

Linked Open Data: Computerlinguistische Ressourcen und Werkzeuge im WWW

Seminar im Modul M-GSW-10
SoSe 2018

Prof. Dr. Udo Hahn

Lehrstuhl für Angewandte Germanistische Sprachwissenschaft /
Computerlinguistik

Institut für Germanistische Sprachwissenschaft

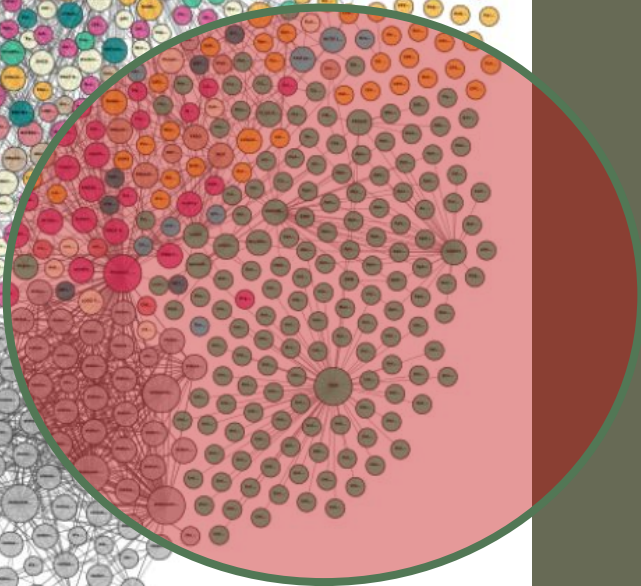
Friedrich-Schiller-Universität Jena

<http://www.julielab.de>

Allgemeine Hinweise

- Termin: Do, 16-18h (Johannisfr.hof 3, SR 2)
- Materialien im Netz
 - <http://www.julielab.de>  „Students“
- Sprechstunde: Mi, 12-13h (bA) (FG 30, R 004)
- Email: udo.hahn@uni-jena.de
- Fachliteratur: durchgängig in Englisch

