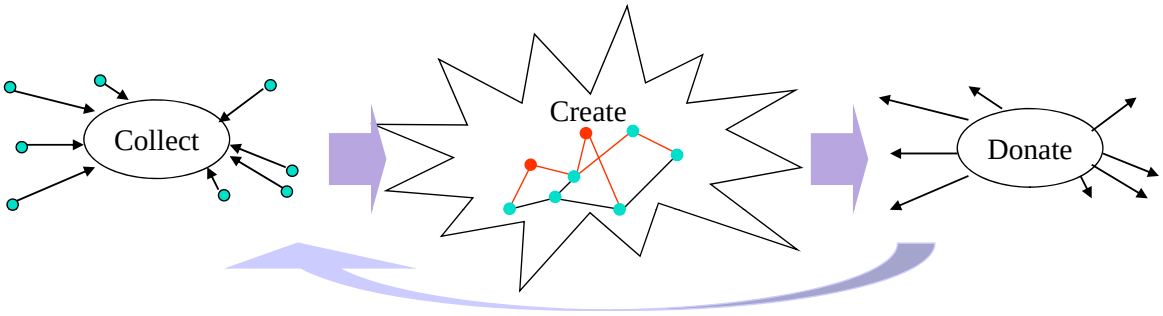
Wikipediaと研究コミュニティ

武田 英明

国立情報学研究所
東京大学　人工物工学研究センター
takeda@nii.ac.jp

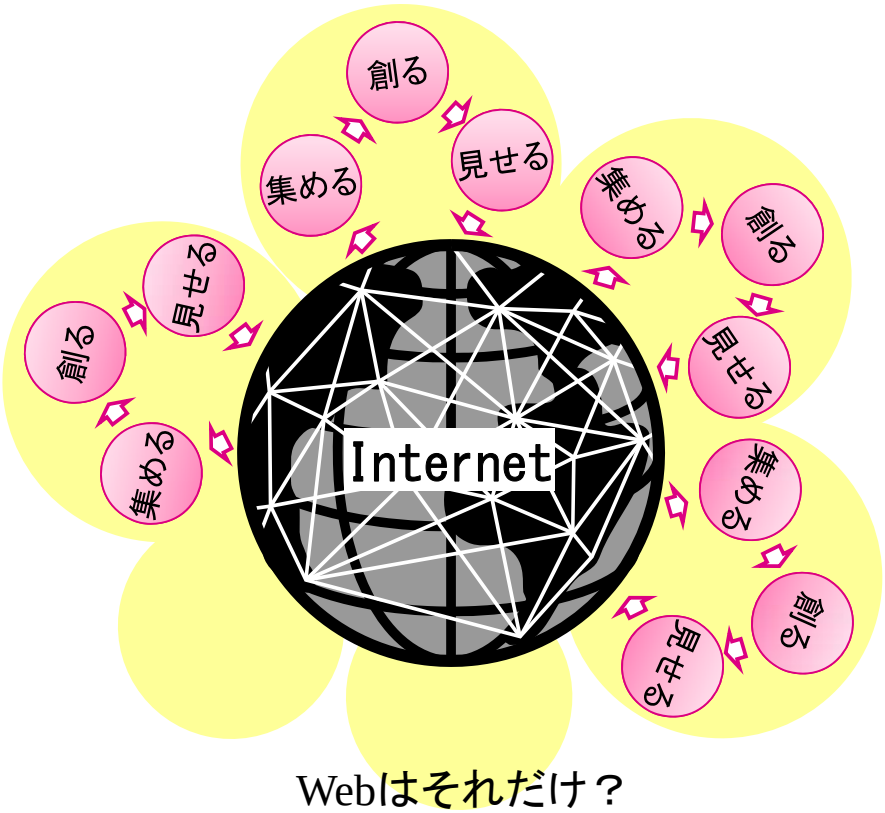
ソーシャルメディアとしてのWeb

情報を創るとは...



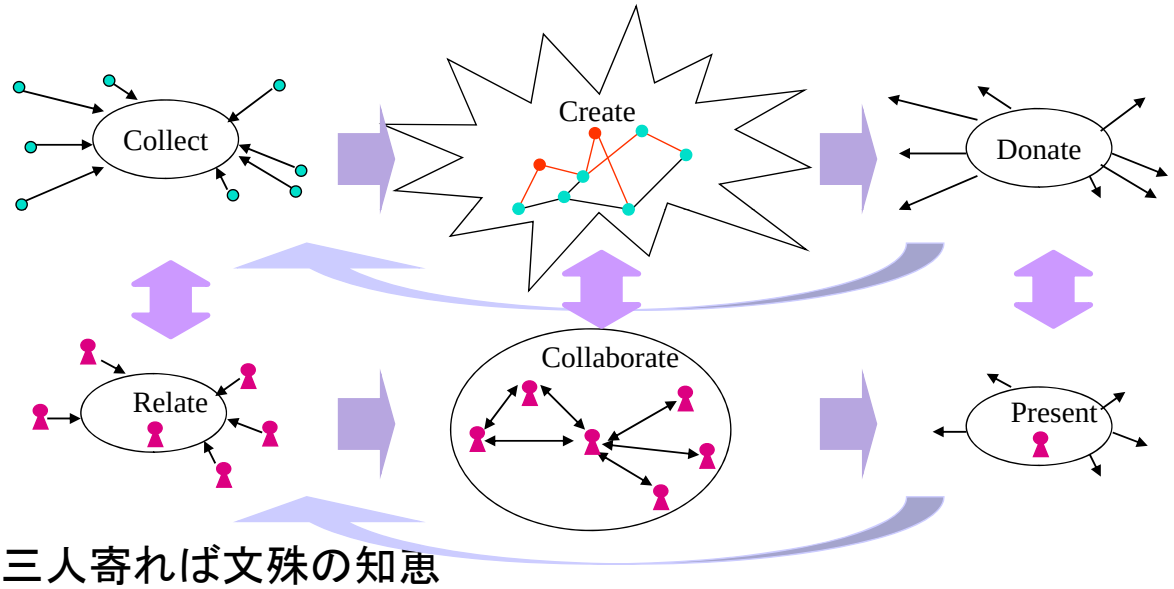
- 創造は無から生じない
- 他者の仕事を知り、理解する
- 他者へ自らの仕事をみせていく
- このサイクルが古今東西普遍のこと、ただし限定されてた
  - 人、速さ、量、範囲
- インターネット技術、ことにWeb技術はこれを解放した
  - 人、速さ、量、範囲

