

Grasping for structured data

Peter Brodersen,
Korruption / Findvej.dk

25. januar 2008
BarCamp Copenhagen

Min baggrund

- Webudvikler siden 1996
 - Selvstændig siden 2003
 - Kommunikation gennem teknik
 - Ophavsmand til Findvej.dk
-
- .. og bruger og sporadisk skribent på Wikipedia

Troværdighed og kildemateriale

- "Kilder til Wikipedias indhold kræves, fordi artiklerne ikke må være original forskning og skal kunne verificeres"
- Alle revisioner og meta-information bevares
- Værktøjer til at se forskelle mellem revisioner
- Mulighed for at henvise til en bestemt revision

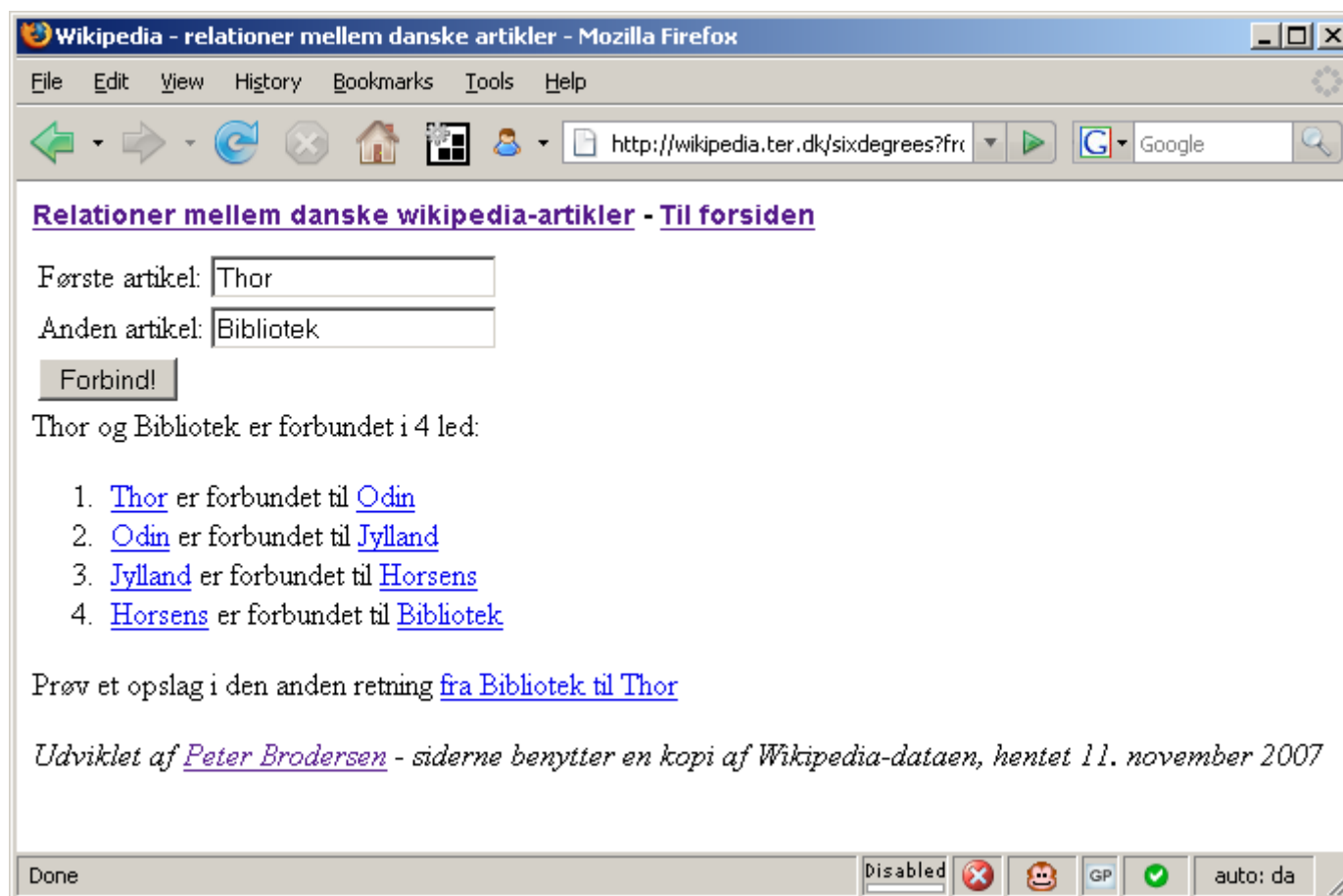
Troværdighed og kildemateriale

- Encyclopædia Britannica vs Wikipedia
- Omtrent samme fejlrate – men hvilke artikler?
- Aktuelle artikler er oftere udsat for hærværk

Strukturering af informationer

- Interne referencer
- Skabeloner og kategorier
- Geografisk metadata
- Revisionsdata

Eksempel: Six Degrees



Eksempel: Mest ønskede artikler

Wikipedia:Mest ønskede artikler - Wikipedia, den frie encyklopædi - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://da.wikipedia.org/wiki/Wikipedia:Mest_%C3%B8nskede_artikler#

henvisninger. Listen er sorteret efter i hvor mange forskellige artikler ønskerne er fordelt.