Linking Open Data Cloud



<http://lod-cloud.net/>



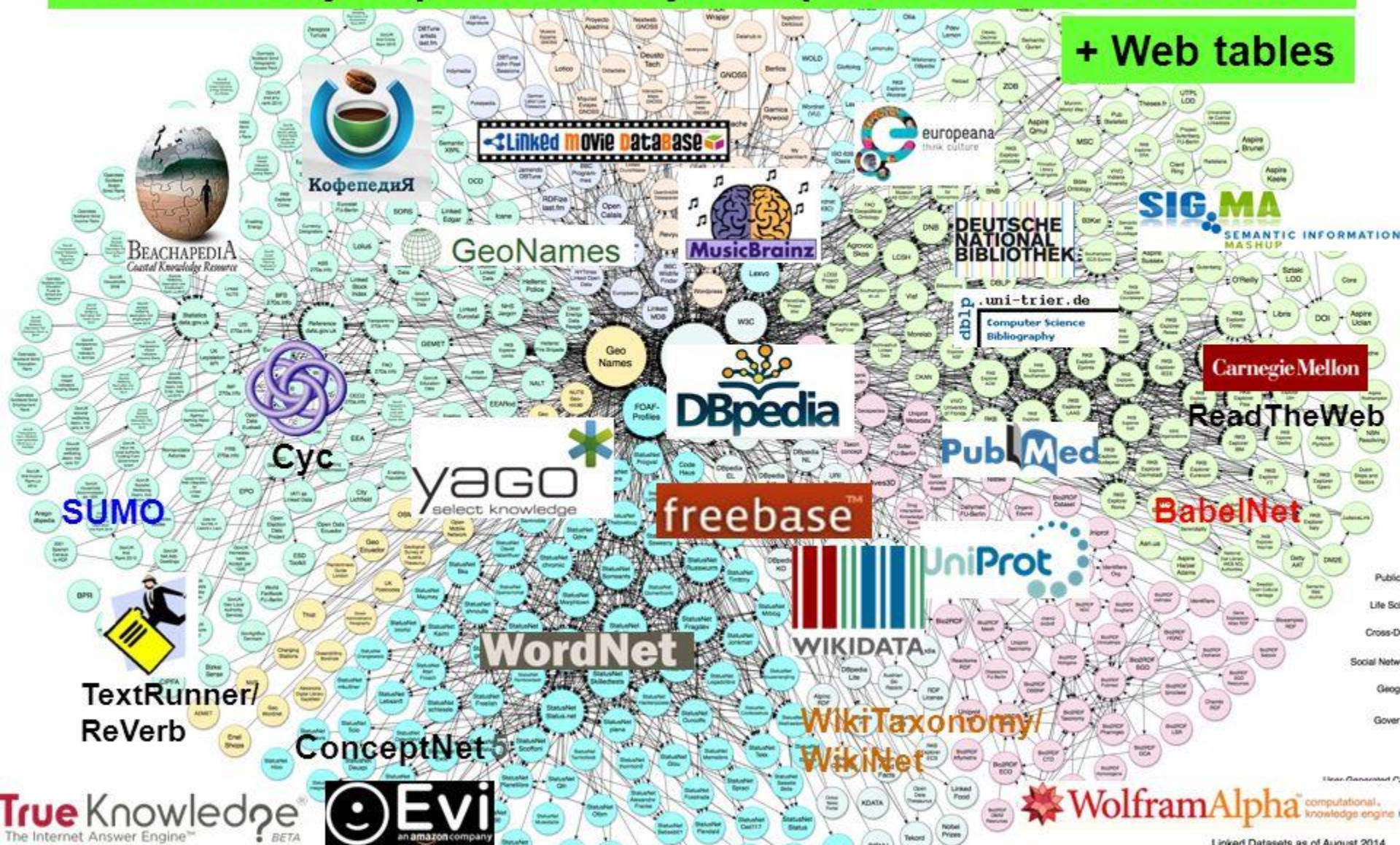
Linked Open Data Cloud



Today's knowledge bases

> 60 Bio. subject-predicate-object triples from > 1000 sources

+ Web tables

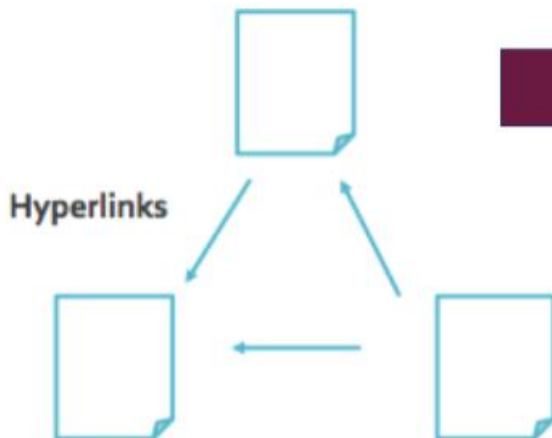


Linked Datasets as of August 2014

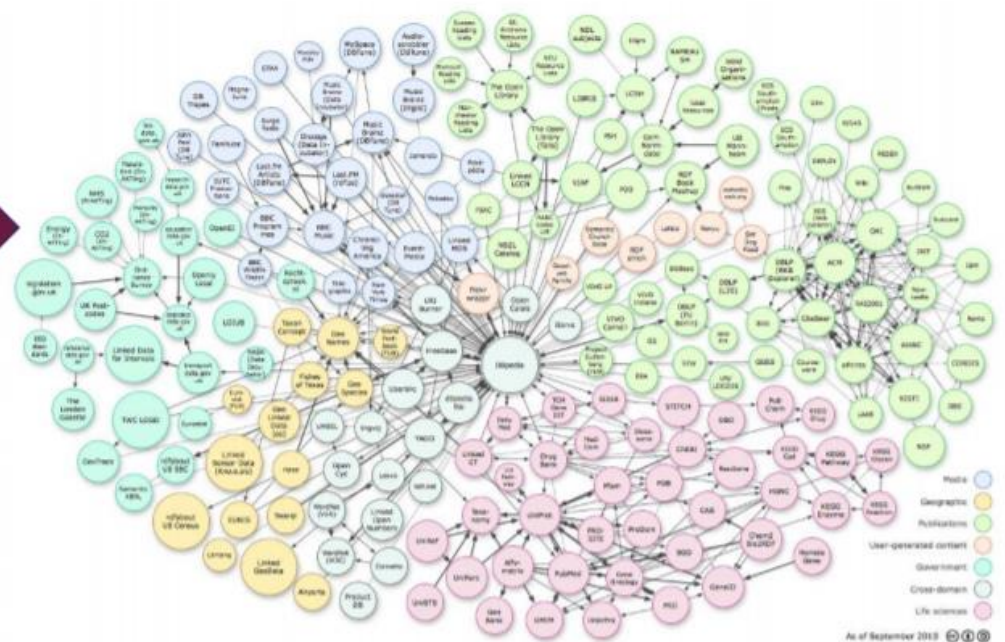
[details>](#)