情報活動としてのInternet



Webはそれだけ？

情報だけあるだけでいいのか…



- 三人寄れば文殊の知恵
- 情報を知ることは人を知ること、またその逆
- 情報を見せることは自らをみせること、またその逆
- 人のネットワークを通じた情報流通

ソーシャルメディア

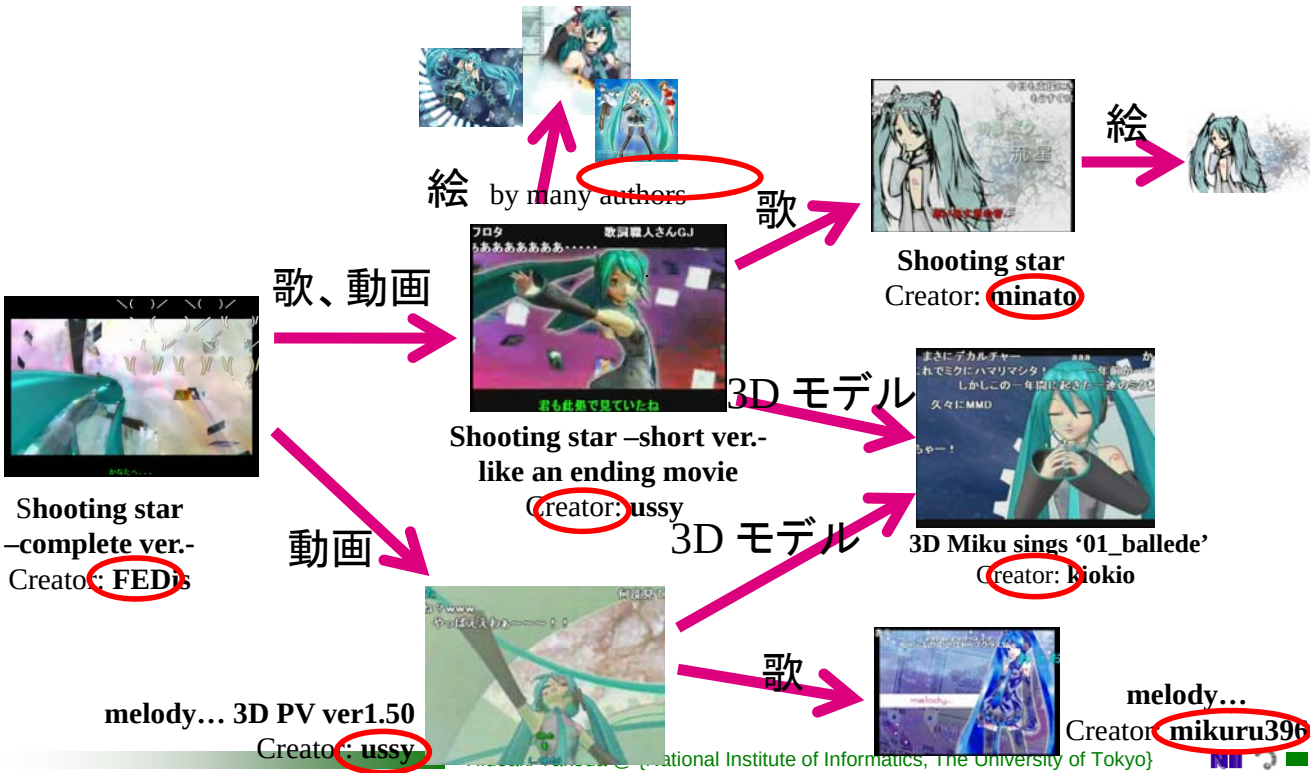
- 社会的に広がりのある参加者によって構成され、参加者間のコミュニケーションによって成り立っているメディア

