Den fremtidige brugers mulighed for genfinding af data

Troels Andresen og Henrik Bulskov
Roskilde Universitet

Agenda

- IR – simpel fritekst
- Meta-data indholdsbeskrivelse
- Ressourcer til indholdsbeskrivelse
- Indeksering
- Ontologi
- Semantisk indeksering
- Summarization, …
- Meta-datas betydning …

Vores baggrund

- Tidligere samarbejde bl.a.
  - DBC
    - "Fleksibel søgning i DanBib" (DanBib og bibliotek.dk)
    - Fuzzy søgning baseret på DK5 indeksering
  - Nationalencyklopædien
    - "OntoQuery: Ontology-based Querying"
    - Ontologibaseret søgning med fokus på Ernæringsartikler i encyklopædien
  - ScanJour
    - "Videnbaseret søgning i ESDH"
    - Avanceret søgning baseret på ontologier knyttet til ESDH
  - Novo Nordisk a/s
    - "SIABO: Semantic Information Access through Biomedical Ontologies"
    - Semantisk indeksering af tekster i videnskabelige artikler
    - Baseret på centrale systematikker/taksonomier/ontologier
      - Biomedicinske UMLS, MESH, SnoMed, …
      - Sproglige WordNet, VerbNet, NomLex, …
    - Semantisk søgning og visualisering baseret på dette

Ord-baseret IR / Fritekst-søgning

- Simpel model: Inverterede filer
  - ”ord peger på dokumenter”
  - en indgang for hvert ord
  - stopord

- Mere avanceret tilgang: vægtet indeksering
  - Vector-space model
  - Fuzzy model

- Bestemmelse af vægtning
  - mål for afstand
  - hvor karakteristisk er et ord (fx hvor mange gange forekommer det i dokumentet)
  - hvor godt er et ord i indeks (meget almindeligt eller sjældent)

Meta-data

- Meta-data
  - stamdata
    - dokument-navn/titel, dato, sagsbehandler, …
  - indholdsbeskrivende data
    - manuelt eller automatisk tilknyttet …
- … det er den sidste form der er den interessante

Ressourcer til indholdsbeskrivelse (som meta-data kan referere til)

- Ressourcer som reference for meta-data
  - manuelt opbygget ”systematik”
  - automatisk genereret

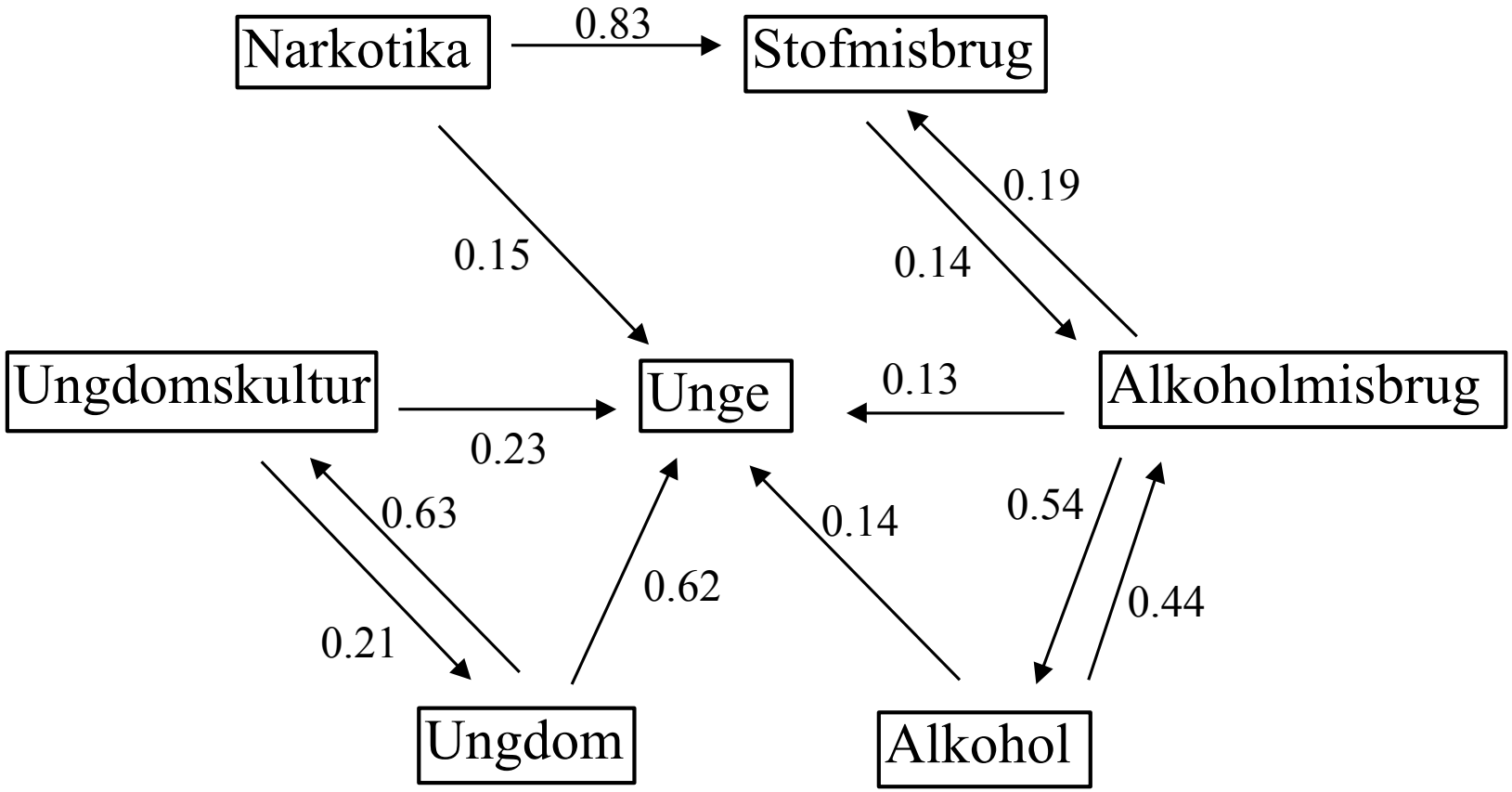
- Ressourcer, manuelt opbyggede
  - Kontrollerede Emneord
  - Klassifikationsystemer
  - Journalplan

