Finn Årup Nielsen,
faan@dtu.dk

Wikidata og Scholia

Det redigerbare Linked Open Data med bibliografiske data

Er det muligt at lave en database, hvor metadata over alverdens udgivelser er frit tilgængelig? Og hvor flersproglige komplekse søgninger, der baserer sig på en stor ontologi, er mulige? Vores Scholia-projekt, der hviler på Wikidata og Linked Open Data, forsøger det.

Scholia er som en web-applikation frit tilgængelig og lader f.eks. brugeren identificere centrale artikler og eksperter inden for et emne eller se oversigter over software og kemikalier benyttet i forbindelse med artikler. Biblioteker ville kunne benytte den åbne data, der vises i Scholia, til at berige deres egne data eller søgeoplevelsen på deres egne services.

Wikidata og bibliografisk data

Wikidata er Wikipedias mindre kendte lillesøster og ligeledes en wiki, hvor enhver kan redigere indholdet. Det, der adskiller Wikidata fra den gængse wiki, er, at indholdet er langt mere struktureret. Det er således kun muligt at indtaste information i fastlagte felter. Disse felter kan indeholde alt fra en persons fødselsdato, et lands regeringsleder, dets befolkningstal, antal ben på et dyr, vægt, højde, et kunstværks inventarnummer osv.

Data fra Wikidata bliver konverteret til et format kompatibelt med Semantisk Web-teknologierne og er således en del af Linked Open Data. Fonden, der driver Wikidata, har sat en Semantisk Web-søgemaskine op, som muliggør komplekse forespørgsler i sproget SPARQL, som f.eks. "byer over 1 million indbyggere med en kvindelig borgmester." Al Wikidatas indhold er åbne data og hele wikiens indhold kan – som Wikipedia – hentes i store dumpfiler.

Som for Wikipedia er det ikke selve Wikimedia Fonden, der står som dem, der tilføjer information til Wikidata. Det er en broget skare af individer og grupper på internettet. Man kan manuelt indtaste hver enkelt informationsenhed, men i Wikidatas økosystem er der en række værktøjer, der enten automatisk eller halvautomatisk kan benyttes til at tilføje information.

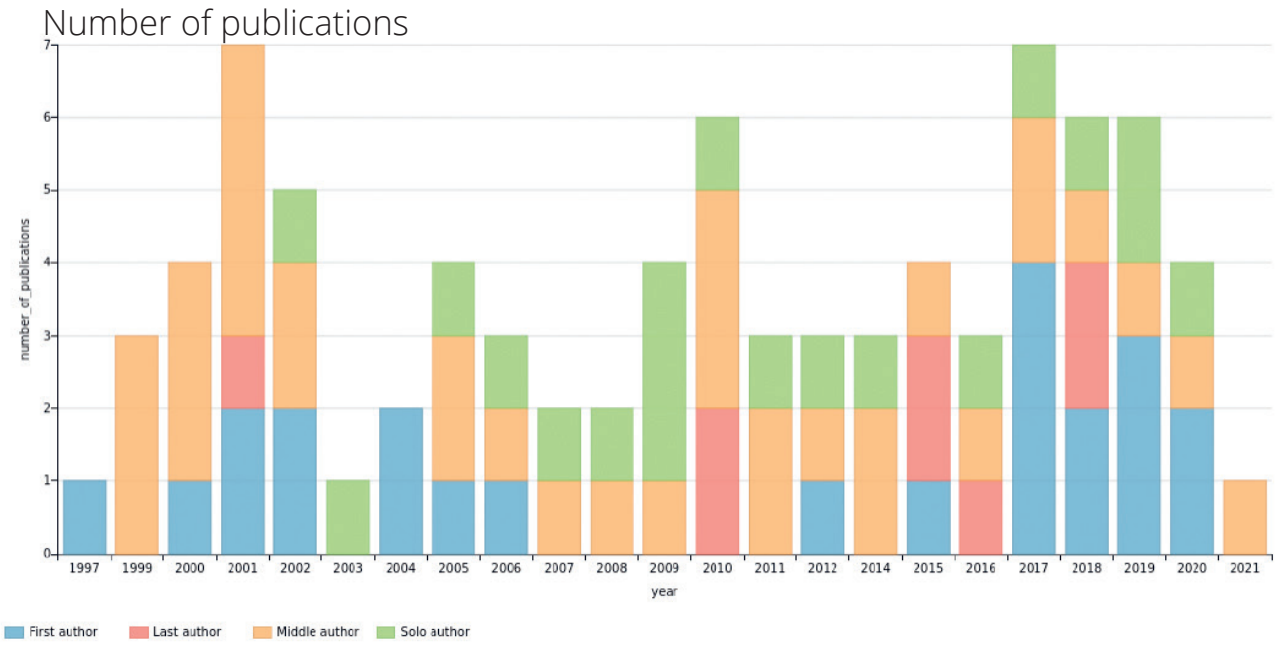
I en forskningsbiblioteksmæssig sammenhæng har især WikiCite-projektet været interessant, som fokuserer på wiki-projekternes kilder og betyder, at wiki-brugerne for alvor er begyndt at benytte Wikidata til at repræsentere bibliografiske data. Over 40 millioner udgivelser, de fleste videnskabelige artikler, er nu beskrevet i Wikidata, en vis del af disse er også beskrevet med citeringer og med emner. For mange af artiklerne gælder, at forfatterne er linket til artiklerne og beskrevet i Wikidata. Vores statistik fortæller os, at vi har over 23 millioner forfatterlinks og over 23 millioner emnelinks. Over 214.000 ISSN'er og cirka 1,8 millioner ORCID'er er også beskrevet.