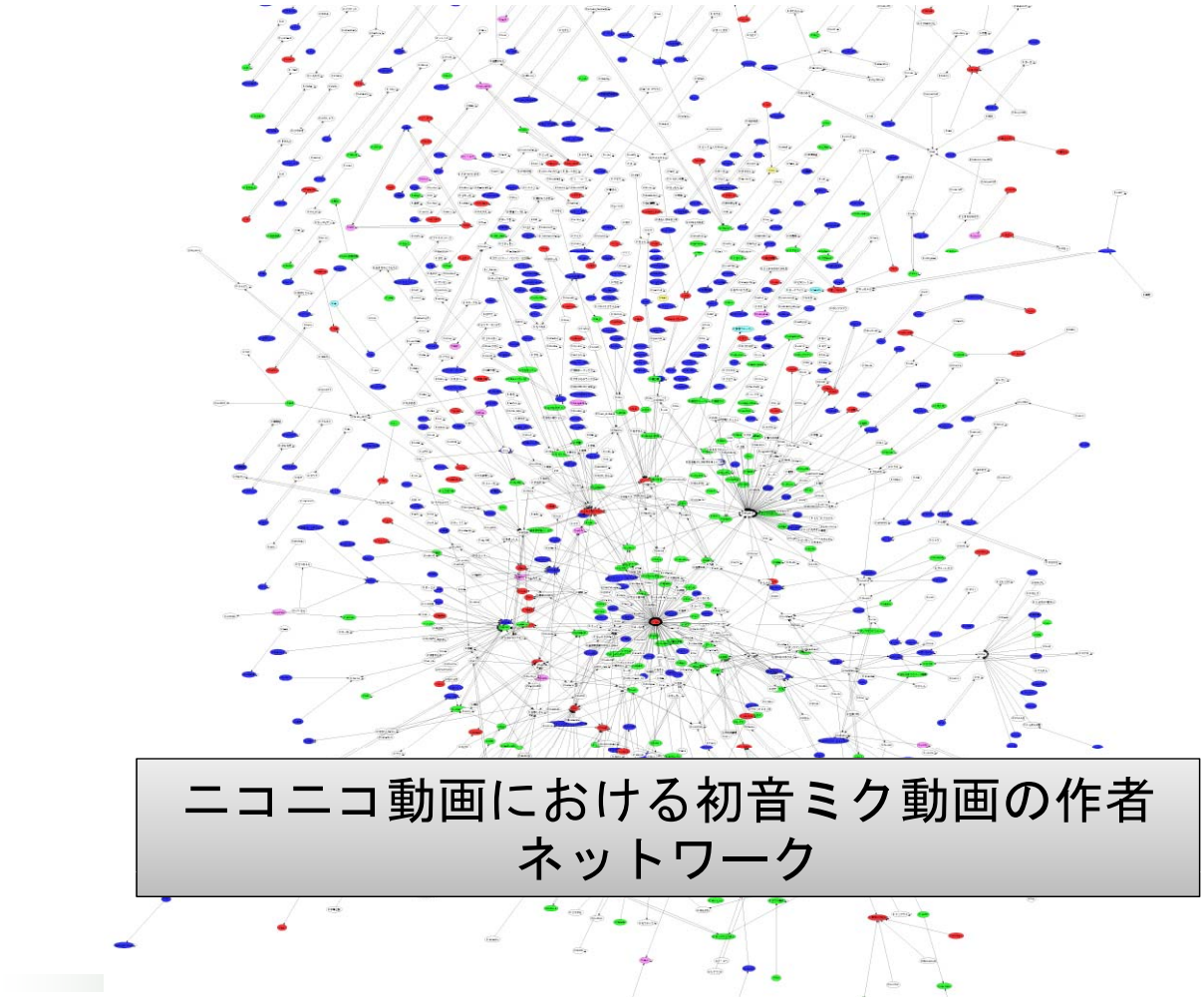
- 例
  - 既存マスメディア（テレビ、新聞等） … No
  - Web…一般にはNo
  - 掲示板 … Yes
  - ブログ … 部分的にYes
  - SNS … Yes
  - ソーシャルタギングサービス … Yes
  -

大規模共創
Massively Collaborative Creation

- 新しいコンテンツを生み出すようなソーシャルメディア
- 例
  - 掲示板
  - Q&Aサイト（Yahoo!知恵袋、教えてGoo、Yahoo!Answers..)
  - Wikipedia
  - ニコニコ動画
    - Cf. Youtube
- 特徴
  - 大規模な参加
  - なんらかのコンテンツを生み出す
  - 参加者の相互作用がコンテンツ作成に影響を与える
    - 相互作用＝コンテンツ（掲示板、Q&Aサイト）
    - 相互作用がコンテンツに影響(Wikipedia, ニコニコ動画）

ニコニコ動画の初音ミク動画の共創プロセス





ニコニコ動画における初音ミク動画の作者ネットワーク

ソーシャルメディアとしてのWikipedia

- どんな特徴があるのか
  - 大規模性
  - 網羅性
  - 共同性
  - データ入手性

Wikipediaと研究

- 研究者が（研究対象として）Wikipediaにどう向かい合うか
  - 分析対象としてのWikipedia：“Wikipedia現象”の分析
  - データとしてのWikipedia：Wikipediaデータの利用
  - Wikipediaの支援や活用

分析対象としてのWikipedia：“Wikipedia現象”の分析

- “Wikipedia現象”の分析：
  - なぜWikipediaはこれだけ大きくなったのか。
  - なぜ書いているのか
  - 誰が書いているのか
  - なにが書かれているのか
  - ．．．

