

De toekomst van het wetenschappelijk document

Gosse Bouma

Informatiekunde
Rijksuniversiteit Groningen

UB 28/09/07

Inleiding

citeseer.org

Articles freely available online are more highly cited. For greater impact and faster scientific progress, authors and publishers should aim to make research easy to access.

Steve Lawrence, NEC Research Institute



Inleiding

Het Web volgens Wikipedia

De uitvinding van het wereldwijde web wordt toegeschreven aan de Engelsman Tim Berners-Lee en de Belg Robert Cailliau als **een methode voor wetenschappers** van het CERN (waar ze beiden werkten) makkelijk met elkaar te kunnen laten communiceren.

Wat wil een wetenschapper?

Als onderzoeker

- Zoeken op auteur, steekwoord, onderwerp, jaar, ...
- Zoeken **in** literatuur
- Gemakkelijk **toegang** tot literatuur

Als auteur

- Publicaties
- Citaties

Wat wil een wetenschapper?

Als onderzoeker

- Zoeken op auteur, steekwoord, onderwerp, jaar, ...
- Zoeken **in** literatuur
- Gemakkelijk **toegang** tot literatuur

Als auteur

- Publicaties
- Citaties

Naar een digitale bibliotheek

Voordelen

- 24/7, wereldwijd, geopend
- Superieure zoekfuncties
- Zeer grote boekenplanken
- Redelijke kosten

Vereisten

- Archivering van elektronische literatuur
- Toegang voor gebruikers
- 'Slimme' zoekfuncties
- Integratie in het wetenschappelijke proces



Naar een digitale bibliotheek

Voordelen

- 24/7, wereldwijd, geopend
- Superieure zoekfuncties
- Zeer grote boekenplanken
- Redelijke kosten

Vereisten

- Archivering van elektronische literatuur
- Toegang voor gebruikers
- 'Slimme' zoekfuncties
- Integratie in het wetenschappelijke proces



Overzicht

- 1 Inleiding
- 2 Elektronische documenten en Archivering
- 3 Nieuwe diensten
- 4 Het Semantisch Web
- 5 E Science

Hoe verwijst je naar een digitaal document

URLs bestaan niet eeuwig...

- Gertjan van Noord. Using Self-Trained Bilexical Preferences to Improve Disambiguation Accuracy. In: *Proceedings of the Tenth International Conference on Parsing Technologies (IWPT)*, 2007, Prague. Pages 1–10.
- `www.let.rug.nl/~vannoord/papers/iwpt2007.pdf`
- `acl.ldc.upenn.edu/W/W07/W07-2201.pdf`
- `acl.eldoc.ub.rug.nl/mirror/W/W07/W07-2201.pdf`



Digital Object Identifiers

DOI

The DOI System is for identifying content objects in the digital environment. Information about a digital object may change over time, including where to find it, but its DOI name will not change.

- Gosse Bouma, Petra Hendriks, and Jack Hoeksema. Focus Particles inside Prepositional Phrases: A Comparison between Dutch, English, and German. *Journal of Comparative Germanic Linguistics*, (2007) 10:1-24.
- www.let.rug.nl/~gosse/papers/Focuspart_web.pdf



dx.doi.org/10.1007/s10828-006-9006-1

SpringerLink - Journal Article - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

Content Types Subject Collections English

Journal Article

Focus Particles Inside Prepositional Phrases: A Comparison of Dutch, English, and German

Journal: The Journal of Comparative Germanic Linguistics
Publisher: Springer Netherlands
ISSN: 1383-4924 (Print) 1572-8552 (Online)
Issue: Volume 10, Number 1 / March, 2007
Category: Original Paper
DOI: 10.1007/s10828-006-9006-1
Pages: 1-24
Subject Collection: Humanities, Social Sciences and Law
SpringerLink Date: Thursday, February 15, 2007

OPEN CHOICE

Remnant Raising and VSO Clausal Architecture
Lee, Felicia

Gosse Bouma¹, Petra Hendriks¹ and Jack Hoeksema¹

(1) Center for Language and Cognition Groningen, University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands

Received: 17 September 2004 **Revised:** 22 June 2005 **Accepted:** 14 August 2005 **Published online:** 14 February 2007

Abstract Partly due to disagreement on acceptability judgements, there is little agreement on the possibility of PP-internal and DP-internal focus particles in languages such as Dutch, English, and German. Our large-scale corpus investigation reveals that PP-internal focus particles are a genuine possibility, not only in English, but also in Dutch and, to a lesser extent, German. These results seem to be incompatible with a number of existing syntactic theories of focus in German. However, our investigation also provides evidence for a strong

Find more options

Within all content
 Within this journal
 Within this issue

Export this article

Export this article as RIS | Text

Text

PDF

The size of this document is 245 kilobytes. Although it may be a lengthier download, this is the most authoritative online format.