- Ressourcer automatisk genererede
  - mængden af forekommende ord
  - vægtet udfra statistik
  - termnet: associationer imellem termer på baggrund af ko-forekomster

Indeksering

- indeksering
  - tildele indholdsbeskrivelse
  - manuelt
    - kontrollerede emneord
    - klassifikation
    - facet
  - automatisk
    - fundne forekomster af ord
- kan bruges til udlede nye ressourcer
  - termnet
  - clustering

Termnet


Narkotika
0.83
Stofmisbrug
0.15
0.19
0.14
Ungdomskultur
Unge
0.13
Alkoholmisbrug
0.23
0.63
0.14
0.54
0.62
0.44
0.21
Ungdom
Alkohol

- Termnet kan være nyttige men kvaliteten afhænger af kvaliteten af indekseringen
- DERFOR: manuelt tilknyttede metadata har potentielt stor værdi
- Termnet kan dannes på baggrund af hele database eller et udsnit fx en tidsperiode

Aktuel DBC artikelsamling 1999

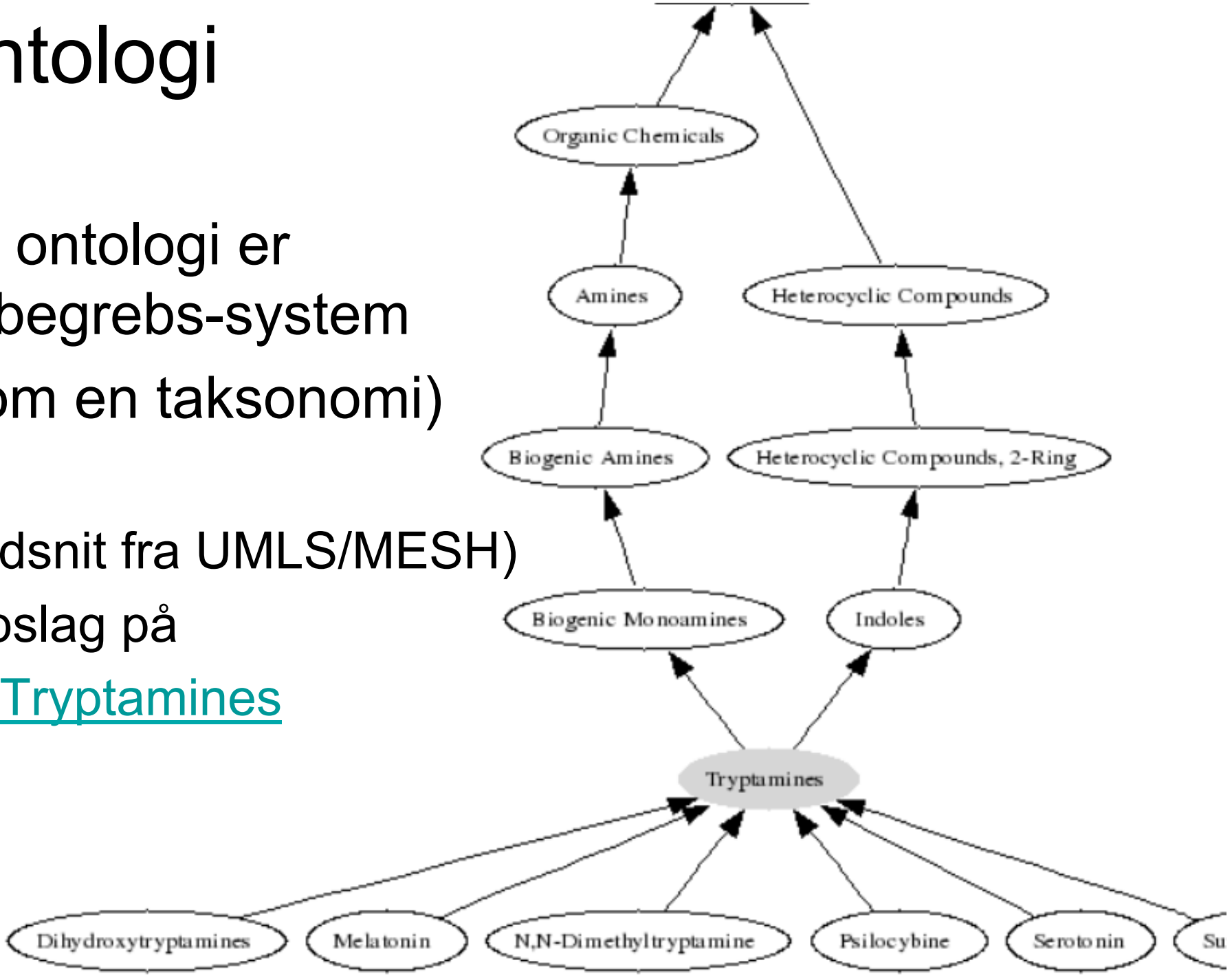
Ontologi

- Generelt
  - specificerer og relaterer “begreber”
  - Gruber (1993):
    - *An ontology is an explicit specification of a conceptualization.*
  - ingen eksakt definition
    - spænder over
      - simple afgrænsede term-lister
        …
      - is-a hierarkier
        …
      - formelle logiske systemer

Ontologi

En ontologi er
et begrebs-system
(som en taksonomi)