分析対象としてのWikipedia：“Wikipedia現象”の分析

- 研究の視点：コンテンツ作成プロセスの分析
  - 合意形成プロセス
    - Wikipediaの各ページは複数の人の貢献によって作られている。このような面識のない人々がいかに協力して一つのコンテンツをつくっていくのか。
  - 集団性、社会性
    - Wikipediaはきわめて多数の人が能動的に参加して成り立っている。この人々が作る社会はどのようなものか。

編集プロセスに注目

- Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination, Aniket Kittur and Robert E. Kraut, Proc. of CSCW 2008
  - ページのクオリティは編集者の数より、編集者のgini係数に相関

		Mean	Median	Std	Quality change	Initial Quality	Article Age	# Article Editors	Editor Concen-tration	# Talk Editors
1	Quality change	.09	.00	.55						
2	Initial Quality	2.36	2.00	1.14	-.20					
3	Article Age	25.90	21.73	17.41	.00	.29				
4	# Article Editors	48.31	11.00	108.73	.08	.43	.51			
5	Editor Concentration	.26	.25	.18	.20	.27	.21	.61		
6	# Talk Editors	6.00	2.00	16.28	.14	.47	.41	.78	.52	

Table 2. Descriptive statistics before log transformation and correlations after log transformation. Quality ranges between 1 (Stub) and 6 (Featured Article). Editor concentration is measured by the gini coefficient, which ranges from 0 (equally distributed) to 1 (highly concentrated).

編集プロセスに注目

- マスコラボレーションにおけるコンテンツ形成プロセスの分析：伊藤諭志, 伊藤貴一, 熊坂賢次, 井庭崇、第20回セマンティックWebとオントロジー研究会、2009
- ページごとの編集者の関係ネットワークを作り分析

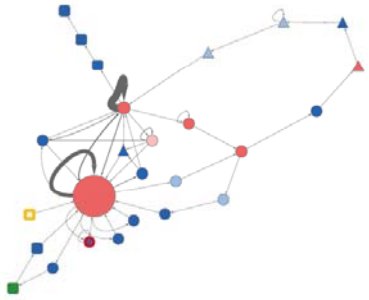


図 7: 秀逸な記事「雨氷」のコラボレーション・ネットワーク

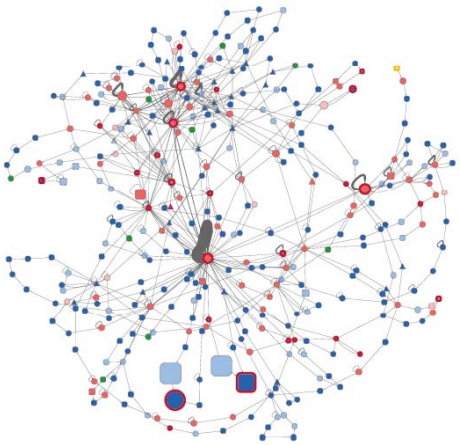


図 9: 秀逸な記事「キリスト教」のコラボレーション・ネットワーク

データとしてのWikipedia：Wikipediaデータの利用

- １．知識の集合として
  - Wikipediaは大量かつ広範な分野に関する比較的均質な知識源。
- ２．（多）言語の集合として
  - Wikipediaのコンテンツは大量かつき広範な分野に関する比較的均質な文章の集合。ここからシソーラスや言語に関する統計的情報などを得ることが期待できる。また、Wikipediaは各国語版があるので、そこから言語間の関係の分析をすることもできる。
- ３．構造化文書の集合として
  - Wikipediaは大量かつ均質なリンクつきの文書の集合。このリンクを分析することで背景的情報、暗黙の情報を得ることが期待できる。

データとしてのWikipedia：1. 知識集合として

- 知識の集合として
  - Wikipediaは大量かつ広範な分野に関する比較的均質な知識源。
    - どんな知識が含まれているのか（分析、発見）
    - この知識を使おう（利用）
- どんな知識源とみるか
  - 構造的知識、オントロジーの抽出
  - 常識、日常知識の抽出と利用
  - 意外な知識の発見

