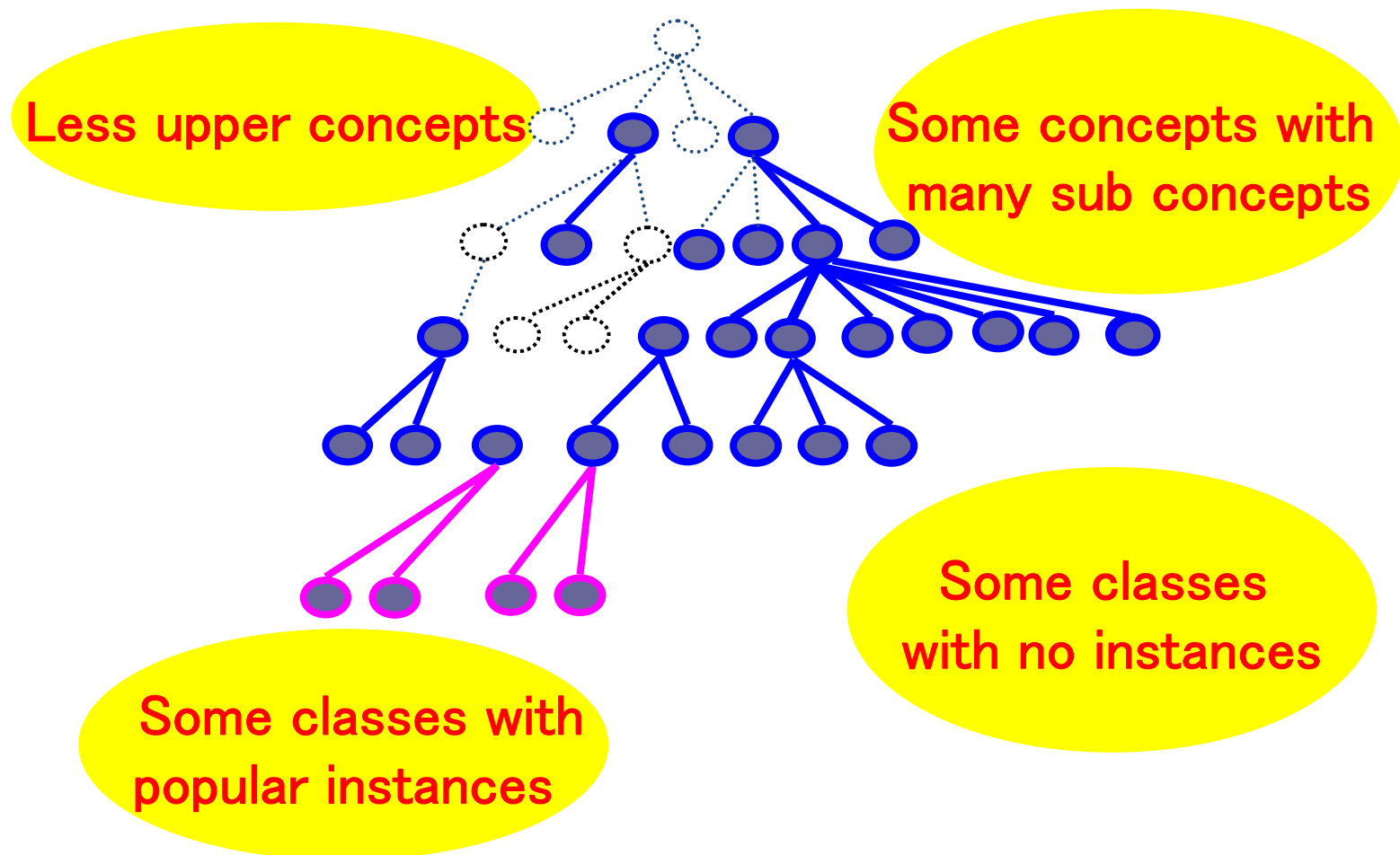
WikipediaJapn ontology

Action ontology

Learning Results from Wikipedia Japan (1)

	Relations	#	Precision
IS-A	AI IS-A	93,322	76.30%
	By string matching	12,558	93.1±1.51%
	By template matching	3,782	95.6±1.09%
	By contents headlines	83,288	72.6±2.74%
Class - Instance		421,989	97.2±1.02%
RDF triple		1,485,751	95.8±1.79%
Property Domains		6,485	95.4±1.22%
Synonyms		106,671	67.0±2.90%
Total		1,834,753	-

How is WJO going ?