- WWW = Netz von Webseiten
- LOD = Netz von (inhaltlichen) Daten
 - → Datenintegration & höhere Datenqualität

Web von Dokumenten...



Web von Linked Data...



Technische Rahmenbedingungen von LOD

● Linked Open Data (LOD)

- Im WWW frei verfügbare Datenbestände
- Identifikation durch Uniform Resource Identifier (URI)
- Abruf per HTTP
- Kodierung der Daten via Resource Description Framework (RDF)
- und darauf aufbauende Standards wie SPARQL und die Web Ontology Language (OWL)

Tim Berners-Lee Credo

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Uniform Resource Identifier (URI)

“Ein Uniform Resource Identifier (URI) ist eine kompakte Reihenfolge von Zeichen, die eine abstrakte oder physische Quelle bezeichnet.”

– ISA’s 10 Rules for Persistent URIs

Ein Land, z.B. Belgien

- <http://publications.europa.eu/resource/authority/country/BEL>



Eine Organisation, z.B. Das Amt für Veröffentlichungen

- <http://publications.europa.eu/resource/authority/corporate-body/PUBL>



Ein Datensatz, z.B. Länder Benannt Behörde Liste

- <http://publications.europa.eu/resource/authority/country/>



RDF & SPARQL

Das **Resource Description Framework** (RDF) ist eine Syntax, um Daten und Ressourcen im Web darzustellen.

RDF gliedert jede Information in **Triples**:

- Subjekt – eine Quelle, die mit einer URI identifiziert werden kann.
- Prädikat – eine URI-identifizierte wiederverwendete Besonderheit einer Beziehung.
- Objekt – eine Ressource oder Symbol, mit dem das Thema verwandt ist.

<http://dbpedia.org/resource/Brussels> ist die Hauptstadt von "Belgien".

OR

<http://dbpedia.org/resource/Brussels> ist die Hauptstadt von <http://dbpedia.org/resource/Belgium>.

Subjekt

Prädikat

Objekt

SPARQL ist eine standardisierte Sprache, um RDF-Daten abzufragen.

Rolle der Computerlinguistik im Rahmen von LOD

- CL: „ressource-heavy“
 - GROSSE Korpora (> 1 Mio. Texte, >100 Mio. Tokens, Giga/Terabyte-Skala)
 - GROSSE Lexika (> 100k Einträge)
 - GROSSE Wissensbasen (> 10-100 Mio. Items)
- Ressourcen im World Wide Web als Grundlage für die computerlinguistische Nutzung
- Aufbau von LOD-Ressourcen durch Verfahren der Computerlinguistik


● Europeana

- Multimediale, multilinguale Digital Library zu europäischem Kulturerbe
- 27 Mio Bilder, 22 Mio Texte, 1,1 Mio Videos, 700k Audios,
- <https://www.europeana.eu/portal/de>


Linked Open Data

Europeana


<http://www.europeana.eu/portal/>



eupeana

Fügen Sie einen Suchbegriff ein  Erkunden

Entdecken Sie 51,350,438 Kunstwerke, Artefakte, Bücher, Videos und Audios aus ganz Europa.