Hvad kan det så bruges til?

Baseret på den bibliografiske data i Wikidata og den Semantisk Web-baserede søgemaskine har vi bygget en dynamisk hjemmeside under navnet Scholia, der sammenstiller bibliografisk data på en række forskellige måder. Søgemaskinen er i stand til ikke blot at returnere data som tekst og tal i tabelform, men også som linjeplot, boblеplot, grafer med mere. Det udnytter vi i Scholia, hvor vi blandt andet kan vise citeringsgrafer, medforfattergrafer og plot over antal artikler udgivet per år for en organisation. Opbygningen af en Scholia-side sker "live" i løbet af nogle sekunder. Den rige annotering i Wikidata tillader os at sammenstille data, der normalt ikke er tilgængelige i bibliografiske databaser. F.eks. kan vi umiddelbart skabe geografisk kort over lokaliteter, der er nævnt for et specifikt defineret emne. Egon Willighagen fra Maastricht Universitet har haft specielt fokus på kemisk information og skabt sider på Scholia, der sammenstiller kemiske og bibliografiske data fra Wikidata.

Scholia voksede ud fra en idé om at skabe en Wikidata-baseret forskerprofil. I modsætning til, hvad der kendes fra Google Scholar, Scopus, Pure, ResearchGate og forskernes egne hjemmesider, vil en sådan profil være frit licenseret, og det vil være muligt at kombinere den bibliografiske data med data fra andre dele af Wikidata.

Finn Årup Nielsen,
faan@dtu.dk



Figur 1: Antal udgivelser per år for Finn Årup Nielsen: Graf på en af Scholias sider: https://scholia.toolforge.org/author/Q20980928.

Efterhånden har vi udvidet Scholia, så vi udover forskerprofiler har profiler for organisationer, udgivere, konferencer, lande med mere. Vi er i stand til at vise de mest citerede personer tilknyttet en organisation. Med den geografiske information i Wikidata kan vi identificere lokale eksperter, f.eks. maskinlæringseksperter i Københavnsområdet. Med citeringsinformation kan vi ud fra en given artikel vise relaterede artikler. Vi har også benyttet Wikidata og Scholia som bibliografisk reference management-system til de videnskabelige artikler, vi har skrevet. Scholia kan også vise resultaterne af en sampublikationsanalyse on-the-fly.

Begrænsninger og videreudvikling

Selvom Wikidata er et ganske interessant projekt, har det også begrænsninger. De bibliografisk data i Wikidata er langt fra komplette, relativt få videnskabelige forfattere er registrerede, og de, der er, er ikke nødvendigvis linket til deres artikler, ligesom citeringsdata mangler i stort omfang. Wikimedia Fondens tidligere forskningschef Dario Taraborrelli søgte at råde bod på sidstnævnte begrænsning ved oprettelsen af initiativet I4OC, der har fået en længere række udgivere til at frigive deres citeringsdata, således at det kan repræsenteres i Wikidata. En række værktøjer udviklet af Magnus Manske hjælper Wikidata-brugere med at flytte data ind i Wikidata fra eksterne kilder. Her er DOI og ORCID af stor nytte. Åbne artikler fra Europe PMC er godt dækket.