データとしてのWikipedia：1. 知識集合として

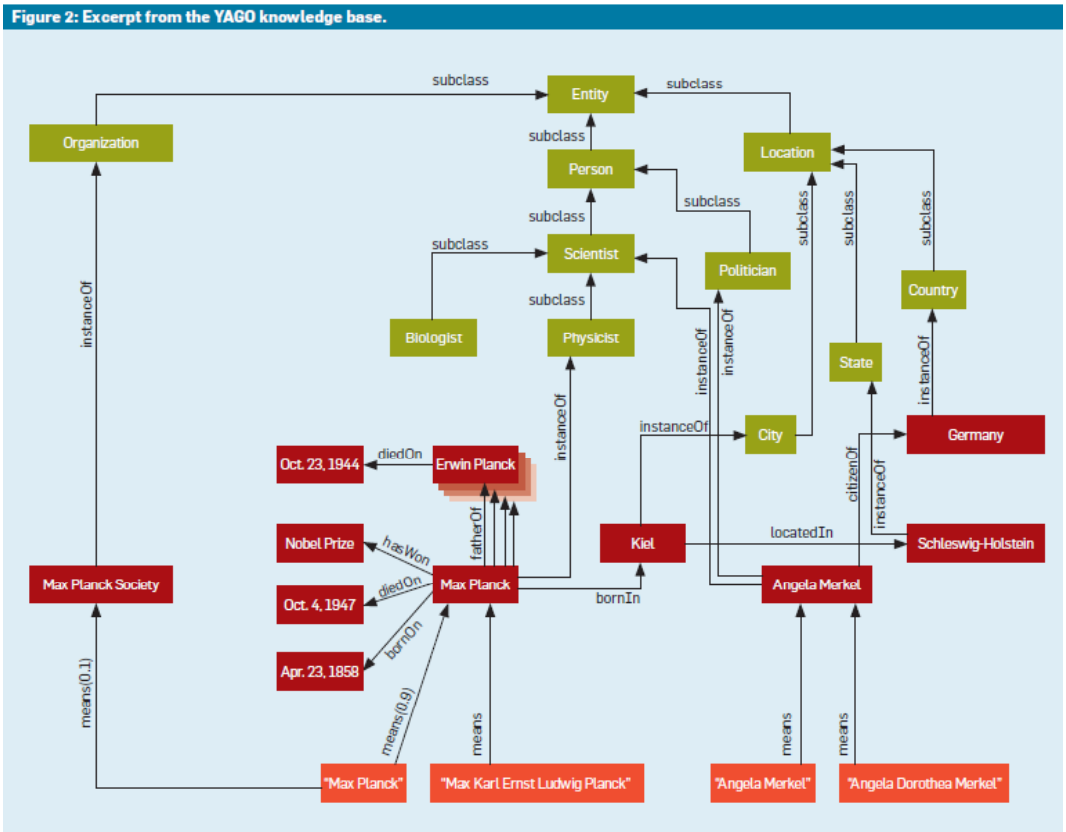
- どんな知識源とみるか
  - 構造的知識、オントロジーの抽出
    - 中山氏：「MediaWikiと構造化知識の抽出」
    - 山口氏：「Wikipediaとオントロジー」
    - Yago, Dbpedia
  - 常識、日常知識の抽出と利用
  - 意外な知識の発見

Hideaki Takeda @ {National Institute of Informatics, The University of Tokyo}

Yago: Yet Another Great Ontology

- WikipediaとWordNetから巨大な知識ベース（オントロジー）
- 手法
  - インスタンス：Wikipediaから
  - インスタンスークラス関係、インスタンス間関係：Wikipediaのカテゴリーの構文解析して生成
    - 例：“American people in Japan” -> Personクラス
      “born in 1954” -> BornInYear 1954
  - クラス：葉のクラスはWikipediaのカテゴリの構文解析。それ以外はWordNetのSynset。それらを機械的に結合

Relations	92
Classes	224,391
Individuals (without words and literals)	1,531,588
People	546,308
Locations	230,988
Institutions/companies	57,893
Movies	33,234



Categories: 1809 births | 1865 deaths | Abraham Lincoln | Assassinated United States Presidents | Deaths by firearm in Washington, D.C. | English Americans | Illinois lawyers | Illinois Republicans | Lincoln family | Members of the Illinois House of Representatives | Members of the United States House of Representatives from Illinois | People from Coles County, Illinois | People from LaRue County, Kentucky | People from Macon County, Illinois | People from Spencer County, Indiana | People from Springfield, Illinois | People murdered in Washington, D.C. | People of Illinois in the American Civil War | People of the Black Hawk War | Postmasters | Presidents of the United States | Religious skeptics | Republican Party (United States) presidential nominees | Smallpox survivors | Union political leaders | United States presidential candidates, 1860 | United States presidential candidates, 1864 | United States Whig Party

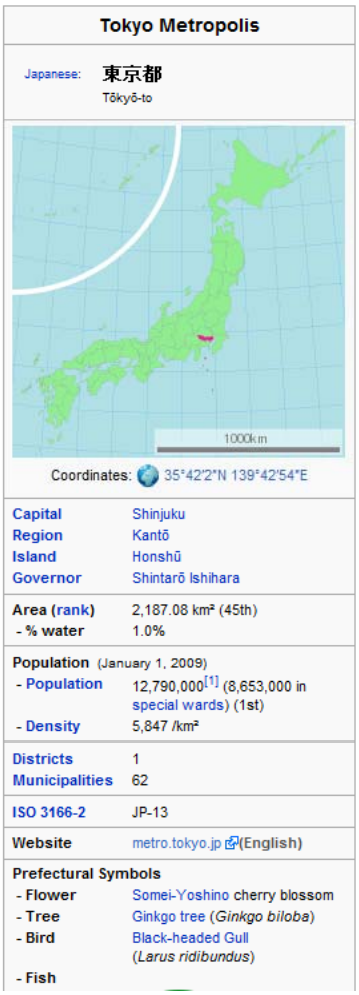
wikic...bolitionists

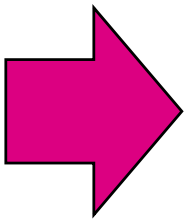
Abraham Lincoln

Type:	individual 21 more
Label:	Линкольн, Авраам 65 more
Alternative label:	16th President of the United States 23 more
[bornIn]:	Hardin County
[bornOnDate]:	1809-02-12
[diedIn]:	DC, The District
[diedOnDate]:	1865-04-15
[hasChild]:	Tad Lincoln 3 more
[hasFamilyName]:	Lincoln
[hasGivenName]:	Abraham
[hasPredecessor]:	James Buchanan
[hasSuccessor]:	Andrew Johnson
[influences]:	Walt Whitman
[isAffiliatedTo]:	Whig Party
[isLeaderOf]:	American Civil War
[isMarriedTo]:	Mary Todd Lincoln
[isNumber]:	16