Study of Blooming Trees in an Orchard, Ladislav Medný
Slovak National Gallery
© Public Domain 

THEMATISCHE SAMMLUNG

EUROPEANA 1914-1918

THEMATISCHE SAMMLUNG

EUROPEANA ART

THEMATISCHE SAMMLUNG

EUROPEANA FASHION

THEMATISCHE SAMMLUNG

EUROPEANA MUSIC

THEMATISCHE SAMMLUNG

EUROPEANA PHOTOGRAPHY

THEMATISCHE SAMMLUNG

EUROPEANA MIGRATION

AUSSTELLUNG

VISIONS OF WAR

DATENSÄTZE

BROWSE ALL OPENLY LICENSED IMAGES



WIKIPEDIA
Die freie Enzyklopädie

Linked Open Data

Beispiel 2

● Wikipedia

- Mehrsprachige Online Enzyklopädie
- Ca. 40 Mio Artikel (2,17 Mio. in Deutsch)
- Ca. 300 Sprachen

The screenshot shows the Wikipedia article for Tim Berners-Lee. The page layout includes a sidebar on the left with navigation links like 'Main page', 'Contents', and 'Help'. The main content area has a title 'Tim Berners-Lee' and a sub-header 'From Wikipedia, the free encyclopedia'. The article text describes him as an English engineer and computer scientist, known for inventing the World Wide Web. It mentions his birth on 8 June 1955 and his current roles at the University of Oxford and MIT. A portrait of him is shown on the right, with a caption identifying him as 'Sir Timothy John Berners-Lee'. The page also features a search bar at the top right and a 'Not logged in' notification.

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | View source | View history | Search Wikipedia

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA FBCS (born 8 June 1955),^[1] also known as **TimBL**, is an English engineer and computer scientist, best known as the inventor of the World Wide Web. He is currently a professor of Computer Science at the University of Oxford.^[3] He made a proposal for an information management system in March 1989,^[4] and he implemented the first successful communication between a Hypertext Transfer Protocol (HTTP) client and server via the internet in mid-November the same year.^{[5][6][7][8][9]}

Berners-Lee is the director of the World Wide Web Consortium (W3C), which oversees the continued development of the Web. He is also the founder of the World Wide Web Foundation and is a senior researcher and holder of the founders chair at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).^[10] He is a director of the Web Science Research Initiative (WSRI),^[11] and a member of the advisory board of the MIT Center for Collective Intelligence.^{[12][13]} In 2011, he was named as a member of the board of trustees of the Ford Foundation.^[14] He is a founder and president of the Open Data Institute.

In 2004, Berners-Lee was knighted by Queen Elizabeth II for his pioneering work.^{[15][16]} In April 2009, he was elected a foreign associate of the United States National Academy of Sciences.^{[17][18]} Named in *Time* magazine's list of the 100 Most Important People of the 20th century, Berners-Lee has received a number of other accolades for his invention.^[19] He was honoured as the "inventor of the World Wide Web" during the 2012 Summer Olympics opening ceremony, in which he appeared in person, working with a vintage NeXT Computer at the London Olympic Stadium.^[20] He tweeted "This is for everyone",^[21] which instantly was spelled out in LCD lights attached to the chairs of the 80,000 people in the audience.^[20] Berners-Lee received the 2016 Turing Award "for inventing the World Wide Web, the first web browser, and the fundamental protocols and algorithms allowing the Web to scale".^[22]

Sir
Tim Berners-Lee
OM KBE FRS FREng FRSA FBCS

Berners-Lee in 2014

Born
Timothy John Berners-Lee
8 June 1955 (age 62)^[1]
London, England

Wiktionary

- Wörterbuch-Pendant zu Wikipedia
- Mehrsprachiges Wörterbuch/Thesaurus
- 23,6 Mio Einträge
- 172 Sprachen



Beispiel 4

- Wikipedia → Dbpedia (U MA, U L, HPI)
 - Semi-strukturierte Daten (Tabellen) und Fließtext aus Wikipedia werden mit computerlinguistischen Verfahren automatisch überführt in vollständig strukturierte Daten
 - Dbpedia (2014): 4,6 Mio. Datensätze, 3 Mrd. Fakten
 - Multilinguale Wikipedia (EN, DE, FR, ES, IT, ...)
 - Verlinkt mit Freebase, Open Cyc, UMBEL, GeoNames, MusicBrainz, CIA World Factbook, New York Times [LOD], Digital Bibliography & Library Project, Project Gutenberg, Jamendo, Eurostat US-Census
 - Datenrepräsentation: RDF

Ein Wikipedia-Eintrag



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

Tools

[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Wikipedia:Items](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#)

Read

[View source](#)

[View history](#)