Open Full document

Elektronische tijdschriften

Door uitgeverij

- Copie van de gedrukte versie
- Vrij toegankelijk voor onderzoekers
- Mits de kennisinstelling betaalt

Open Access tijdschriften

- Verschijnt (meestal) niet in druk
- Vrij toegankelijk
- Extra diensten en mogelijkheden
 - Sneller publiceren
 - Langere artikelen
 - Aandacht voor wetenschappelijke niches

Elektronische tijdschriften

Door uitgeverij

- Copie van de gedrukte versie
- Vrij toegankelijk voor onderzoekers
- Mits de kennisinstelling betaalt

Open Access tijdschriften

- Verschijnt (meestal) niet in druk
- Vrij toegankelijk
- Extra diensten en mogelijkheden
 - Sneller publiceren
 - Langere artikelen
 - Aandacht voor wetenschappelijke niches

neerlandistiek.nl



Waarom een elektronisch tijdschrift?

De bestaande neerlandistische tijdschriften zijn en blijven in zwaar weer....kostenloze verspreiding.... Zodra een artikel de beoordelingsprocedure doorlopen heeft, staat het ter beschikking van de lezers ...een elektronische publicatie maakt bovendien mogelijk dat ruwe onderzoeksgegevens voor de lezer raadpleegbaar zijn

Ondersteuning voor Elektronische tijdschriften

Open Access Journals

- 900 tijdschriften in 10 talen
- Annals of Genealogical Research, **Archivos Latinoamericanos de Produccion Animal**, Cosmos and History, **HISTOIRE(S) de l'Amérique latine**, Poetry and Science, **Transnational Curriculum Inquiry**, Vanderbilt Undergraduate Research Journal, **Wildlife Biology in Practice**, Women in Judaism: A Multidisciplinary Journal



Directory of Open Access Journals (2300 ts)

The screenshot shows the DOAJ website interface. The browser window title is "Directory of open access journals - Mozilla Firefox". The address bar shows "Google News Nederland". The DOAJ logo is prominently displayed at the top left. A sidebar on the left contains navigation links: "Find Journals", "New titles", "Find articles", "Suggest a journal", "About", "FAQ", "News", "Sponsors", "Membership", "Contact", "For journal owners", and "For authors". The main content area features a "Find articles" section with a search bar and a "Presenting page 1/7" indicator. Below this, three article listings are shown, each with a title, author, journal, year, volume, issue, and page/record number, followed by "View record" and "Fulltext" links.

Find articles
Found 68 documents matching your query :
All Fields=digital libraries

Presenting page 1/7

The Dede Korkut digital library: a student project faces the real world
Author: Bayram, Ozlem Journal: Library Student Journal Year: 2007 Vol: Issue: Pages/record No.:
[View record](#) [Fulltext](#)

Factors influencing publication choice: why faculty choose open access
Author: Warlick Stefanie ; Vaughan KTL Journal: Biomedical Digital Libraries Year: 2007 Vol: 4 Issue: 1 Pages/record No.: 1
[View record](#) [Fulltext](#)

Biblioteca universal: um sonho antigo da humanidade
Author: Joana Coeli Ribeiro Garcia Journal: Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação Year: 2007 Vol: Issue: 23 Pages/record No.: 62-72
[View record](#) [Fulltext](#)

CAMbase – A XML-based bibliographical database on Complementary and Alternative Medicine (CAM)
Author: Ostermann Thomas ; Zillmann Hartmut ; Raak Christa ; Buessing Arndt ; Matthiessen Peter Journal: Biomedical Digital Libraries Year: 2007 Vol: 4 Issue: 1 Pages/record No.: 2
[View record](#) [Fulltext](#)

The strike rate index: a new index for journal quality based on journal size and the h-index of citations
Author: Barendse William Journal: Biomedical Digital Libraries Year: 2007 Vol: 4 Issue: 1 Pages/record No.: 3
[View record](#) [Fulltext](#)

Verbeterde informatievoorziening

ScienceDirect (Elsevier)

- Gebruik **e-mail** en textcolorredRSS feeds om onderzoekers te informeren
- nieuwe **nummers** van geselecteerde tijdschriften
- nieuwe **artikelen** over geselecteerde onderwerpen
- **citaties**

berners lee - ResearchIndex document query - Mozilla Firefox

File Edit View History Bookmarks Tools Help delicio.us

CiteSeer Find: Documents Citations

Searching for **PHRASE** berners lee.

