Om brugen af Big Data til at identificere middelalderfragmenter

Holck, Jakob Povl

Publication date:
2015

Document version
Forlagets udgivne version

Citation for pulished version (APA):
Holck, J. P. (2015, dec 15). Om brugen af Big Data til at identificere middelalderfragmenter.

Om brugen af Big Data til at identificere middelalderfragmenter

Af Jakob Povl Holck

Overalt på de europæiske national- og universitetsbiblioteker findes bevarede fragmenter fra middelalderen. I nogle samlinger findes mange fragmenter – i andre er der få. Fragmenterne kan eksempelvis være brugt som omslag/bogindbindingsmateriale, og af den grund kan fragmenternes tekster være mere eller mindre ødelagte. Pergamentstykkerne vil ofte være klippet til på forskellig vis, for at kunne bruges af bogbinderen. I nogle tilfælde er der tale om decideret makulatur, ja nærmest konfetti. Perioden lige efter reformationen var særligt voldsom ved mange middelalderlige værker, der simpelthen blev skåret i stykker for siden at blive genbrugt til andre – med datidens øjne mere lødige – formål.

Forkortelser og sammenskrivninger

Samtidig kan tidens tand have været forholdsvis hård ved både pergament og blæk, alt efter opbevaringsforholdene gennem flere århundreder. Den forsker, der interesserer sig for at læse eller afkode teksten på et fragment, kan desuden risikere, at denne er skrevet med udbredt brug af middelalderens abbreviaturer (særlige forkortelser, hvor eksempelvis ord kan trækkes sammen, fx *IHS* = *Jesus*) og ligaturer (sammenskrivninger af enkelttegn, fx *&*, oprindelig fra latin: *et*). Selvom disse tegn og specielle konstruktioner ofte vil kunne slås op i standardværker, vil de i kombination med de andre, ovennævnte udfordringer kunne gøre en tekst nærmest ulæselig for forskeren, der ikke nødvendigvis sidder med den middelalderlige munks forkundskaber og særlige referenceramme.

Gavn af talrige editioner

Hvad kan man så gøre, når man som forsker er interesseret i at afkode latinen på et vanskeligt fragment? Ja, der er i hvert fald én faktor, der kommer forskeren til hjælp: I middelalderens verden var afskrift efter afskrift, kopier af kopier, den vanlige modus i munkenes skrivestuer. Det betyder i praksis, at meget af det, der blev produceret i et scriptorium, slet ikke var unikt, men fandtes i talrige editioner. Munkene blev uddannet til at gengive skrifterne ordret, så værkerne ikke blev forvanskede. Dette var ikke mindst alfa og omega i forhold til den korrekte gengivelse af de liturgiske skrifter i den romersk-katolske kirke. Naturligvis begik afskriverne alligevel fejl eller ændrede på ting, men i princippet blev værkernes særlige frekvens og distribution af ord typisk bibeholdt.

'Digital Humanities'

Mange af middelalderens værker er i dag transskriberet og uploadet på internettet. Som forsker kan man være heldig, at der er tale om Open Access. Transskriptionen og placeringen på nettet betyder, at et værks særlige frekvens og distribution af ord bliver søgbar i en søgemaskine som fx Google. Hvor en forsker tidligere havde brug for en ret stor og indgående viden om passager fra mange forskellige værker, om deres editioner og om forfatterne for at blive i stand til at identificere sparsom tekst fra et fragment, kan søgemaskinen i dag ofte nøjes med få udvalgte ord, eventuelt en sætning – helst flere i kombination. I løbet af få sekunder er nettet afsøgt efter tilsvarende ord med en lignende indbyrdes placering, og forskeren vil i mange tilfælde kunne identificere teksten fra middelalderfragmentet, fordi den allerede er kendt på nettet.