(Udsnit fra UMLS/MESH)
Opslag på
Tryptamines



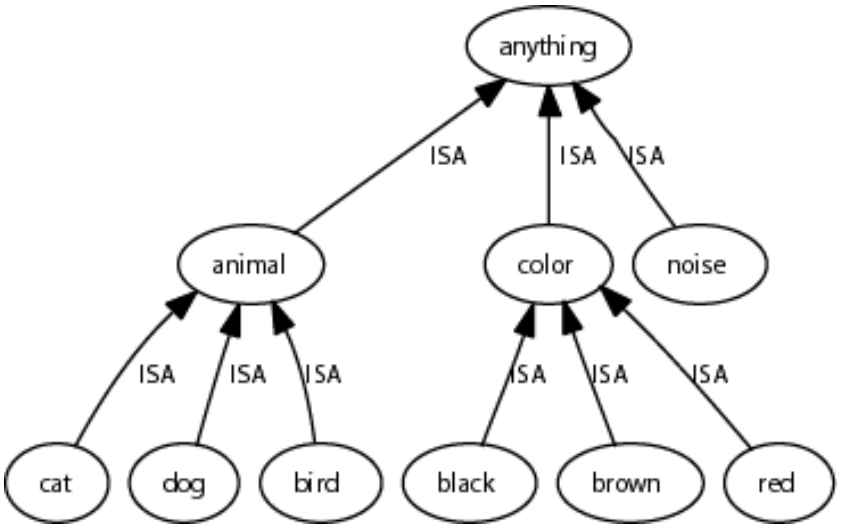
Sammensatte begreber

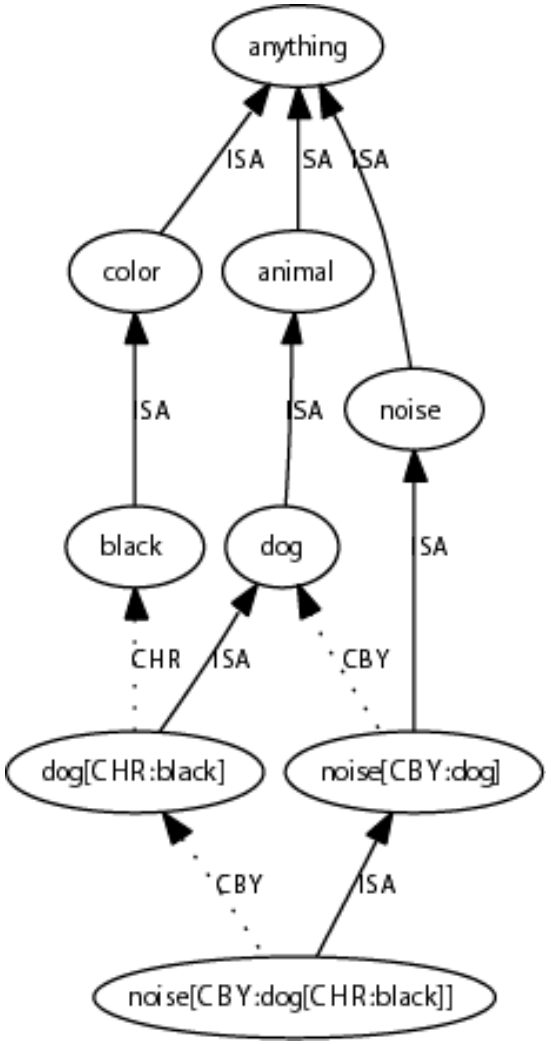
- som udgangspunkt er begreber relaterede med ‘en relation:
  - **isa:** begrebsinklusion
- med en samling af supplerende relationer fx
  - **loc**: location, position
  - **wrt**: with respect to
  - **chr**: characteristic (property ascription)
  - **cby**: caused by
  - **agt**: agent of act or process
  - **pnt**: patient of act or process
  - …
- dannes sammensatte begreber, fx
  - woman
  - woman[**chr** : pregnant]
  - treatment[**pnt**: woman[**chr**: diabetic, **chr**: pregnant]]

- begreber kan altså sammensættes og dermed danne nye begreber,

Ontologi

- Baggrunds-ontologi
- Udvidet med genkendte begreber





Ontologisk semantik

- at “afbilde” ord og fraser ind i begreber i ontologien
- i princippet uafhængigt af forskellige sproglige iklædninger, således at samme begreb, fx
  - barn[**chr**: overvægtig]
- kan genkendes udfra:
  - ”overvægtige børn”
  - ”børn med overvægt”
  - ”børn som er overvægtige”

Ontologisk beskrivelse / Semantisk indeksering

- beskrivelsen (indekseringen) for en tekst er en struktur af (udvalgte) begreber der indgår i denne

- for teksten
  - ”behandling af overvægtige børn med alternativ medicin”
- er følgende mulige alternative beskrivelser
  - {behandling, overvægtig, barn, alternativ, medicin}
  - {behandling, {overvægtig, barn}, {alternativ, medicin}}
  - {behandling, barn[**chr**: overvægtig], medicin[**chr**: alternativ] }
  - {behandling[**ptn**: barn[**chr**: overvægtig]], medicin[**chr**: alternativ] }

ptn: patient of act or process
chr: characteristic

- semantisk indeksering
  - erstatter / udvider konventionel ord-baseret indeksering
  - semantisk indeks =
    ontologisk beskrivelse =
    mængde af identificerede begreber

Semantisk indeksering
Eksempel

- begrebs-baseret indeks =
  ontologisk beskrivelse =
  mængde af identificerede begreber

- fx kan
  - ” 10,594 pregnant women registered at 47 primary health care (PHC) centers in Al-Hassa”
- beskrives med
  - { care[chr: primary, wrt: health], center[wrt: Al-Hassa], woman[chr: pregnant] }

Ekstraktion af begreber til semantisk indeksering

- The newest drugs, selective serotonin reuptake inhibitors (SSRIs), relieve depression similar to the TCAs and MAOIs, but have a lower rate of unpleasant side effects.