Search Wikipedia

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee *OM KBE FRS FREng FRSA FBCS* (born 8 June 1955),^[1] also known as **TimBL**, is an English [engineer](#) and [computer scientist](#), best known as the inventor of the [World Wide Web](#). He is currently a professor of Computer Science at the [University of Oxford](#).^[3] He made a proposal for an information management system in March 1989,^[4] and he implemented the first successful communication between a [Hypertext Transfer Protocol \(HTTP\)](#) client and [server](#) via the [internet](#) in mid-November the same year.^{[5][6][7][8][9]}

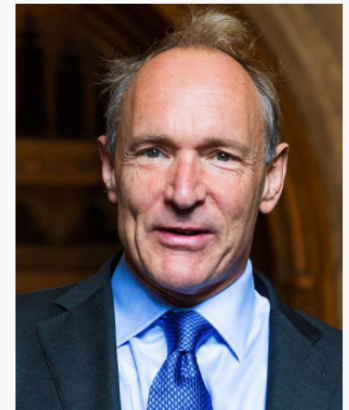
Berners-Lee is the director of the [World Wide Web Consortium \(W3C\)](#), which oversees the continued development of the Web. He is also the founder of the [World Wide Web Foundation](#) and is a senior researcher and holder of the [founders chair](#) at the [MIT Computer Science and Artificial Intelligence Laboratory \(CSAIL\)](#).^[10] He is a director of the [Web Science Research Initiative \(WSRI\)](#),^[11] and a member of the advisory board of the [MIT Center for Collective Intelligence](#).^{[12][13]} In 2011, he was named as a member of the board of trustees of the [Ford Foundation](#).^[14] He is a founder and president of the [Open Data Institute](#).

In 2004, Berners-Lee was [knighted](#) by Queen [Elizabeth II](#) for his pioneering work.^{[15][16]} In April 2009, he was elected a foreign associate of the [United States National Academy of Sciences](#).^{[17][18]} Named in *Time* magazine's list of the [100 Most Important People of the 20th century](#), Berners-Lee has received a number of other accolades for his invention.^[19] He was honoured as the "Inventor of the World Wide Web" during the [2012 Summer Olympics opening ceremony](#), in which he appeared in person, working with a vintage [NeXT Computer](#) at the [London Olympic Stadium](#).^[20] He [tweeted](#) "This is for everyone",^[21] which instantly was spelled out in [LCD](#) lights attached to the chairs of the 80,000 people in the audience.^[20] Berners-Lee received the 2016 [Turing Award](#) "for inventing the World Wide Web, the first web browser, and the fundamental protocols and algorithms allowing the Web to scale".^[22]

Sir

Tim Berners-Lee

OM KBE FRS FREng FRSA FBCS




Berners-Lee in 2014

Born

Timothy John Berners-Lee
 8 June 1955 (age 62)^[1]
[London, England](#)

Ein Wikipedia-Eintrag und seine Dbpedia-Form



Not logged in | Talk | Contributions | Create account | Log in

Article | **Talk** | Read | View source | View history

Search Wikipedia

<Weaving the Web> <is written by> <Tim Berners-Lee> .
<Tim Berners-Lee> <has first name> "Tim" .
<Tim Berners-Lee> <has last name> "Berners-Lee" .
<Tim Berners-Lee> <is born on> "06/08/1955" .
<Tim Berners-Lee> <is born in> <London> .
<London> <is located in> <England> .
<London> <has population> "7825200" .
<London> <hat Fläche> "130395 km²" .