Big Data er dog ikke nogen mirakelløsning på den måde, at der stadig vil være et arbejde med at identificere fragmentets proveniens – hvor og hvem kommer det fra – ligesom forskeren også er nødt til at

ty til konventionelle metoder til at datere fragmentet. Der kræves et specialkendskab, hvis et fragment eventuelt skal genkendes som del af et større håndskrift. Flere andre forhold vil kræve specialistkompetencer. Men der er næppe nogen tvivl om, at udviklingen går i retning af en sammenkobling af tekst(OCR – Optical Character Recognition)- og billedgenkendelsesteknologier sammen med AI (Artificial Intelligence), hvor meget arbejde vil kunne overtages af maskinerne som forskerens forlængede arm.

Man kan betragte denne udvikling som et led i dannelsen af den digitale humaniora (DH = Digital Humanities, se fx: https://en.wikipedia.org/wiki/Digital_humanities) i erkendelsen af det potentiale, som teknologien rummer som støtte for – og videreudvikling af – de klassiske, humanistiske discipliner.

Eksempler på vellykket identifikation
På Syddansk Universitetsbibliotek er Big Data i 2015 allerede anvendt til at identificere flere middelalderhåndskriftfragmenter på latin, der ikke tidligere har været identificeret, hvoraf nogle har været meget svære at læse. Det gælder det nyligt identificerede fragment fra Herlufsholm 958.16, der netop både er meget slidt (blækket er visse steder nærmest væk) og samtidig har en næsten stenograferet tekst med masser af abbreviaturer og ligaturer. Uden tvivl har skriveren i middelalderen her skullet spare på pergamentet og måske samtidig haft en deadline at skulle leve op til, så det hele er nedskrevet lidt hurtigt. Alligevel kunne søgemaskinen – retfærdigvis efter flere søgninger og med brug af flere forskellige ordrækkefølger – knække koden og finde mønstret, med i virkeligheden meget lidt at arbejde med.

Teksten viste sig at være et udsnit fra Thomas Aquinas' (ca. 1225-1274) Questiones de Quodlibet VIII – del af et større værk (I-XII) med filosofiske spørgsmål om "hvad som helst" og tilhørende svar i artikelform.

Et andet fragment, der på denne vis lod sig afkode, var et udsnit fra Bedas (ca. 672-735) Homilia XIX. In dominica XII post pentecosten, der er placeret indvendigt på bagpermen til Herlufsholm 71.7. Her er skriften dog meget læsbar og stort set uden forkortelser, men desværre skæmmet af en, set med moderne øjne, uhensigtsmæssig placering af Herlufsholms ex libris ind over store dele af den i øvrigt meget velbevarede og smukt udformede tekst. Forpermen har i øvrigt et udsnit af kirkefaderen Origenes' (ca. 185-254) homilier (IV, kommentar til Matthæus). Skriften ser ved første øjekast ensartet ud på begge Herlufsholm 71.7-fragmenter, der muligvis har udgjort dele af et større værk med homilier, måske fra 900-tallet. Men der *er* tale om to forskellige hænder.

I dette blad kunne der for noget tid siden (10.08.10.2015) også skrives om et fragment fra Syddansk Universitetsbiblioteks samlinger i Esbjerg (specifikt RARA 47), der i tekst og noder hylder Den hellige Cecilie. Også dette fragment kunne forholdsvis hurtigt placeres ved brug af Big Data.