Restrict to: [Title](#) [No restrictions](#) Order by: [Expected citations](#) [Hubs](#) [Usage](#) [Date](#) Try: [Google \(CiteSeer\)](#)
[Google \(Web\)](#) [Yahoo!](#) [MSN](#) [CSB](#) [DBLP](#)

Results restricted to the header field. Processing 524 documents.
45 documents found. **Order: number of citations.**

[Hypertext Transfer Protocol - HTTP/1.1 - Fielding, Gettys, Mogul, Frystyk, ... \(1998\)](#) (Correct) (230 citations)
L. Masinter, Xerox P. Leach, Microsoft T. **Berners-Lee**, W3C/MIT Expires May 18, 1999 November 18, 1436, University of Minnesota, March 1993. 3] **Berners-Lee**, T.Universal Resource Identifiers in WWW, [www.w3.org/Protocols/HTTP/1.1/diff-v11-Rev05ToRev06.ps.gz](#)

[Ontology-based Web Agents - Luke, Spector, Rager, Hendler \(1997\)](#) (Correct) (72 citations)
web page and its links. Related Work HTML 2.0 (**Berners-Lee** and Connolly 1995) includes several weak [stanford.edu/knowledge-sharing/papers/kif.ps](#) **Berners-Lee**, T. and D. Connolly. 1995. Hypertext Markup [gopher.cs.umbc.edu/courses/graduate/891/current/shoe.ps](#)

[The Semantic Web: The Roles of XML and RDF - Decker, Melnik, Van Harmelen.. \(2000\)](#) (Correct) (42 citations)
but the next-generation Web, which Tim **Berners-Lee** and others call the "Semantic Web, aims at languages. REFERENCES 1. T. **Berners-Lee**, Weaving the Web, Harper, San Francisco, [www.cs.vu.nl/~frankh/postscript/IEEE00.pdf](#)

[Emerging Trends in the WWW User Population - Pitkow, Kehoe \(1996\)](#) (Correct) (22 citations)
been devoted towards the World Wide Web (WWW) **Berners-Lee** 94]but relatively little research has been revolution since Guttenberg. References **Berners-Lee**, T.Cailliau, R.Luotonen, A.Nielsen, H. [www.cc.gatech.edu/gvu/user_surveys/papers/cacm.ps](#)

citeseer.org

Citations: Ontology-based Web agents - Luke, Spector, Rager, Hendler (ResearchIndex) - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

75 citations found. Retrieving documents...

Sean Luke, Lee Spector, David Rager, and Jim Hendler. *Ontology-based Web Agents*". In Proceedings of First International Conference on Autonomous Agents, 1997.
(<http://www.cs.umd.edu/projects/plus/SHOE/>).

CiteSeer [Home/Search](#) [Document Details and Download](#) [Summary](#) [ACM](#) [TOC](#) [Related Articles](#)
[Check](#)

This paper is cited in the following contexts:

[First 50 documents](#) [Next 50](#)

[Annotation of the Shallow and the Deep - Handschuh, Staab](#) (Correct)

....and improper references. **The** approaches from this line of research that are closest to CREAM is the SHOE Knowledge Annotator , WebKB, and the MnM annotation tool. **The** SHOE Knowledge Annotator is a Java program that allows users to mark up web pages with the SHOE ontology. **The SHOE system [26] defines additional tags that can be embedded in the body of HTML pages.** The SHOE Knowledge Annotator is rather a little helper (like our earlier OntoPad [12] 6) than a full fledged annotation environment. **WebKB [27]** uses conceptual graphs for representing the semantic content of Web

S. Luke, L. Spector, D. Rager, and J. Hendler. *Ontology-based Web Agents*. In Proceedings of the First International Conference on Autonomous Agents, Marina del Rey, CA, USA, February 5-8, 1997.