- … selective serotonin reuptake inhibitors …

- selective/JJ serotonin/NN reuptake/NN inhibitor/NNS

- NP[selective/JJ serotonin/NN reuptake/NN inhibitors/NNS]

- inhibitor[chr: selective, wrt: reuptake, wrt: serotonin]

Ekstraktion af begreber til semantisk indeksering

- The newest drugs, selective serotonin reuptake inhibitors (SSRIs), relieve depression similar to the TCAs and MAOIs, but have a lower rate of unpleasant side effects.

- The/DT new/JJS drug/NNS ,/, selective/JJ serotonin/NN reuptake/NN inhibitor/NNS (/(SSRI/NNS)/) ,/, relieve/VB depression/NN similar/JJ to/TO the/DT TCA/NNS and/CC MAOI/NNS ,/, but/CC have/VBP a/DT low/JJR rate/NN of/IN unpleasant/JJ side/NN effect/NNS ./.

- **NP[**The/DT new/JJS drug/NNS**]** ,/, **NP[**selective/JJ serotonin/NN reuptake/NN inhibitor/NNS**]** (/(**NP[**SSRI/NNS**]**)/) ,/, relieve/VB **NP[**depression/NN**]** similar/JJ to/TO **NP[**the/DT TCA/NNS**]** and/CC **NP[**MAOI/NNS**]** ,/, but/CC have/VBP **NP[**a/DT low/JJR rate/NN**]** **PP[**of/IN unpleasant/JJ side/NN effect/NNS**]** ./.

Ekstraktion af begreber til semantisk indeksering

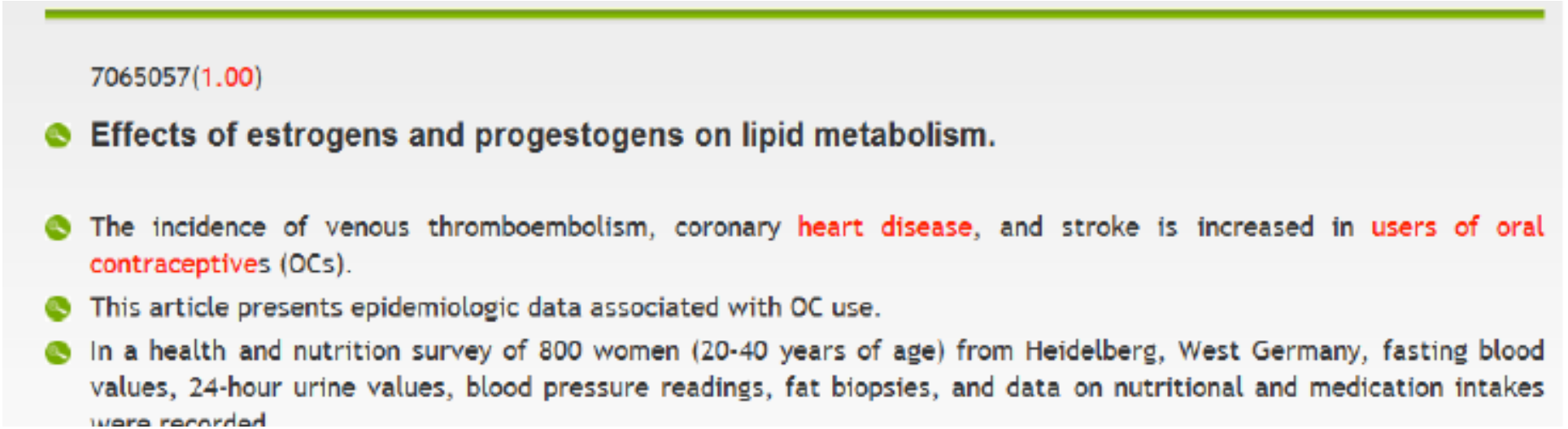
- The newest drugs, selective serotonin reuptake inhibitors (SSRIs), relieve depression similar to the TCAs and MAOIs, but have a lower rate of unpleasant side effects.