Contact page

Tools

What links here

Related changes

Upload file


Special pages

Permanent link

Page information

Wikipedia items

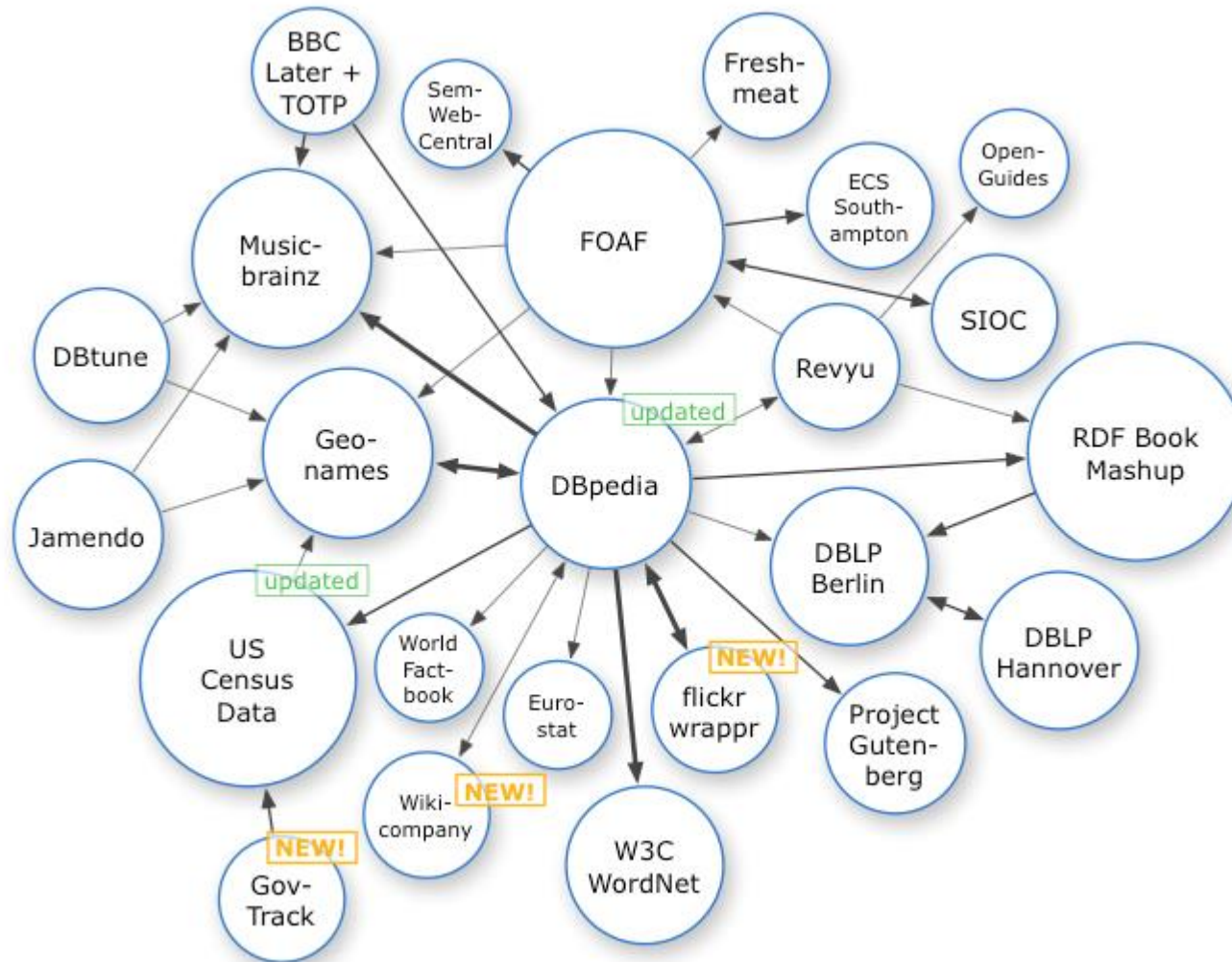
In 2004, Berners-Lee was knighted by Queen Elizabeth II for his pioneering work.^{[15][16]} In April 2009, he was elected a foreign associate of the United States National Academy of Sciences.^{[17][18]} Named in *Time* magazine's list of the 100 Most Important People of the 20th century, Berners-Lee has received a number of other accolades for his invention.^[19] He was honoured as the "Inventor of the World Wide Web" during the 2012 Summer Olympics opening ceremony, in which he appeared in person, working with a vintage NeXT Computer at the London Olympic Stadium.^[20] He tweeted "This is for everyone",^[21] which instantly was spelled out in LCD lights attached to the chairs of the 80,000 people in the audience.^[20] Berners-Lee received the 2016 Turing Award "for inventing the World Wide Web, the first web browser, and the fundamental protocols and algorithms allowing the Web to scale".^[22]



Berners-Lee in 2014

Born Timothy John Berners-Lee
8 June 1955 (age 62)^[1]
London, England

Datenquellen von DBpedia



◉ Wikipedia → YAGO (MPI SB)