Dbpedia

- WikipediaのInfoboxを使って知識ベースをつくる
- データとして利用可能にする。
- 手法
  - 上位クラスはYagoから
  - インスタンスの関係はInfoboxから

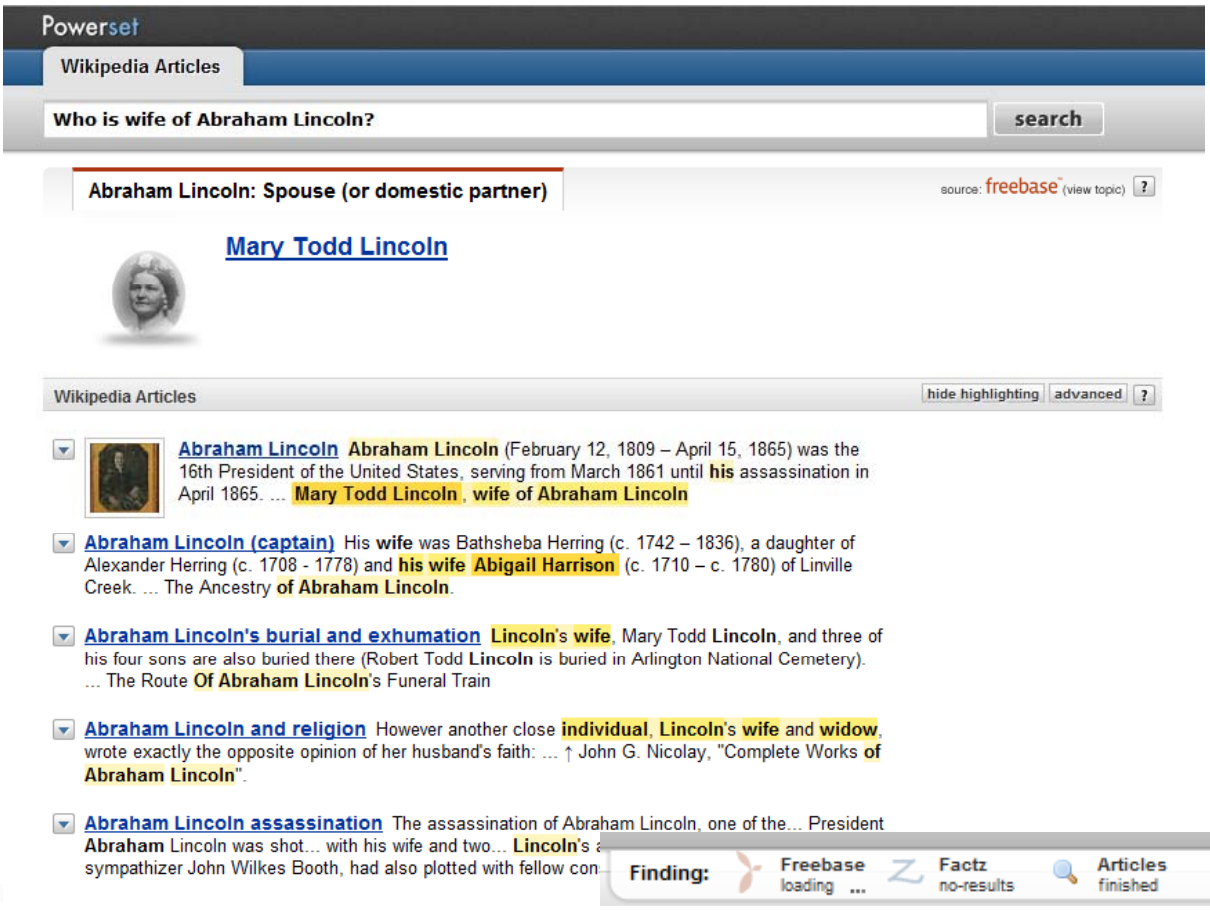




Dataset	Description	Triples
Articles	Descriptions of all 1.95 million concepts within the English Wikipedia including titles, short abstracts, thumbnails and links to the corresponding articles.	7.6M
Ext. Abstracts	Additional, extended English abstracts.	2.1M
Languages	Additional titles, short abstracts and Wikipedia article links in German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish and Norwegian.	5.7M
Lang. Abstracts	Extended abstracts in 13 languages.	1.9M
Infoboxes	Data attributes for concepts that have been extracted from Wikipedia infoboxes.	15.5M
External Links	Links to external web pages about a concept.	1.6M
Article Categories	Links from concepts to categories using SKOS.	5.2M
Categories	Information which concept is a category and how categories are related.	1M
Yago Types	Dataset containing rdf:type Statements for all DBpedia instances using classification from YAGO [16].	1.9 M
Persons	Information about 80,000 persons (date and place of birth etc.) represented using the FOAF vocabulary.	0.5M
Page Links	Internal links between DBpedia instances derived from the internal pagelinks between Wikipedia articles.	62M
RDF Links	Links between DBpedia and Geonames, US Census, Musicbrainz, Project Gutenberg, the DBLP bibliography and the RDF Book Mashup.	180K