NP[The new drug]	drug[chr: new]
NP[selective serotonin reuptake inhibitor]	inhibitor[chr: selective wrt: reuptake, wrt: serotonin]
NP[SSRI]	SSRI
NP[depression]	depression
NP[TCA]	TCA
NP[MAOI]	MAOI
NP[a low rate] PP[of][NP[unpleasant side effect]]	rate[chr: low, wrt: effect[chr: unpleasant, wrt: side]]

Semantisk søgning

- … en prototype
- Søgespecifikation
  - Et tekstuelt udtryk (sætning, frase, opremsning, …)
- Søgeevaluering
  - bedste match i en ”fuzzy” evaluering
  - ontologi inddrages ved
    - similaritetsmål over begreber i ontologien samt
    - ekspansion af forespørgsel med ”similare” begreber
- Eksempel, søgning på:

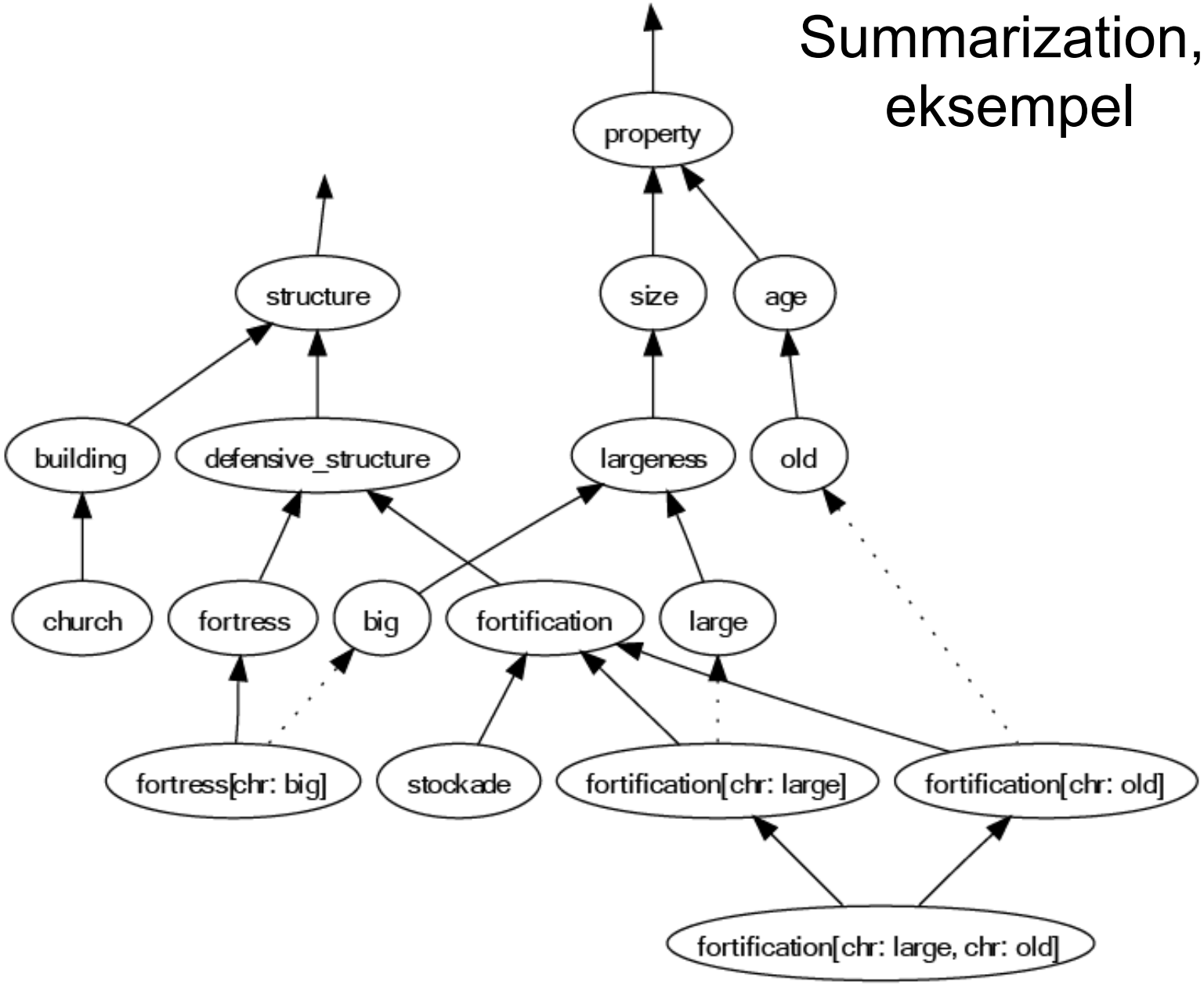
cardiovascular disease in users of birth control pills