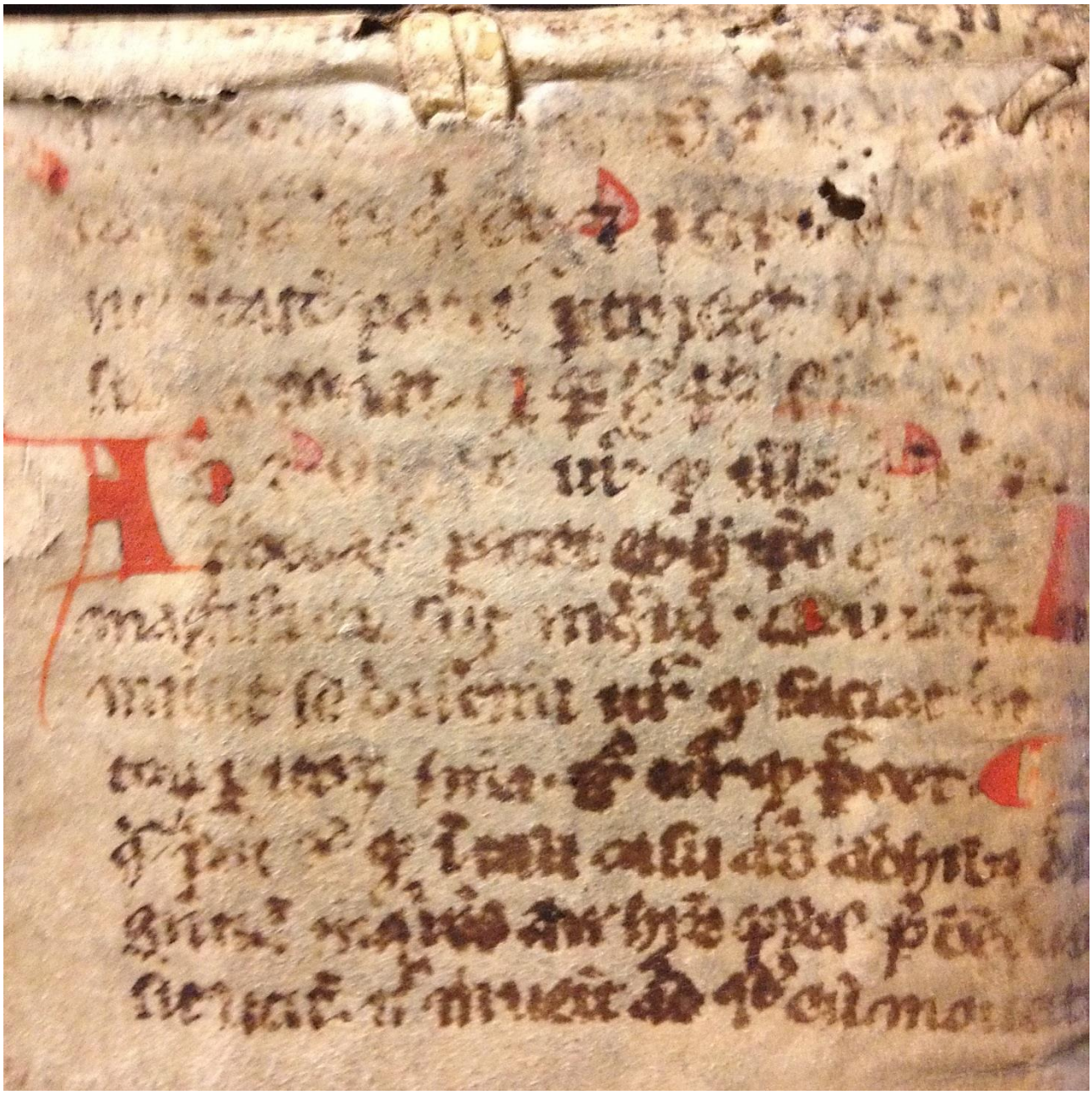
Tre vigtige betingelser
Forudsætningen for den vellykkede brug af søgemaskiner til afkodning af middelalderfragmenter er, at:

1) værket ikke er unikt

2) at en venlig sjæl har transskriberet det i én eller anden udgave, så det findes tilgængeligt på nettet i søgbar form

3) at det er muligt at udlede tilstrækkeligt med anvendelige ord fra fragmentet til at fodre søgemaskinen

Hvis disse ting er opfyldt, er der gode chancer for at få noget ud af det digitale detektivarbejde.



Fragmentet på Herlufsholm 958.16 lider under flere vanskelige karakteristika, der gør den umiddelbare læsning problematisk. Her kan Big Data komme til hjælp – ud fra 'skyen' af de ord, der trods alt kan plukkes.