データとしてのWikipedia： 1. 知識集合として

- どんな知識源とみるか
  - 構造的知識、オントロジーの抽出
  - 常識、日常知識の抽出と利用
    - 「Wikipediaからの拡張クエリ生成によるWeb検索とその評価」堀憲太郎他、SIG-SWO-A803-13
    - Powerset
  - 意外な知識の発見



データとしてのWikipedia： 1. 知識集合として

- どんな知識源とみるか
  - 構造的知識、オントロジーの抽出
  - 常識、日常知識の抽出と利用
  - 意外な知識の発見
    - 「Wikipediaカテゴリネットワークからの意外性のある関連性の抽出」 野田陽平他、SIG-SWO-A901-04

データとしてのWikipedia：2. (多)言語の集合として

- Wikipediaのコンテンツは大量かつき広範な分野に関する比較的均質な文章の集合。ここからシソーラスや言語に関する統計的情報などを得ることが期待できる。また、Wikipediaは各国語版があるので、そこから言語間の関係の分析をすることもできる。
  - 「リンク共起性解析によるWikipediaからの連想シソーラス構築手法」伊藤雅弘他、SIG-SWO-A803-05
  - 「Wikipediaを利用した音声認識用言語モデルの構築および評価」田中和紀他、 SIG-SWO-A803-11
  - 「Wikipedia概念体系を用いた日本語ブログ空間のトピック分布推定」川場真理子他、 SIG-SWO-A803-12
  - 「Wikipediaを用いた多言語情報アクセスに関する研究：言語間リンクの分析と応用、新井嘉章他、 SIG-SWO-A803-15
  - 「多言語に展開するWikipediaの特徴の比較調査」、森竜也他、 SIG-SWO-A901-05

データとしてのWikipedia：3. 構造化文書として

- Wikipediaは大量かつ均質なリンクつきの文書の集合である。このリンクを分析することで背景的情報、暗黙の情報を得ることが期待できる。
  - 「「縁」 の発見と可視化による理解の支援」
  - 「リンク共起性解析によるWikipediaからの連想シソーラス構築手法」中山浩太郎他
  - 「多型トピックモデルを用いたWikipedia検索」江口浩二他、SIG-SWO-A803-14

Wikipediaの利用

- 情報検索（Wikipedia内／Web）
  - クエリ拡張等
- 情報推薦（Wikipedia内／Web）
  - 関連性推定等
- 信頼性推定（Wikipedia内／Web）
  - ページ信頼度、人物信頼度
- 2次データ生成
  - 事実（ファクト）データ
    - Linked Data化
  - シソーラス
  - 知識ベース
  - オントロジー