Mest ønskede artikler [\[redigér\]](#)

- [Amatør](#) (23 artikler med i alt 25 henvisninger)
- [Nomade](#) (23 artikler med i alt 24 henvisninger)
- [Parlophone](#) (23 artikler med i alt 24 henvisninger)
- [Planetring](#) (22 artikler med i alt 24 henvisninger)
- [Vedbæk](#) (21 artikler med i alt 25 henvisninger)
- [Billboard Hot 100](#) (21 artikler med i alt 21 henvisninger)
- [Bob Ramsing](#) (21 artikler med i alt 21 henvisninger)
- [F-blok](#) (21 artikler med i alt 21 henvisninger)
- [Hofjunker](#) (21 artikler med i alt 21 henvisninger)
- [Monument](#) (21 artikler med i alt 21 henvisninger)
- [American Recordings](#) (20 artikler med i alt 26 henvisninger)
- [Parameter](#) (20 artikler med i alt 21 henvisninger)
- [Universal Music](#) (20 artikler med i alt 21 henvisninger)
- [Designer](#) (20 artikler med i alt 20 henvisninger)
- [Mosaik](#) (20 artikler med i alt 20 henvisninger)
- [Myndighed \(institution\)](#) (20 artikler med i alt 20 henvisninger)
- [Proterozoikum](#) (19 artikler med i alt 21 henvisninger)
- [Dåbsfad](#) (19 artikler med i alt 19 henvisninger)
- [Elektra Records](#) (19 artikler med i alt 19 henvisninger)
- [Motiv](#) (19 artikler med i alt 19 henvisninger)
- [Transportfly](#) (19 artikler med i alt 19 henvisninger)
- [Kunstskejteløb](#) (18 artikler med i alt 25 henvisninger)
- [Albert R. Broccoli](#) (18 artikler med i alt 21 henvisninger)
- [Generalsekretær](#) (18 artikler med i alt 19 henvisninger)
- [Mølle](#) (18 artikler med i alt 19 henvisninger)
- [Tekstkritik](#) (18 artikler med i alt 19 henvisninger)

Done Disabled auto: da

Eksempel: Strukturerede opslag

Semantic Wikipedia Query: Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabita...

File Edit View History Bookmarks Tools Help

http://wikipedia.aksw.org/index.php?qid=1#

Google

Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | [Most popular](#) | [Upcoming](#)

[Tennis players from Moscow](#)
[Pictures of Clint Eastwood's Movies](#)
[Sitcoms set in NYC](#)
[Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants](#)
[People influenced by Friedrich Nietzsche](#)
[Films longer than 5 hours](#)
[Space Missions](#)
[Film music composer born 1965](#)
[What connects Innsbruck and Leipzig](#)
[People being 1.80m tall](#)
[List of Web browser software](#)
[Mayors of US cities higher than 1000m](#)
[Scientists and their doctoral advisors](#)

Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants

[Modify this query or create your own!](#)

Click on a column header to sort results on this page. Results: 10

10 results found in 0.001s

Nr.	?player	?club	?country	>40000	>10000000
1	Cicinho	Real Madrid	Brazil	80354	187560000
2	Gonzalo Fierro	Colo-Colo	Chile	62000	16432674
3	Lukas Podolski	FC Bayern Munich	Poland	69901	38536869
4	Mark González	Liverpool F.C.	South Africa	45362	47432000
5	Michael Thurk	Eintracht Frankfurt	Germany	52000	82438000
6	Ramón Morales	Chivas de Guadalajara	Mexico	72480	107784179
7	Robin van Persie	Arsenal F.C.	Netherlands	60432	16336346
8	Stefano Mauri	S.S. Lazio	Italy	82656	58751711
9	Stein Huysegems	Feyenoord Rotterdam	Belgium	51177	10419000
10	Tarek El-Said	Al-Ahly	Egypt	74100	78887007

Done Disabled GP 1 Error auto: en-US

Eksempel: Strukturerede opslag

Semantic Wikipedia Query: Sitcoms set in NYC - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://wikipedia.aks.wiki.org/index.php?qid=18

Google

Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | [Most popular](#) | [Upcoming](#)

[Tennis players from Moscow](#)
[Pictures of Clint Eastwood's Movies](#)
[Sitcoms set in NYC](#)
[Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants](#)
[People influenced by Friedrich Nietzsche](#)
[Films longer than 5 hours](#)
[Space Missions](#)
[Film music composer born 1965](#)
[What connects Innsbruck and Leipzig](#)
[People being 1.80m tall](#)
[List of Web browser software](#)
[Mayors of US cities higher than 1000m](#)
[Scientists and their doctoral advisors](#)
[Pictures of American guitarists](#)
[Battles in Saxony](#)