[CREAM - CREating Metadata for the Semantic Web - Handschuh, Staab \(2003\)](#) (1 citation) (Correct)

....in the scenario section. **The** approaches from this line of research that are closest to CREAM is the SHOE Knowledge Annotator 5. WebKB, and the MnM annotation tool. **The** SHOE Knowledge Annotator is a Java

citeseer.org

Ontology-based Web Agents - Luke, Spector, Rager, Hendler (ResearchIndex) - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://citeseer.ist.psu.edu/luke97ontologybased

S. Luke, L. Spector, D. Rager, and J. Hendler. Ontology-based Web agents. In W. L. Johnson, editor, Proceedings of the 1st International Conference on Autonomous Agents, pages 59–66. ACM, 1997. <http://citeseer.ist.psu.edu/luke97ontologybased.html> [More](#)

```
@inproceedings{ luke97ontologybased,  
  author = "Sean Luke and Lee Spector and David Rager and James Handler",  
  title = "Ontology-based Web Agents",  
  booktitle = "Proceedings of the First International Conference on Auto",  
  publisher = "ACM Press",  
  address = "Marina del Rey, CA, USA",  
  editor = "W. Lewis Johnson and Barbara Hayes-Roth",  
  pages = "59--68",  
  year = "1997",  
  url = "citeseer.ist.psu.edu/luke97ontologybased.html" }
```

Citations (may not include all citations):

- 30 [Computer Science Department \(context\)](#) - Logic- - 1995
- 1 [HyperText Markup Language Specification Version 3 \(context\)](#) - http, cs et al. - 1995

Year of Publication of Citing Articles

Year	Number of Citations
1995	8
1996	8
1997	7
1998	8
1999	12
2000	6
2001	11
2002	4

Geboortedata

1977 - Wikipedia - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

W http://nl.wikipedia.org/wiki/1977

Geboren [bewerken]

januari

- 1 - [Keeley Hawes](#), Brits actrice
- 5 - [Hadewych Minis](#), Nederlands actrice
- 11 - [Anni Friesinger](#), Duits schaatsster
- 13 - [Orlando Bloom](#), Brits acteur
- 15 - [Marja Vis](#), Nederlands schaatsster
- 16 - [Ariel Zeevi](#), Israëlisch judoka
- 24 - [Minke Booij](#), Nederlands hockeyinternational
- 24 - [Michelle Hunziker](#), Nederlands tv-presentatrice
- 25 - [Hatem Trabelsi](#), Tunesisch voetballer
- 25 - [Luke Roberts](#), Australisch wielrenner
- 26 - [Nathan Vecht](#), Nederlands cabaretier
- 28 - [Takuma Sato](#), Japans Formule 1-coureur
- 31 - [Torri Edwards](#), Amerikaanse atlete

februari

- 1 - [Sonja Silva](#), Nederlands actrice en tv-presentatrice
- 2 - [Marc Bernaus](#), Andorrees voetballer



Hoe vindt een machine geboortedata?

```
<title>1977 - Wikipedia</title>
...
<h2>Geboren</h2>
<dl>
<dt><b>januari</b></dt>
</dl>
<ul>
<li>
  <a href="/wiki/11_januari" title="11 januari">
    11</a> -
    <a href="/wiki/Anni_Friesinger"
      title="Anni Friesinger">
      Anni Friesinger</a>,
Duits schaatsster
</li>
```



Het Semantisch Web

Een web voor mensen en machines

- Gestructureerde informatie
- Betekenis van informatie wordt vastgelegd
- Software kan informatie verzamelen, combineren, afleiden

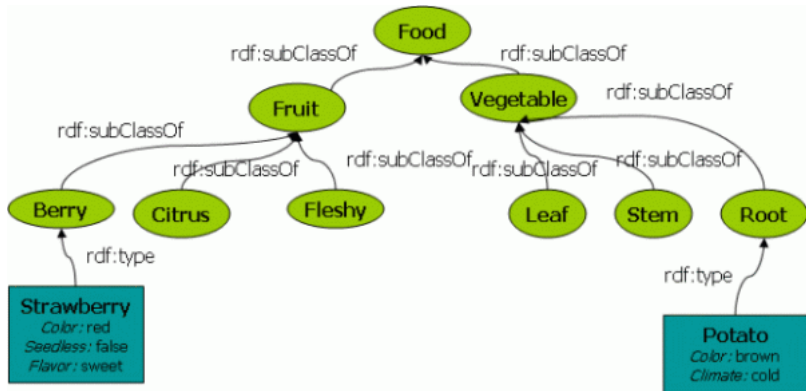


Hoe vindt een machine geboortedata?

```
<pagina categorie='jaar'>  
  <jaar>1977</jaar>  
  <geboren>  
    <naam wiki='wiki/Anni_Friesinger'>  
      Anni Friesinger</naam>  
    <datum>1977-01-11</datum>  
  </geboren>  
  <geboren>  
    <naam wiki='Marja Vis'>Marja Vis</naam>  
    <datum>1977-01-15</datum>  
  </geboren>
```