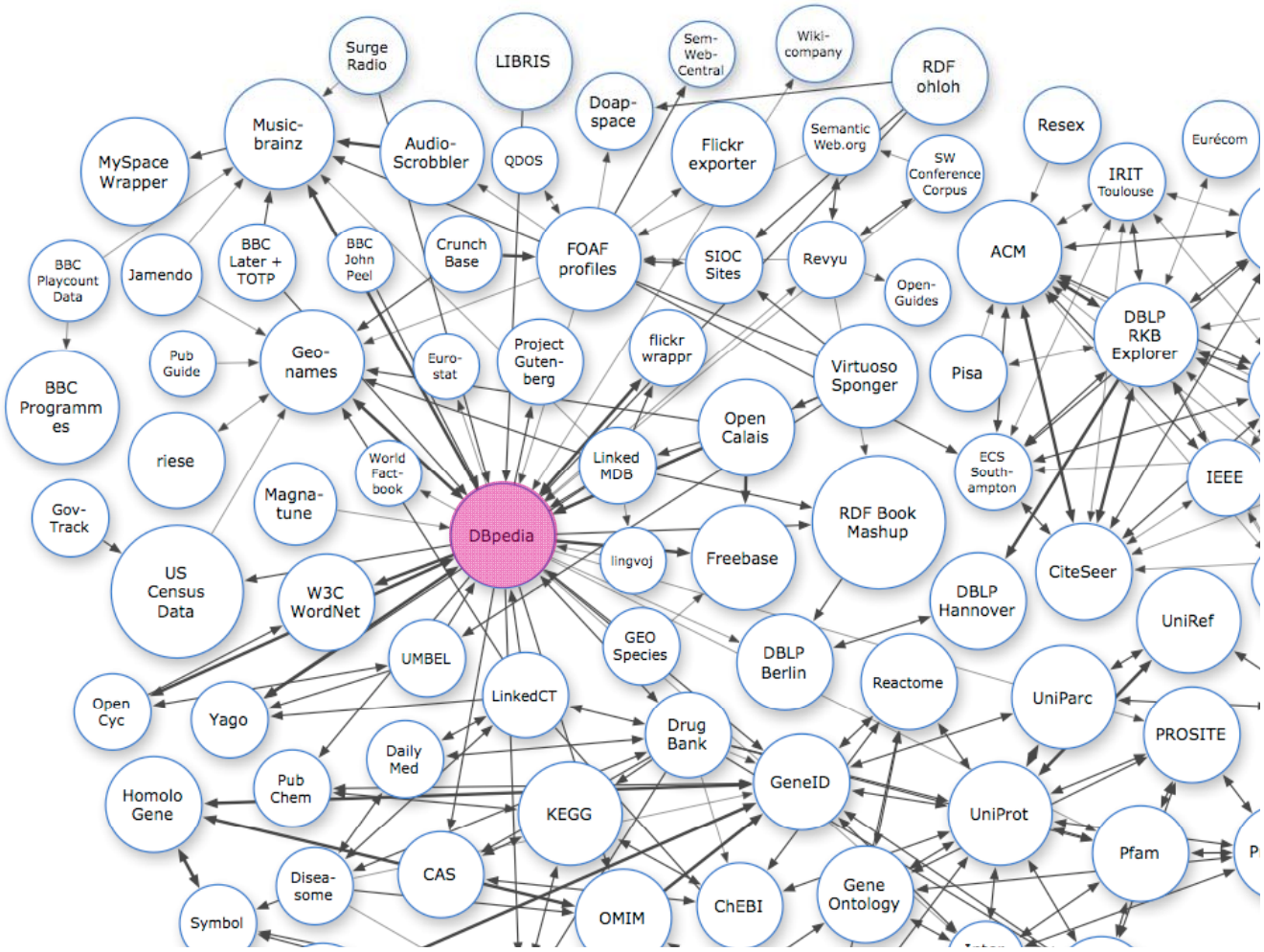
Linked Data

- Linked Dataとは *"Web of Data"*
  - RDFで公開されるデータ
  - 外部から参照可能
- Linked Dataのための4条件
  - 事柄の名前にURIを使うこと
    - *すべてのモノ，コトにURIを！*
  - 名前の参照がHTTP URIでできること
    - *DOIとかいったURNは使わないでね*
  - URIを参照したときに関連情報が手に入るように
    - *理解可能なデータを提供してね.*
  - 外部へのリンクも含めよう
    - *Webのようにリンクでつながるデータを作ろう*

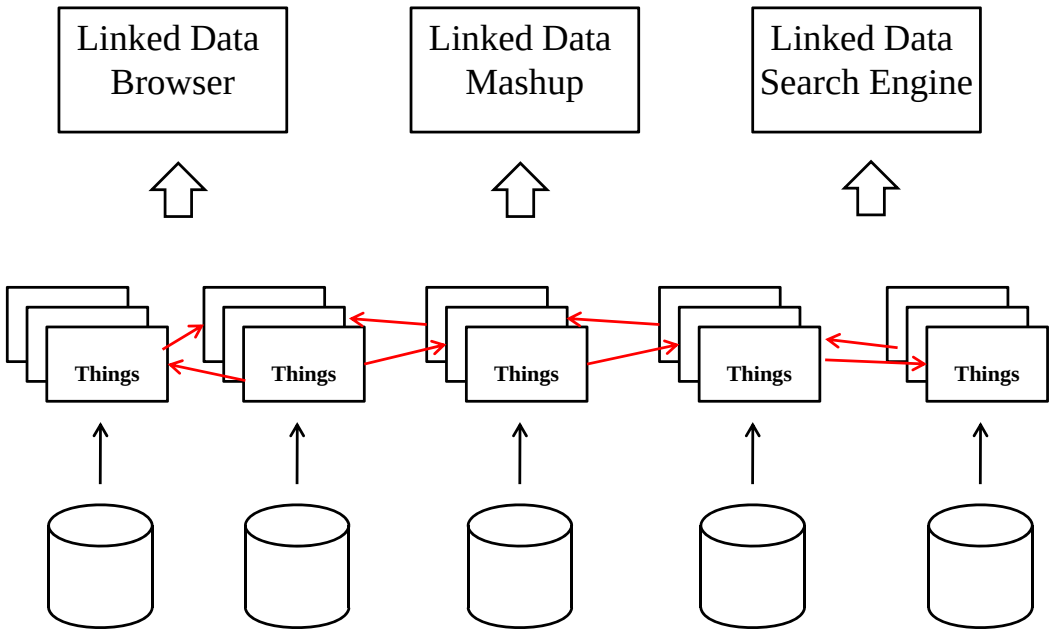
Linked Data, TBL, http://www.w3.org/DesignIssues/LinkedData.html

Linking Open Data (LOD)

- 公開されたLinked Dataを集めるプロジェクト
- 主要なLinked Data
- （データ変換）
  - Dbpedia (Wikipedia) : 百科事典, 2.7億文
  - Geonames：地名と緯度経度, 9300万文
  - MusicBrainz：音楽
  - WordNet：辞書
  - DBLP bibliography：論文の書誌，2800万文
  - US Census Data: 米国国勢調査(2000年), 10億文
- （クロール）
  - FOAF (Friend Of A Friend)：個人と個人関係のプロファイル
- （ラッパー）
  - Flickr Wrapper



How to use Linked Data



DBpedia Mobile

まとめ

- Wikipediaはソーシャルメディア、大規模共創として貴重かつ重要な存在
- Wikipediaは研究者に魅力的なものである
  - なぜ：大規模性、網羅性、共同性、データ入手可能
  - どんな対象として
    - 社会現象として
    - データとして：知識源、言語源、構造化文書源
- Wikipediaはいろいろ利用されている
  - 検索、推薦、マッシュアップ