Vi arbejder fortsat med at udvikle Scholia, og mængden af bibliografisk data stiger stadig. Wikidata-brugere har ivrigt kopieret åbne citeringsdata til Wikidata, og vi har over 285 millioner citeringer registreret. Daniel Mietchen, som er involveret i Scholias udvikling, har kopieret en stor mængde data til Wikidata, der er fokuserede redigeringer omkring Zika-virus. Det betyder, at videnskabelige artikler om Zika-virusset er grundigt repræsenteret i Wikidata, så Scholia kan vise en ganske bred oversigt over emnet.
En udfordring er artikler fra konferencer og tidsskrifter, der ikke benytter DOI eller ORCID. For disse artikler skal specielle programmer udvikles til at udtrække data fra deres hjemmesider.

Bag om Scholia

Scholia var i første omgang blot en enkelt hjemmeside med information om Finn Årup Nielsens udgivelser og cv-information.

Siden konverterede han Scholia til en Python web-applikation, så den blev generel og en hvilken som helst forfatter kunne vises.

Scholia blev sat op på en hjemmeside, og Finn Årup Nielsen flyttede koden til GitHub under en Open Source-licens. Da det var sat op, begyndte flere at hjælpe til med at udvide koden og promovere Scholia bredere.

Kernen af Scholia-gruppen består nu, foruden Finn Årup Nielsen, af Daniel Mietchen, Egon Willighagen og Lane Rasberry. Derudover har en række andre bidraget med udvidelser, og Scholia har fået generøs økonomisk støtte fra den amerikanske Sloan Foundation til at lønne folk.

Scholia kan i princippet køre på en hvilken som helst server, men fungerer godt på Toolforge – en cloud-infrastruktur som Wikimedia Fonden stiller gratis til rådighed for Wikimedia-relevante projekter.

Mere information
https://scholia.toolforge.org - Scholias hjemmeside

Finn Årup Nielsen, Daniel Mietchen og Egon Willighagen, "Scholia, Scientometrics and Wikidata", ESWC 2017, DOI 10.1007/978-3-319-70407-4_36

For links til andet materiale om Scholia se https://scholia.toolforge.org/about og Scholias side om Scholia selv https://scholia.toolforge.org/topic/Q45340488 der oplister artikler som beskriver brug af Scholia.

Figur 2: Emner som personer med tilknytning til forskningssektionen Kognitive Systemer på Danmarks Tekniske Universitet har udgivet i: Tabel på en af Scholias sider: https://scholia.toolforge.org/topic/Q24283660

Topics that employees and affiliates have published on

Topics of publications by past and present employees, affiliates, and members, ranked by number of individuals having published on the topic.

Show 10 entries

Search:

Researchers	Topic	Zoom	Topic description	Samplework
22	electroencephalography		electrophysiological monitoring method	The Smartphone Brain Scanner: A Portable Real-Time Neuroimaging System
18	machine learning		scientific study of algorithms and statistical models that computer systems use to perform tasks without explicit instructions	Archetypal analysis for machine learning and data mining
18	smartphone		multi-purpose mobile computer	The Smartphone Brain Scanner: A Portable Real-Time Neuroimaging System
18	brain		organ that serves as the center of the nervous system in all vertebrate and most invertebrate animals	Similar brain networks for detecting visuo-motor and visuo-proprioceptive synchrony
17	functional magnetic resonance imaging		MRI procedure that measures brain activity by detecting associated changes in blood flow	On clustering fMRI time series
15	deep learning		branch of machine learning	Inferring visual semantic similarity with deep learning and Wikidata: Introducing imagesim-353
14	Bayesian statistics		theory in the field of statistics	Bayesian correlated component analysis for inference of joint EEG activation
12	independent component analysis		in signal processing, a computational method	Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG
12	neuroimaging		set of techniques to measure and visualize aspects of the nervous system	Finding related functional neuroimaging volumes
11	Gaussian process		particular kind of statistical model where observations occur in a continuous domain, e.g. time or space. In a Gaussian process, every point in some continuous input space is associated with a normally distributed random variable	Heterogeneous Multi-output Gaussian Process Prediction

Wikidata Query Service organization: topics.sparql

Showing 1 to 10 of 500 entries

Previous 1 2 3 4 5 … 50 Next

I Scholia er der værktøjer til at udtrække metadata repræsenteret i tidsskrifter, der benytter Open Journal Systems, og det er i nogen udstrækning benyttet til artikler på tidsskrift.dk.

Biblioteker kan bidrage til wiki-verden, dels ved at benytte den åbne information i Wikidata og dels ved at stille metadata til rådighed. I det omfang bibliotekerne stiller deres autoritetsdata åbent til rådighed, kan det blive inkluderet i Wikidata, hvor Wikidata så kommer til at virke som en hub mellem forskellige autoritetsdatabaser. Der er allerede mange databaser, der er linket fra Wikidata, og bibliotekerne kan f.eks. benytte det til at holde rede på tvetydige forfatternavne.