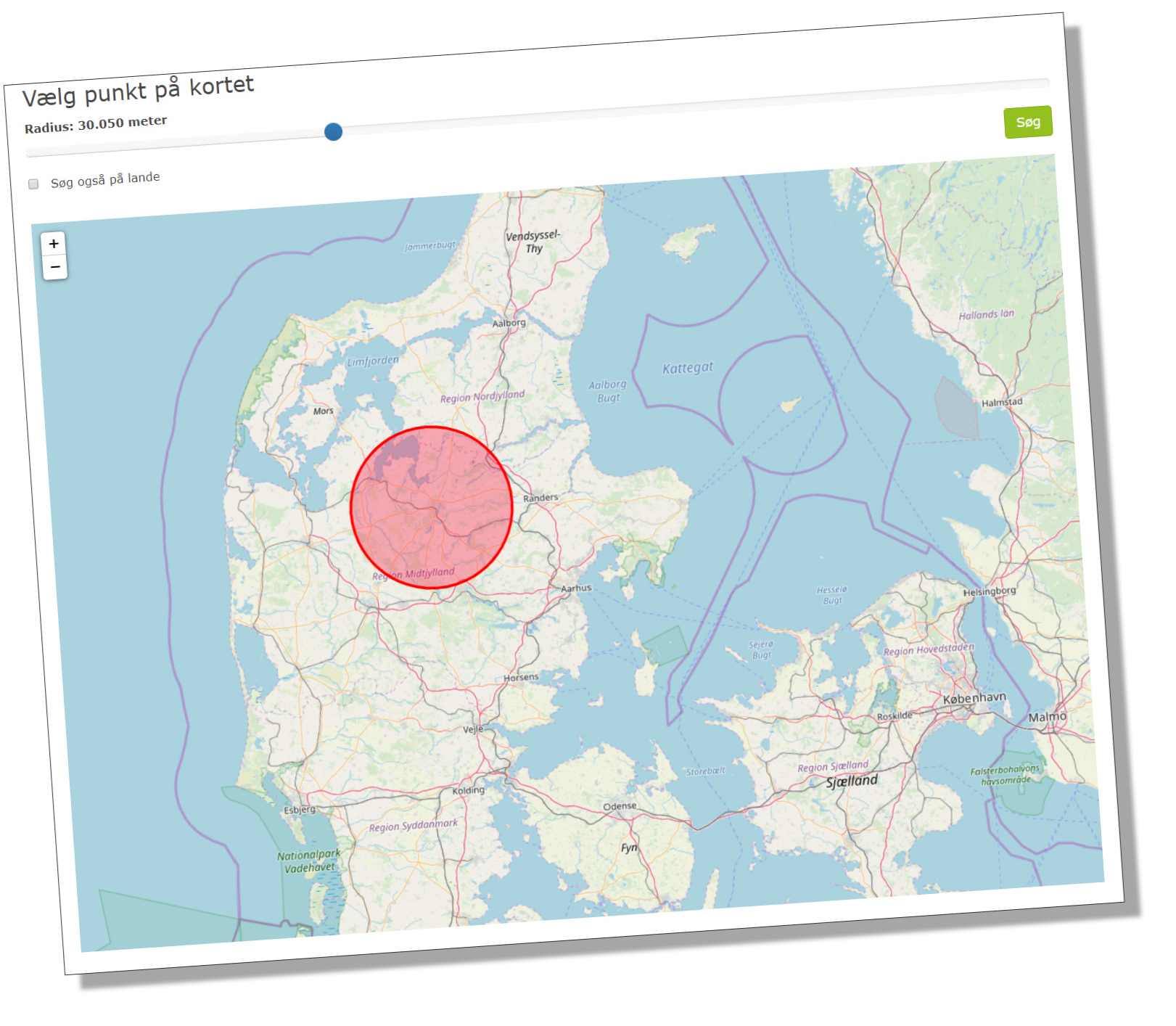
**Forfattere:** Thomas Rosgaard Thomas Høj Thykjær