<< 1 >>

More Information: at [dbpedia.org](#) and in the paper [What have Innsbruck and Leipzig in common? Extracting Semantic from Wiki Content.](#)

Sören Auer and Jens Lehmann//--> [AKSW Workgroup](#) @ BIS / Universität Leipzig //-->

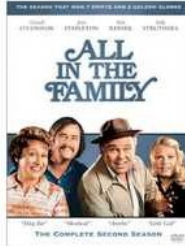
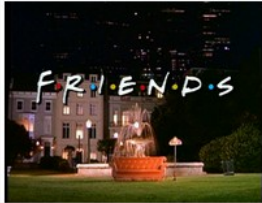
Sitcoms set in NYC

[Modify this query or create your own!](#)

Click on a column header to sort results on this page.

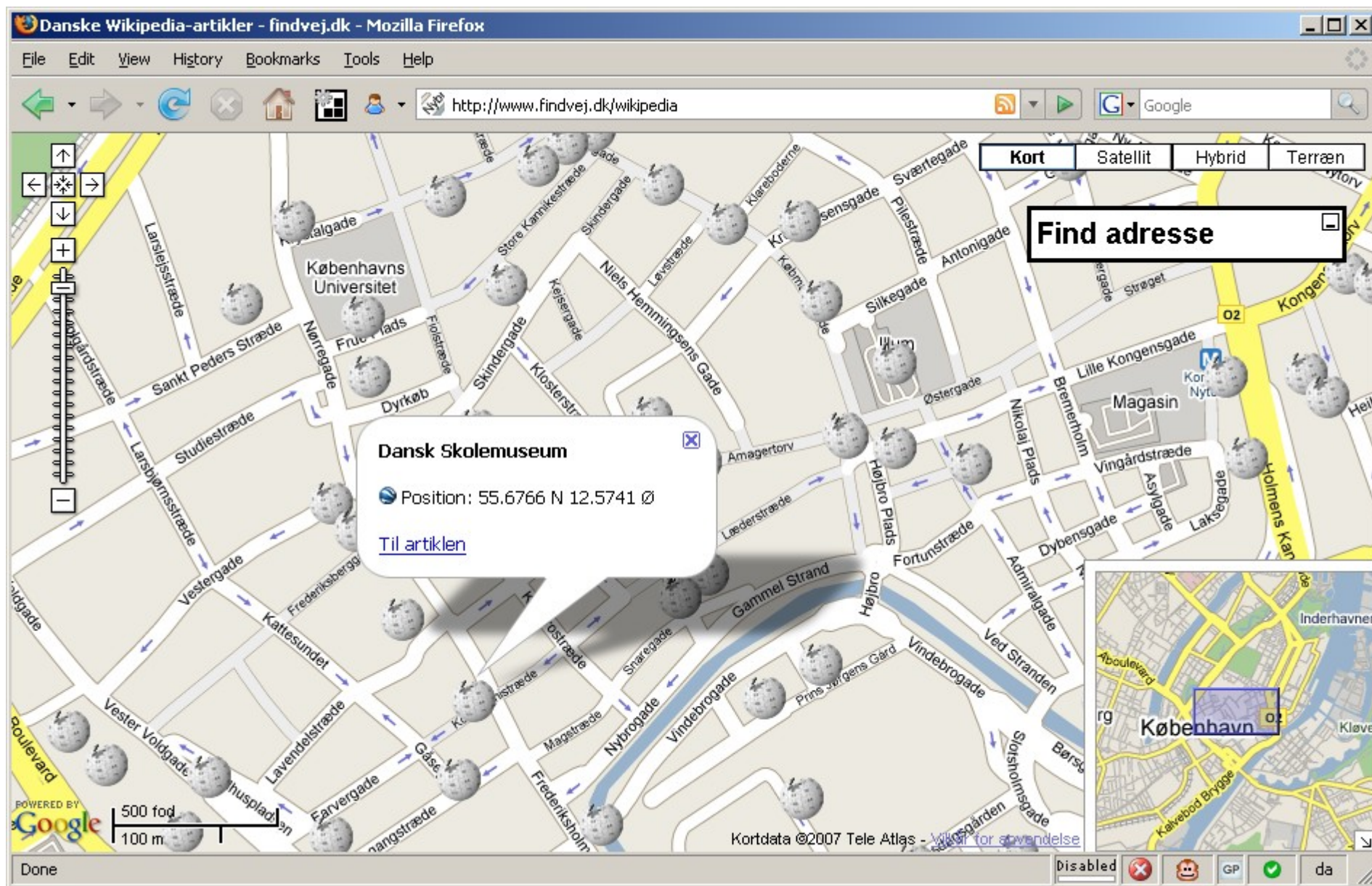
Results: 10

27 results found in 0.001s

Nr.	?show	?image
1	All in the Family	
2	Barney Miller	
3	Becker (TV series)	
4	Diff'rent Strokes	http://upload.wikimedia.org/wikipedia/commons/8/80/differentstrokes.jpg
5	Everybody Loves Raymond	http://upload.wikimedia.org/wikipedia/commons/3/3e/everybodylovesray.jpg
6	Friends	