- Automatisch aus Wikipedia erzeugte Ontologie und Wissensbasis (computerlinguistische Methodik)
- Unter Verwendung von Geonames, WordNet
- Links zu Dbpedia und SUMO ontology
- 10 Mio. Einträge und 120 Mio. Fakten
- <https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/>

- Repräsentationsaspekte und technischer Hintergrund zu LOD
 - RDF, URI, SPARQL, SKOS, OWL, HTTP ... Triple-Stores
- Textressourcen
 - Wikipedia, Europeana, Projekt Gutenberg, ...
- Sprachressourcen (Lexika)
 - Wiktionary, WordNet, LemonUBY, ...
- Computerlinguistische Anwendungen
 - Wikipedia → DBpedia
 - Wikipedia → YAGO

◎ Vortrag (mündlich)

- 1-stündig
- Elektronische Version (PDF, PPT) verfügbar machen

◎ Referat (schriftlich)

- 15-20 Seiten Kerntext (mit Standardformaten)
- Elektronische Version (PDF, DOC) verfügbar machen
- Eidesstattliche Erklärung zur Eigenautorenschaft
 - Wir prüfen mit Plagiatserkennungs-Software
- Abgabe: Anfang Juli 2018

Bemerkungen zu Referaten

● Aufbaumuster:

- Deck- bzw. Titelblatt mit vollständigen Angaben
- Inhaltsverzeichnis
- Einführung ins Thema, Motivation
- Themenabhandlung: grundlegende Formalisierungen, Verfahrensbeschreibungen (Algorithmen), Systemfunktionalitäten, Ressourcenmerkmale, Experimente/Evaluationen usw.
- Fazit mit kritischer Würdigung, offene Probleme ansprechen
- Bibliographie

● Zitationen:

- Alle verwendeten Quellen zitieren
 - Mit einem bibliographisch korrektem Zitat die jeweilige Quelle eindeutig beschreiben
 - Fachartikel nicht mit <http://...foo.pdf>-Link zitieren
 - Online-Quellen mit URLs und Datum des letztem Zugriffs
- **Wikipedia** ist keine zitierfähige wissenschaftliche Quelle !

● Eigenleistungen (Literatur, Beschäftigung mit konkreten Ressourcen/Systemen usw.) sind sehr erwünscht → unabdingbar !

Wege zum Vortrag und Referat

- Email: Anmeldung von **drei** nach fallender Priorität geordneten Themenwünschen
 - First-come, first-served
- Email: Themenvergabe durch Dozenten
- Erste Literaturhinweise als „Saat“ nach Bestätigung der Themenauswahl
- Themenbearbeitung durch Referenten
 - Mündlicher Vortrag zum vereinbarten Termin
 - Schriftliches Referat (unter Einhaltung der organisatorischen Verabredungen) zum vereinbarten Termin

- Felix Ostrowski; Pascal Christoph: Einführung in Linked Open Data (2011)
 - <http://swib.org/swib11/vortraege/swib11-felix-ostrowski.pdf>

Wichtige Zeitschriften

- Journal of Web Semantics: Science, Services and Agents on the World Wide Web
- Semantic Web – Interoperability, Usability, Applicability

Wichtige Konferenzen

- Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linked Data Science @ LREC 2018 [und frühere Workshops; üblich: 2-jährlich mit LREC]
- LREC 2018 – Proceedings of the 11th International Conference on Language Resources and Evaluation [und frühere Konferenzen, 2-jährlich: 2016, 2014, ...]
- The Semantic Web. Proceedings of the 17th International Semantic Web Conference – ISWC 2018 [und frühere Konferenzen; jährlich]
- WWW '18 – Proceedings of the 27th International Conference on World Wide Web [und frühere Konferenzen; jährlich]
- The Semantic Web: Proceedings of the 15th European Semantic Web Conference – ESWC 2018 [und frühere Konferenzen; jährlich]

Ablaufplan

12.4. Hahn
19.4. ---
26.4. ---
03.5. Hahn – Themenvergabe
10.5. ---
17.5. ---
24.5. ---
31.5. Gesprächstermin
07.6. Gesprächstermin
14.6. ---
21.6. ---
28.6. --- (Retreat)
05.7. Patrick Zerrer
12.7. Natalia Sulaberidze

Sprachressourcen: Lexika
Textressourcen:
Wikipedia, Europeana