**Vejleder:** Kristian Torp

**Uddannelse:** Master i IT, på linjen Softwarekonstruktion, Aalborg Universitet, Cassiopeia, Institut for Datalogi

**Dato:** Januar 2020

**Titel: Spatio-tekstuel integration -** Integration af eksterne datakilder med interne datakilder i Microsoft SQL Server 2017 for at give slutbrugere nye muligheder for spatiale søgninger



**Indholdsfortegnelse**

[Abstract 7](#_TOC_250069)

[Resumé 8](#_TOC_250068)

[Termer 9](#_TOC_250067)

1. [Indledning 11](#_TOC_250066)
   1. [Problemformulering 12](#_TOC_250065)
   2. [Arbejdsspørgsmål 12](#_TOC_250064)
2. [Metode 13](#_TOC_250063)
   1. [Teknologivalg 13](#_TOC_250062)
      1. [Database 13](#_TOC_250061)
         1. [Spatiale datatyper 14](#_TOC_250060)
      2. [Programmeringssprog 14](#_TOC_250059)
      3. [Kortmaterialer 14](#_TOC_250058)
      4. [Dataformater 14](#_TOC_250057)
   2. [Afgrænsninger 15](#_TOC_250056)
   3. [Fremgangsmåde 15](#_TOC_250055)
3. [Redegørelse og teori 16](#_TOC_250054)
   1. [Redegørelse 16](#_TOC_250053)
      1. [Tabeller 17](#_TOC_250052)
         1. Tabellen MARCRecord 17
         2. Tabellen BiblRecord 18
         3. Cache-tabeller 19
         4. Tabellen BiblRecordCacheTitle 19
         5. Tabellen BiblRecordCacheBeskrivelse 19
         6. Tabellen BiblRecordCachePaedNoter 20
         7. Datamængder 21
      2. Danmarkskort 21
      3. Verdenskort 24
      4. Hardware 24
   2. Teori 24
      1. Indlæsning af data fra disk 25
      2. Indeksering af data 25
      3. B+ træer (og B træer) 28
      4. Full-text-indeks 30
      5. Spatiale indeks 33
         1. Tessellation 35
         2. Grids til lineære B-træer 35
         3. Space-division vs. data-division 36
      6. XML Indeks 36
         1. Primære XML indeks 36
         2. Sekundære XML indeks 37
         3. Selektive XML indeks 37
4. [Analyse 39](#_TOC_250051)
   1. [Indlæs kortdata i databasen 39](#_TOC_250050)
      1. [Kontrol af importerede data 40](#_TOC_250049)
      2. [Forsøg med nyt kortmateriale 41](#_TOC_250048)
      3. [Oprettelse af spatiale indeks 41](#_TOC_250047)
   2. [Spatio-tekstuel integration 42](#_TOC_250046)
      1. [Direkte integration 42](#_TOC_250045)
         1. [Oprettelse af “BiblRecordCacheSpatial” tabel 43](#_TOC_250044)
      2. [Begrænsning i forhold til konteksten 44](#_TOC_250043)
         1. [Begrænsning med negativ-liste 44](#_TOC_250042)
         2. [Sproglige udfordringer i full-text-søgningen 46](#_TOC_250041)
      3. [Full-text-søgning med neutralt sprog 48](#_TOC_250040)
      4. [Indekserede views 49](#_TOC_250039)