Done Disabled GP 1 Error auto: en-US

Eksempel: Wikipedia-kort



Eksempel: Wikipedia-kort



Eksempel: Wikipedia-scanner

Wikipedia - scanner for den danske udgave - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://wikipedia.ter.dk/scanner?netname=ministeriet

Scanner for den danske Wikipedia - Til forsiden

Søg efter firmanavn/organisation:	Udvalgte firmaer/netværk:	IP-adresse:	IP-range:	Wikipedia-brugernavn
ministeriet	== Vælg fra listen ==			

Eksempel: 147.29.32.0-147.29.33.255

Tjekker efter "ministeriet" i registrerede netværk:

Viser 1-50 ud af 80
Side: 1 2

ip-range	ændringer	navn på netværk	beskrivelse
194.255.107.64-194.255.107.127	47	UDENRIGSMINISTERIET	Foreign Ministry in Denmark
87.54.19.0-87.54.19.15	20	KULTURMINISTERIET-NET	1780 Koebenhavn V
147.29.215.0-147.29.215.255	10	IndenrigsOgSundhedsministeriet	IndenrigsOgSundhedsministeriet
217.157.46.160-217.157.46.167	6	SKATTEMINISTERIET5-DSL-DK	2920 Charlottenlund
217.10.22.0-217.10.22.39	5	KM-DK	Kirkeministeriet
195.192.252.0-195.192.252.255	1	UNDERVISNINGSMINISTERIET	the Router Network for the Ministry of Education in Denmark
147.29.150.192-147.29.150.223	1	Skatteministeriet	Skatteministeriet
85.89.224.128-85.89.224.191	1	TM-DK	Trafikministeriet
83.91.91.104-83.91.91.111	1	KULTURMINISTERIET-NET	1780 Koebenhavn V
193.3.60.0-193.3.61.255	0	LFMSDDK	Landbrugs- og Fiskeriministeriet
193.163.80.0-193.163.83.255	0	TRAFIKMIN-DK	Trafikministeriet
195.249.139.192-195.249.139.207	0	STATSMINISTERIET-NET	1218 Koebenhavn K

Done Disabled GP auto: da

Eksempel: Wikipedia-scanner

Wikipedia - scanner for den danske udgave - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://wikipedia.ter.dk/scanner?iprange=194.182.139.48-194.182.139.63

Scanner for den danske Wikipedia - Til forsiden

Søg efter firmanavn/organisation:	Udvalgte firmaer/netværk:	IP-adresse:	IP-range:	Wikipedia-brugernavn
	Egmont Entertainment		194.182.139.48-194.182.139.63	

Eksempel: 147.29.32.0-147.29.33.255

Tjekker 194.182.139.48 - 194.182.139.63:

Viser 1-12 ud af 12:

ip	title	revision	comment	time
194.182.139.62	Tivi Magnusson ^[oversigt]	1604145		2007-09-21 16:57:19
194.182.139.62	Kim Magnusson ^[oversigt]	1604130		2007-09-21 16:52:49
194.182.139.62	Jacob Houind ^[oversigt]	1546038		2007-08-21 14:28:55
194.182.139.62	Jacob Houind ^[oversigt]	1546035		2007-08-21 14:28:15
194.182.139.62	Jacob Houind ^[oversigt]	1546033		2007-08-21 14:27:30
194.182.139.62	G-streng ^[oversigt]	1340919		2007-05-07 10:01:04
194.182.139.62	Frankrig ^[oversigt]	1302166		2007-04-20 09:34:05
194.182.139.62	Eton College ^[oversigt]	1300651		2007-04-19 13:32:27
194.182.139.62	Eton College ^[oversigt]	1300644		2007-04-19 13:31:33
194.182.139.62	Tivi Magnusson ^[oversigt]	1015923		2006-12-06 15:50:13
194.182.139.62	Olsen-banden ^[oversigt]	145296	/* Ekstern henvisning */	2004-11-17 11:29:46
194.182.139.62	Olsen-banden ^[oversigt]	139623	/* Ekstern henvisning */	2004-11-17 11:29:02

[Læs mere om scanneren og resultaterne.](#)