Ontologie



Het Semantisch Web voor Onderzoekers

'slim' zoeken

- woordenboek → lexicon
- nier → kidney
- Washington → de persoon of de plaats of de staat?
- David Warren → D.H.D. Warren (informaticus), D.S. Warren (informaticus)
- Michael Jackson → de zanger, de acteur, de generaal, de basketballer, ..., (27 totaal)

Het Semantisch Web voor Onderzoekers

Uitleg over concepten

- In dit model werd humaan amnionvlies gebruikt als modelsubstraat voor **peritoneum**. Met behulp van scanning- en transmissie-electronenmicroscopie werd ...
- Het buikvlies of **peritoneum** (niet te verwarren met perineum) is de laag die van binnen de buikholte bekleedt en de daarin gelegen organen. (wikipedia)

Het Semantisch Web voor Onderzoekers

Linked Data

- unieke verwijzingen voor auteurs en artikelen
- artikelen indiceren op concepten
 - → **metadata**
- semantische relaties tussen concepten



Informatie-extractie

Oorzaak-gevolg relaties (Wikipedia, 15000)

oorzaak	patroon	gevolg
een algemene immunodeficiëntie	leidt tot	Aids
HIV	de oorzaak van	aids
hiv	de oorzaak van	de ziekte aids
Hiv	veroorzaakt	aids
Hiv-2	veroorzaakt	aids maar niet altijd en meestal in minder ernstige vorm
het HIV-virus	veroorzaakt	de ziekte aids
de aanwezigheid van onwetendheid	leidt tot	het lijden

Digitale Informatie en E Science

E Science

- Delen van data en computer resources (grid computing)
- Wetenschappelijke blogs
- On-line verzamelen en annoteren van relevante literatuur



SW en E Science: Data Grid

CombeChem

- semantic web technology to **describe and link chemical information**
- Semantic Data Grid: tens of millions of RDF triples across multiple stores
- techniques are needed to disseminate the data
- software support that improves the human experience of conducting the research

SW en E Science: Annotatie

SWAN

- a semantic web application in neuromedicine
- Alzheimer disease research
- support for personal data organization, **hypothesis generation**, experimentation, lab data organization, and **digital pre-publication collaboration**,
- **organized and interconnected** using a common semantic framework



SW en E Science: Annotatie

World Wide Argument Web

- a web of **interconnected arguments**
- an open platform for representing arguments and for building linked and dynamic argument networks on the Semantic Web



Digitale bibliotheken en E Science

Geannoteerde Wetenschappelijke literatuur

- Literatuur is digitaal beschikbaar
- Onderliggende data is beschikbaar
- Onderzoekers **annoteren** relevante literatuur
 - concepten, relaties, argumenten voor en tegen een hypothese, datasets, software
- Semantic Web technieken
 - **Unieke verwijzingen** voor auteurs, artikelen, concepten, relaties
 - **Metadata** gekoppeld aan **ontologische informatie**
 - slim zoeken
 - definities, relaties

Tot slot

Wat denkt de UB?

Dankzij significante vorderingen in de technieken van tekstanalyse, in slimme zoekmachines bijvoorbeeld die zoeken op concepten in plaats van op woorden, en in de toch zeer fraaie ontwikkelingen op het gebied van het semantisch web, kan het universum van documenten omgevormd worden tot een goed doorzoekbaar web van documenten met onderlinge relaties

H. Ellerman, Pictogram, sept 2007



Tot slot

Wat denkt de UB?

Dankzij significante vorderingen in de technieken van tekstanalyse, in slimme zoekmachines bijvoorbeeld die zoeken op concepten in plaats van op woorden, en in de toch zeer fraaie ontwikkelingen op het gebied van het semantisch web, kan het universum van documenten omgevormd worden tot een goed doorzoekbaar web van documenten met onderlinge relaties

H. Ellerman, Pictogram, sept 2007