      5. [Integration med fuldt OSM-datasæt for Danmark 51](#_TOC_250038)
      6. [Yderligere begrænsninger 51](#_TOC_250037)
         1. [Positiv-liste 52](#_TOC_250036)
         2. [Geografiske navne på et enkelt ord 54](#_TOC_250035)
         3. [Numeriske navne 54](#_TOC_250034)
         4. [Fjern navne-dubletter over typer 55](#_TOC_250033)
         5. [Spatialt afstandskrav 55](#_TOC_250032)
         6. [Resultater af begrænsninger 57](#_TOC_250031)
         7. [Eksakte case-sensitive søgninger 58](#_TOC_250030)
      7. [Verdens lande 60](#_TOC_250029)
   3. [Opsummering af den spatio-tekstuelle integration 61](#_TOC_250028)
   4. [Test af integrationen 62](#_TOC_250027)
   5. [Udvidelse af integrationen 64](#_TOC_250026)
      1. [Udvidelse af spatiale data 64](#_TOC_250025)
      2. [Udvidelse af tekstuelle data / metadata 66](#_TOC_250024)
         1. [Valg af dataformat 67](#_TOC_250023)
         2. [XML-schema til undertekster 68](#_TOC_250022)
         3. [Indeksering af undertekster 69](#_TOC_250021)
         4. [Selektive XML indeks 71](#_TOC_250020)
         5. [Inklusion af undertekster i den spatio-tekstuelle integration 72](#_TOC_250019)
      3. [Opsummering af udvidelser 73](#_TOC_250018)
   6. [Frontend prototype 73](#_TOC_250017)
      1. [Grafisk brugerflade 74](#_TOC_250016)
      2. [Spatiale søgninger 75](#_TOC_250015)
      3. [Indeks performance 77](#_TOC_250014)
5. [Diskussion 79](#_TOC_250013)
   1. [MSSQL og spatiale data 79](#_TOC_250012)
   2. [Full-text-søgninger 79](#_TOC_250011)
   3. [Begrænsning af spatiale data 79](#_TOC_250010)
   4. [Målinger 80](#_TOC_250009)
   5. [Generel anvendelse 80](#_TOC_250008)
6. [Konklusion 81](#_TOC_250007)
7. [Perspektivering 83](#_TOC_250006)
8. [Bilag 84](#_TOC_250005)
   1. [danMARC2 format eksempel 84](#_TOC_250004)
   2. [Sammenligning og kontrol af importerede data 85](#_TOC_250003)
   3. [Direkte integration 86](#_TOC_250002)
   4. [Oprettelse af indexed views og full-text-index 91](#_TOC_250001)
   5. [Stored Procedure - LimitSearch 93](#_TOC_250000)
   6. Stored Procedure - InsertSubtitle 107
   7. Stored Procedure - InsertIntoBiblRecordCacheSpatial 108
   8. Test af spatio-tekstuel integrationsproces 119
      1. Test med “Nordkraft” 119
      2. Test med “Hans og Kirsten i 2. verdenskrig” 120
      3. Test med “Kulturkanon” 123
      4. Test med “Den arabiske verden” 125
   9. Positivlist - fuld klassifikation 129
   10. Kildekode til import af undertekster 137
   11. Liste over filer 139
9. Litteraturliste 140
   1. Bøger 140
   2. Hjemmesider 140