Done Disabled GP auto: da

Eksempel: Wikipedia-scanner

Wikipedia - scanner for den danske udgave - Mozilla Firefox

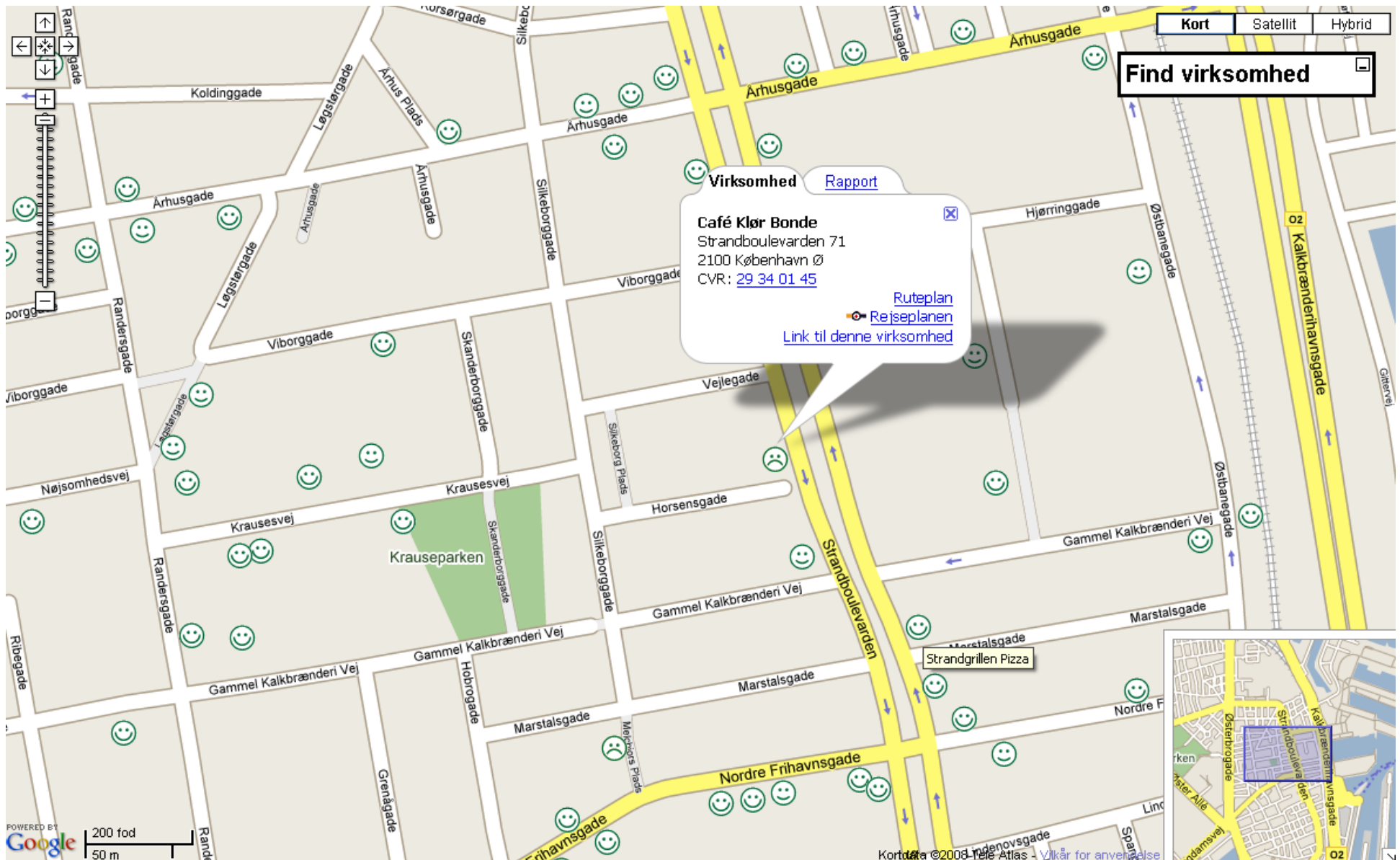
File Edit View History Bookmarks Tools Help

http://wikipedia.ter.dk/scanner?article=Jeppe+Nybroe

193.88.88.10	TV2 TV2 / Danmark	1568673	Fyret fra DR	2007-09-03 18:50:53
		Rune X2 1568462	/* Manipulation af lydspor */	2007-09-03 15:31:34
		Rune X2 1568460		2007-09-03 15:30:30
		Rune X2 1568416		2007-09-03 15:09:57
193.88.88.10	TV2 TV2 / Danmark	1560550	indslag fra tv-avisen . d 29 aug	2007-08-30 10:00:38
193.88.88.10	TV2 TV2 / Danmark	1560538	blog gendannet på dr.dk	2007-08-30 09:52:03
193.88.88.10	TV2 TV2 / Danmark	1558998		2007-08-29 09:45:35
193.88.88.10	TV2 TV2 / Danmark	1558996	blog slettet	2007-08-29 09:45:16
	Villy	1530403	wikistandard	2007-08-11 20:43:20
	Fink			
	Isaksen			
	Nico	1530400	gendanner efter fejlsletning	2007-08-11 20:38:56
	Nico	1530395	Gendannelse til seneste version ved [[Bruger:81.19.246.10 81.19.246.10]], fjerner ændringer fra [[Bruger:80.199.162.140 80.199.162.140]] ([[Speciel:Contributions/80.199.162.140 bidrag]])	2007-08-11 20:33:52
80.199.162.140		1530213	nye oplysninger	2007-08-11 17:47:42
80.199.162.140		1529961	henvisning til citat	2007-08-11 12:35:00
81.19.246.10	JAYNET-DK-DIX-POLITIKEN A/S Dagbladet Politiken	1524547		2007-08-08 08:11:47