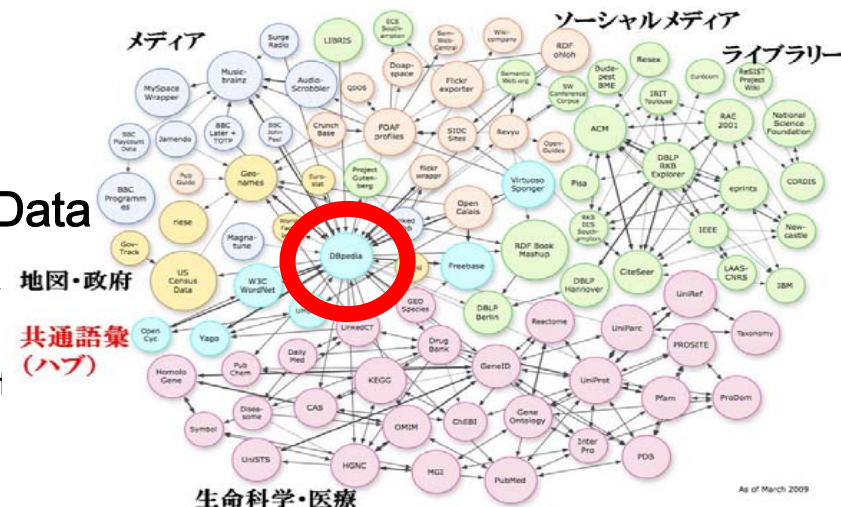
Related Work

		IS-A	Property		Instance	
			Property	Property domain	Class-Instance	RDF triple
Bizer	DBpedia		○			○
Suchanek	YAGO		○	△	○	△
Ponzetto	—	○				
Keio Univ.	Wikipedia Japan2 Ontology	○	○	○	○	○

DBpedia - A Crystallization Point for the Web of Data

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer
 Christian Becker, Richard Cyganiak Sebastian Hellmann
 Journal of Web Semantics: Science, Services and Agents on the
 World Wide Web, Issue 7, Pages 154–165, 2009.

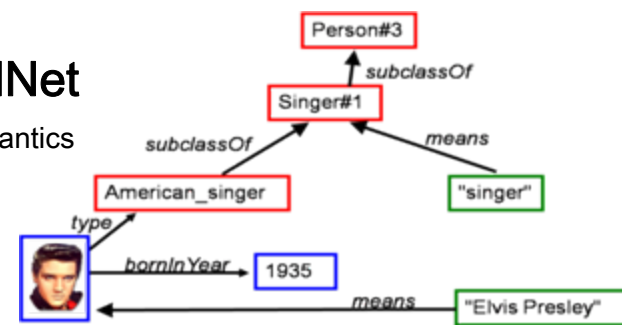
<http://wiki.dbpedia.org/>



YAGO - A Large Ontology from Wikipedia and WordNet

Fabian M Suchanek, Gjergji Kasneci, Gerhard Weikum Elsevier Journal of Web Semantics

<http://www.mpi-inf.mpg.de/yago-naga/yago/>



Conclusions and Future work

Conclusions

- Wikipedia Japan works for light-weight (not heavy-weight) ontology development.
- Wikipedia Japan Ontology works for knowledge creation support and HRI.

Future Work

Manage the right issues of Wikipedia Japan Ontology by with Upper Ontologies.