**Tabelfortegnelse**

T3.A: Et uddrag fra MARCRecord-tabellen for materialet “Danske naturrekorder” 18

T3.B: Uddrag fra BiblRecord-tabellen 18

T3.C: Et uddrag fra BiblRecordCacheTitle-tabellen for materialet “Danske naturrekorder” 19

T3.D: Et uddrag fra BiblRecordCacheBeskrivelse-tabellen for materialet “Danske naturrekorder” 19

T3.E: Beskrivelse for materialet “Danske naturrekorder” 20

T3.F: Et uddrag fra BiblRecordCachePaedNoter-tabellen for materialet “Danske naturrekorder” 20

T3.G: Et uddrag fra BiblRecordCacheEmneord-tabellen for materialet “Danske naturrekorder” 20

T3.H: Antal rækker i cache-tabeller 21

T3.I: Beskrivelse af de enkelte filer i OSM Danmarkskortet fra GeoFabric.de 22

T3.J: Uddrag fra gis\_osm\_pois\_free\_1.DBF 23

T3.K: Uddrag fra gis\_osm\_pois\_free\_1.DBF 24

T3.L: Simpel tabel med ustruktureret tekst 31

T3.M: Inverted index til FTS på vores fiktive tabel T3.L 31

T4.A: Uddrag fra data importeret med Shape2SqlServer 40

T4.B: Uddrag af BiblRecordCacheSpatial cache-tabellen 43

T4.C: Negativ-listen i tabellen GISInvalidCodes 45

T4.D: Top ti over geografiske navne efter indførsel af negativ-liste 46

T4.E: Indhold af cache-tabeller for materialet “Norge” 47

T4.F: Indhold af BiblRecordCacheSpatial for materialet “Norge” 47

T4.G: Resultat af full-text-parseren for “fakta Q” sætningen, med dansk sprog 48

T4.H: Indhold af BiblRecordCacheSpatial for materialet “Norge” med angivelse af neutralt sprog 49

T4.I: Uddrag af BiblRecordCacheSpatial for materialet “Norge” med hele OSM-datasættet 51

T4.J: Klassifikation af OSM code-fclass-type kombinationer. 53

T4.K: Resultat af spatio/tekstuel integration med positiv-liste 53

T4.L: Top 5 match i undergruppen “artwork” 54

T4.M: Resultat med simple begrænsninger 57

T4.N: Resultat af avancerede begrænsninger 57

T4.O: Indhold af BiblRecordCacheSpatial for materialet “Norge” 60

T4.P: Top ti over geografiske navne 60

T4.Q: Opsummering af resultaterne 63

T4.R: Spatial match på tv-udsendelsen “1 døgn, 2 hold, 3 dyr” 72

T4.S: Opsummering af nøgletal fra udvidelser 73

T4.T: Test af søgning med og uden spatialt indeks. 77

T4.U: Uddrag af resultat fra Kodeblok K4.O med buffer på 10 km og 100 km 78

T8.A: danMARC2 85

T8.B: Uddrag fra data importeret med GDAL ogr2ogr 85

T8.C: Uddrag fra data importeret med Shape2SqlServer 85

T8.D: Uddrag af BiblRecordCacheSpatial cache-tabellen 89

T8.E: Top ti over geografiske navne som har flest match i vores undervisningsmaterialer 90

T8.F: Indhold af cache-tabeller for materialet “Nordkraft” 119

T8.G: Manuelt og automatisk fundne geografiske navn for “Nordkraft” 119

T8.H: Udfaldet af de fundne match 120

T8.I: Indhold af cache-tabeller for materialet “Hans og Kirsten i 2. verdenskrig” 121

T8.J: Manuel og automatisk fundne geografiske navne i “Hans og Kirsten i 2. verdenskrig” 122