Done Disabled GP auto: da

Findvej.dk/smiley



Biblitekspenge.dk

[Forside](#) [Alfabetisk oversigt](#) [Årsoversigter](#) [Søg efter p](#)

Josefine Ottesen

(Alice J Ottesen Schønemann)

2007	375.647,87 kr. (28)
2006	263.537,85 kr. (61)
2005	187.020,74 kr. (108)
2004	146.532,24 kr. (142)
2003	103.235,73 kr. (221)
2002	93.467,05 kr. (316)
2001	84.875,46 kr. (336)
2000	65.230,00 kr. (445)
1999	54.812,55 kr. (525)
1998	49.529,80 kr. (556)

[Annoncer fra Google](#) [Tekst Forfatter](#) [HC Andersen Eventyr](#) [Blive Forfatter](#)

Mere info (fra Wikipedia):

Josefine Ottesen (4. februar 1956), uddannet skuespiller og dramaturg. Debuterede omkring tres titler for børn og unge.

Har modtaget adskillige nationale priser og har været nomineret til flere internationale.

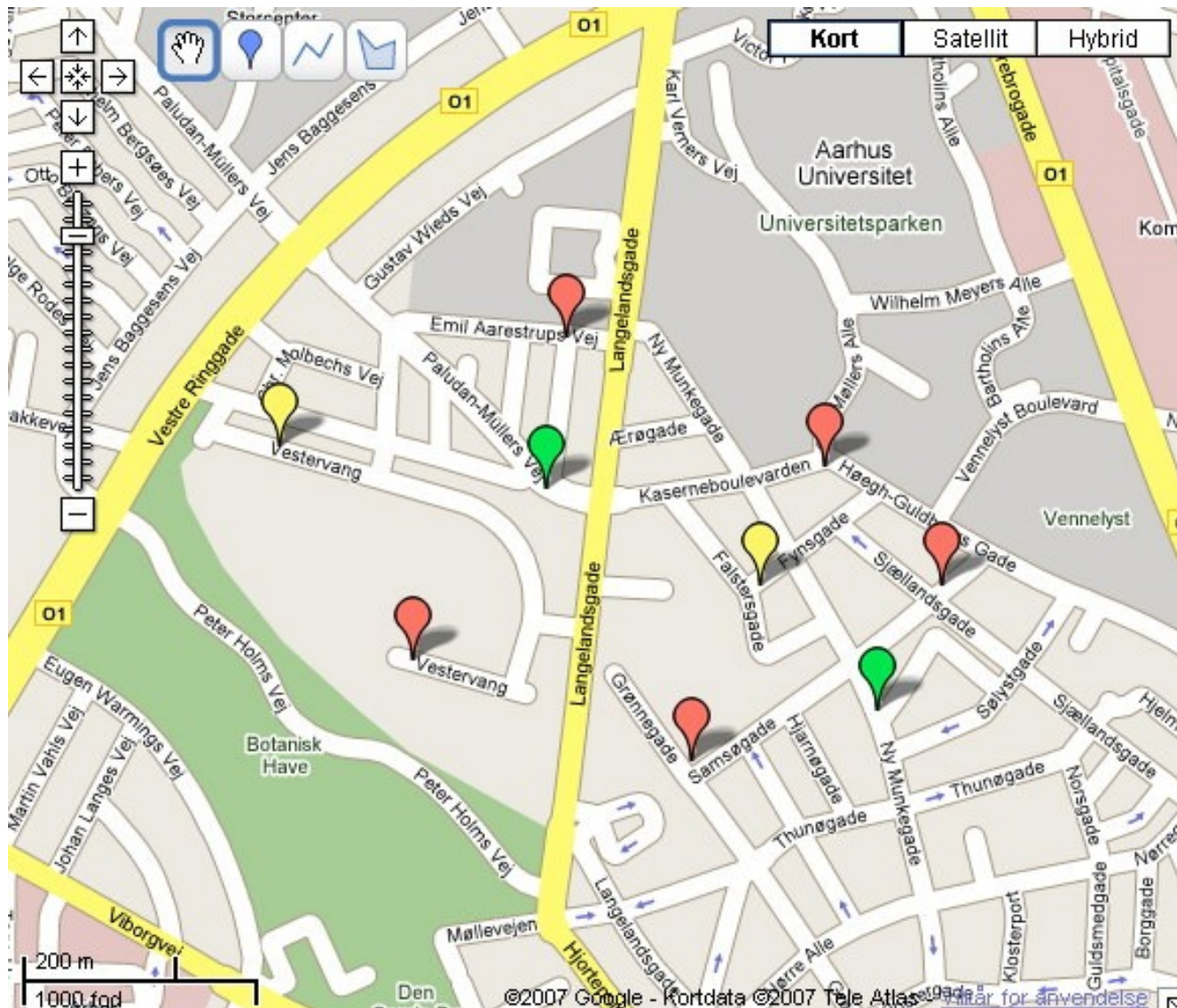
Litteraturliste

- *Prinsesse Morgenrøde og Lindormen*, Mallings 1983
- *Eventyret om Fjeren og Rosen*, Mallings 1986/Carlsen 1998
- *Eventyret om Påfuglen, der ville synge*, Mallings 1987
- *Eventyret om Drengen og Morgenstjernen*, Mallings 1988
- *Dragens Kys*, Mallings 1994/Høst og Søn 1999
- *Roselil og hendes venner*, Mallings 1995/Sesam 2006
- *Roselil og Blomsterkrigen*, Mallings 1996/Sesam 2006
- *Roselil og den lille dragen*, Mallings 1997/Sesam 2006

Højdespringere i 2006

1	Morten Ramsland	82.366,83 kr. (305)
2	Christian Jungersen	78.410,78 kr. (176)
3	Josefine Ottesen	76.517,11 kr. (61)
4	Sara Blædel	56.947,62 kr. (357)
5	John Jørgensen	47.626,11 kr. (122)
6	Elsebeth Egholm	45.399,75 kr. (102)
7	Jonna Løvdahl Gensbøl	44.093,13 kr. (601)
8	Hanna Louisa Lützen	41.897,20 kr. (27)
9	Jørn Martin Jensen	40.454,66 kr. (46)
10	Susanne Torpe	34.367,41 kr. (545)
11	Mette Skot Perschke	31.402,34 kr. (150)
12	Henrik Einspor	29.821,00 kr. (157)
13	Jan Sneum	28.457,88 kr. (219)
14	Anders Johansen	26.995,65 kr. (47)
15	Kim Langer	24.962,41 kr. (130)
16	Kamilla Jørgensen	23.851,31 kr. (599)
17	Lene Bodil Schjøtt	23.687,32 kr. (577)
18	Gretelise Holm	22.985,19 kr. (100)
19	Mich Vraa	22.420,14 kr. (55)
20	Anne Marie Løn	21.762,70 kr. (35)
21	Lene Kaaberbøl	21.041,07 kr. (20)
22	Klavs Brøndum	21.015,27 kr. (240)
23	John Nehm	20.356,66 kr. (85)
24	Louis Jensen	19.350,53 kr. (71)
25	Daniel Zimakoff	19.033,29 kr. (236)

Geotagging af artikler



Artikler med geografisk information

Savnet ældre kvinde fundet død i å

Den 71-årige kvinde fra Grimstrup øst for Esbjerg, som politiet har ledt efter siden natten til mandag den 14. januar, er i dag blevet fundet død

Ritzau /112 - 13:29 - 25. jan. 2008

Send Print

Den 71-årige kvinde fra Grimstrup øst for Esbjerg, der forlod sit hjem i nedtrykt stemning natten til mandag den 14. januar, er i dag blevet fundet død. Det skriver Jydske Vestkystens netavis

Kvinden er fundet i Sneum Å - cirka 10 kilometer fra det sted, hvor hun sidst blev set i live.

- Vi genoptog vores eftersøgning i går på grund af den faldende vandstand. I går uden resultat, men i formiddags fandt folk fra Beredskabsstyrelsen hende i åen, oplyser vagtchef Anders Hansen hos Syd- og Sønderjyllands Politi.

De pårørende til den ældre kvinde er orienteret.



Kvinden blev fundet i Sneum Å. (Arkivfoto: Claus Bonnerup)

LÆS OGSÅ ...

[Leder efter liqet i åen](#)

.. hver eneste dag ..

Tank udsat for pistolrøveri

En Shell-tank i Ringsted blev klokken fire fredag morgen udsat for væbnet røveri, da en ekspedient blev truet med pistol til at udlevere 2000 kroner

Lisbeth K. Larsen - 10:19 - 25. jan. 2008

Send Print

En tankrøver stak tidligt fredag morgen af med 2000 kroner fra en Shell-tank i Ringsted, da han havde truet den 40-årige ekspedient med en pistol.

Gerningsmanden forsvandt efterfølgende fra stedet til fods. Der gik dog en halv time fra, at røveriet havde fundet sted, før politiet blev alarmeret.

- Ekspedienten er ikke blevet ordentligt informeret om alarmproceduren, så hun trykkede på en knap, men det var ikke overfaldsknappen, derfor vidste hun ikke, at hun skulle tilkalde politiet. Så da vi nåede frem var gerningsmanden stukket af, siger Carsten Andersen fra Midt- og Vestsjællands Politi til ekstrabladet.dk.