T8.K: Udfaldet af de fundne match 122

|  |  |
| --- | --- |
| T8.L: Indhold af cache-tabeller for materialet “Kulturkanon” | 124 |
| T8.M: Manuel og automatisk fundne geografiske navne i “Kulturkanon” | 124 |
| T8.N: Udfaldet af de fundne match | 125 |
| T8.O: Indhold af cache-tabeller for materialet “Den arabiske verden” | 127 |
| T8.P: Manuel og automatiske fundne geografiske navne i “Den arabiske verden” | 127 |
| T8.Q: Udfaldet af de fundne match | 128 |
| T8.R: Fuld klassifikation af positivlisten | 136 |
| T8.S: Liste over filer i medleveret ZIP-fil | 139 |
| **Kodeblokfortegnelse** |  |
| K3.A: Et eksempel på danMARC2 som læsbart linjeformat | 16 |
| K3.B: CRS WKT beskrivelse af det spatiale referencesystem for dataene i shape-filen | 22 |
| K3.C: Vis indholdet af et full-text indeks | 31 |
| K3.D: T-SQL Inflectional eksempel | 32 |
| K3.E: Eksempel på oprettelse af selektive XML indeks | 38 |
| K4.A: Kommandolinje til import af geospatiale data med ogr2ogr | 39 |
| K4.B: Oprettelse af et spatialt indeks på OSM tabellen “gis\_osm\_pois\_free\_1” | 42 |
| K4.C: Tilrettet SQL i CTE\_Tabellen fra Kodeblok K8.C | 45 |
| K4.D: Oprettelse af et indexeret view samt full-text-index | 50 |
| K4.E: Eksempel på håndtering af afstandskrav | 56 |
| K4.F: Eksempel på kombineret full-text-søgning og LIKE-søgning | 58 |
| K4.G: Udvidelse af SP. LimitSearch | 65 |
| K4.H: Udvidelse af SP. InsertIntoBiblRecordCachceSpatial | 66 |
| K4.I: Udsnit af en XML undertekst fra vores streaming-leverandør | 67 |
| K4.J: Oprettelse af XML Schema til validering af undertekster | 68 |
| K4.K: Oprettelse af tabel til undertekster samt full-text-indeks | 69 |
| K4.L: Eksempel på XML og full-text-søgning | 71 |
| K4.M: Oprettelse af Selective XML Index til de tekstuelle data i XML underteksterne | 71 |
| K4.N: Genopbygning af spatialt indeks | 75 |
| K4.O: Test af spatialt filter | 76 |
| K4.P: Måling af effektivitet på et spatialt indeks | 77 |
| K8.A: Kontrol af de importerede data | 86 |
| K8.B: Eksempel på brug af funktionen ST\_IsValid fra PostGIS | 86 |
| K8.C: Spatio/tekstuel integration med én enkelt OSM tabel | 88 |
| K8.E: T-SQL til oprettelse af indexed views | 92 |
| K8.F: Stored Procedures - LimitSearch | 106 |
| K8.G: Stored Procedures - InsertSubtitle | 107 |
| K8.H: Stored Procedures - InsertSubtitle | 118 |
| K8.I: Kildekode til import af undertekster | 139 |

**Figurfortegnelse**

F1.A: VIA CFU’s lager i Viborg 11

F3.A: E/R Diagram - Udsnit af mitCFU’s database 17

F3.B: Relationer mellem indeks, tabel og data på disk 27

F3.C: Et simpelt eksempel på et B+ træ (branching factor / fanout = 3) 29

F3.D: Eksempel på Hilbert kurve med punkter og en linjestring 35

F4.A: Skærmbillede af ShapeFile2SqlServer 40

F4.B: QGIS visualisering af Shape-filen gis\_osm\_pois\_free\_1.shp 41

F4.C: Den undervisningsrelevante delmængde 44

F4.D: Execution plan fra kodeblok K4.F med @name = “Giver” 59

F4.E: Oversigt over integrationsprocessen. Se også bilag 8.5 og 8.7 61

F4.F: Billede af grafisk brugerflade til geografisk filter 74

## Problemformulering

Hvordan er det muligt at benytte full-text-søgningen i Microsoft SQL Server 2017 til at finde geografiske referencer, fra frie og åbne kortdata, i store mængder tekstuelle data, for automatisk at kunne tilføje relevante geospatiale metadata?

## Arbejdsspørgsmål

Kan der laves nogle justerbare regler for præprocessering af kortmaterialerne, således at mængden af spatio-tekstuelle navne som der skal udføres full-text-søgninger på kan begrænses i forhold til en undervisningsrelevant kontekst?

Kan database indeksering effektivisere den spatio-tekstuelle integrationsproces?

Er det muligt at foretage løbende udvidelser af den spatio-tekstuelle integration med yderligere tekstuelle og spatiale data?

Kan de automatisk fundne og tilføjede spatiale metadata bruges til at give CFU’ernes brugere en ny søgemulighed, til at finde lokationsrelevante undervisningsmidler, ved at angive et punkt og en radius på et kort