Politiet vil gerne høre fra vidner til røveriet eller personer, der ud fra signalementet kan genkende gerningsmanden. Man kan kontakte politiet på 114 eller 46351448 hvis man har oplysninger i sagen.



En Shell-tank i Ringsted var tidligt fredag morgen udsat for røveri, efter en ung mand havde truet sig til 2000 kroner med en pistol. (Foto: Paul Sakuma)

LÆS OGSÅ ...

- [Tankpasser truet med pistol i Voje](#)
- [Ny tankrøver på spil i København](#)
- [Ældre kvinde sparkede pistolrøver](#)

EKSTRA FACTS

Signalment

Gerningsmanden beskrives som i starten af 20'erne, 165-170 centimeter høj, iført fleecetrøje med lyseblåt på brystet, blå cowboybukser og kondisko. Manden var maskeret med en hue med huller til øjnene.

.. med adresser i teksten

Mor til tre dræbt i trafikulykke

En 35-årig mor til tre mistede torsdag livet, da hun kørte ind i et vejtræ på Aggersvoldvej i Jyderup

Lisbeth K. Larsen - 08:39 - 25. jan. 2008

Send Print

Tre børn mistede torsdag eftermiddag deres mor, da hun kørte ind i et vejtræ på Aggersvoldvej i Jyderup og mistede livet.

Kvinden mistede herredømmet over bilen i en højrekurve, hvor hendes venstre hjulpar kom over i rabatten i den modsatte kørebane, og hun ramte ind i træet.

- Hun kørte uden sikkerhedssele, og hun kunne muligvis være reddet, hvis hun havde haft sikkerhedssele på, siger Carsten Andersen fra Midt- og Vestsjælland Politi til ekstrabladet.dk.

Kvinden var 35 år og alenemor til tre. Politiet måtte efter ulykken hente kvindes 3 børn, en datter på 13 år, en 10-årig søn og en 5-årig datter, der alle var i deres respektive institutioner. Holbæk Kommune er gået ind i sagen og sørgede for psykologhjælp. Børnene er overgivet til den dræbtes broder.



En 35-årig mor til tre mistede torsdag eftermiddag livet, da hun kørte ind i et vejtræ i Jyderup. (Foto: Colourbox)

LÆS OGSÅ ...

[69-Årig dræbt i trafikulykke](#)

[19-Årig dræbt i vanvidsoverhaling](#)

Hvordan opdrager vi journalister?

- Indtastning af koordinater
- Klik på kort
- Indtastning af adresser
- GPS i bilen
- GPS i kamera (EXIF-data – ikke for arkivfotos)
- Søgning efter adresser i brødtekst (regulære udtryk)

Match vejnavne

- Regulære udtryk
- Et eller flere ord med stort begyndelsesbogstav (og evt. tal)
- Aggersvoldvej
 - København
 - Holbæk/Mørkøv
- Holbæk Kommune

Mor til tre dræbt i trafikulykke

En 35-årig mor til tre mistede torsdag livet, da hun kørte ind i et vejtræ på Aggersvoldvej i Jyderup

Lisbeth K. Larsen · 08:39 · 25. jan. 2008

Send Print

Tre børn mistede torsdag eftermiddag deres mor, da hun kørte ind i et vejtræ på Aggersvoldvej i Jyderup og mistede livet.

Kvinden mistede herredømmet over bilen i en højrekurve, hvor hendes venstre hjulpar kom over i rabatten i den modsatte kørebane, og hun ramte ind i træet.

- Hun kørte uden sikkerhedssele, og hun kunne muligvis være reddet, hvis hun havde haft sikkerhedssele på, siger Carsten Andersen fra Midt- og Vestsjælland Politi til ekstrabladet.dk.

Kvinden var 35 år og alenemor til tre. Politiet måtte efter ulykken hente kvindes 3 børn, en datter på 13 år, en 10-årig søn og en 5-årig datter, der alle var i deres respektive institutioner. Holbæk Kommune er gået ind i sagen og sørgede for psykologhjælp. Børnene er overgivet til den dræbtes broder.



En 35-årig mor til tre mistede torsdag eftermiddag livet, da hun kørte ind i et vejtræ i Jyderup. (Foto: Colourbox)

LÆS OGSÅ ...

69-årig dræbt i trafikulykke

19-årig dræbt i vanvidsoverhaling

Kritisk masse

- Hold journalistens interesse ("why bother?")
- Først interessant efter den kritiske masse
- Korrelationer i kort
- Teknikken kan bruges med tilbagevirkende kraft
- Hurtigt at geotagge nogle hundrede artikler
- 100 % er næppe nødvendigt

Sammenkobling med eksterne data

- Business Intelligence Light
- Kundekartotek:
 - Hvornår blev kunden oprettet?
 - Hvilken branche tilhører kunden?
 - Hvor mange pænge har de lagt?
 - CVR/Virk/WebReg/Erhvervs- og Selskabsstyrelsen
 - Antal medarbejdere
 - Årsregnskab

Tak for opmærksomheden!

Spørgsmål?