andre mulige anvendelser …

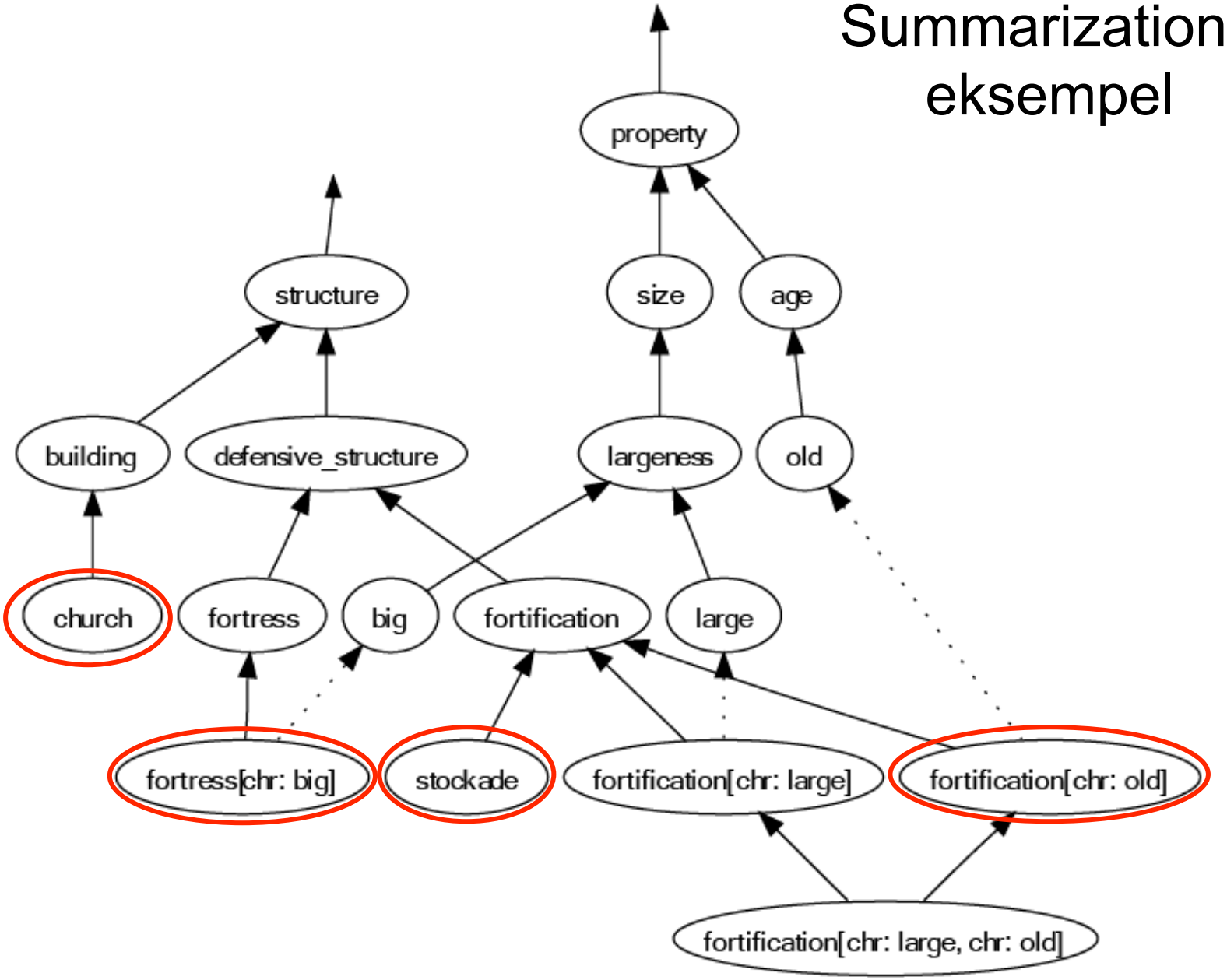
- Udover søgning kan en ontologi anvendes til bl.a.
  - zoom og summarization (opsummering af svar)
  - clustering (samling af dokumenter i grupper)
  - automatisk klassifikation
  - grafisk visualisering (præsentation af dokumenter, såsom svar, fx igennem graf over systematikken)
  - …

Summarization, eksempel



C = {church, fortification[CHR: *old*]*, stockade, fortress*[CHR: *big*]*}*

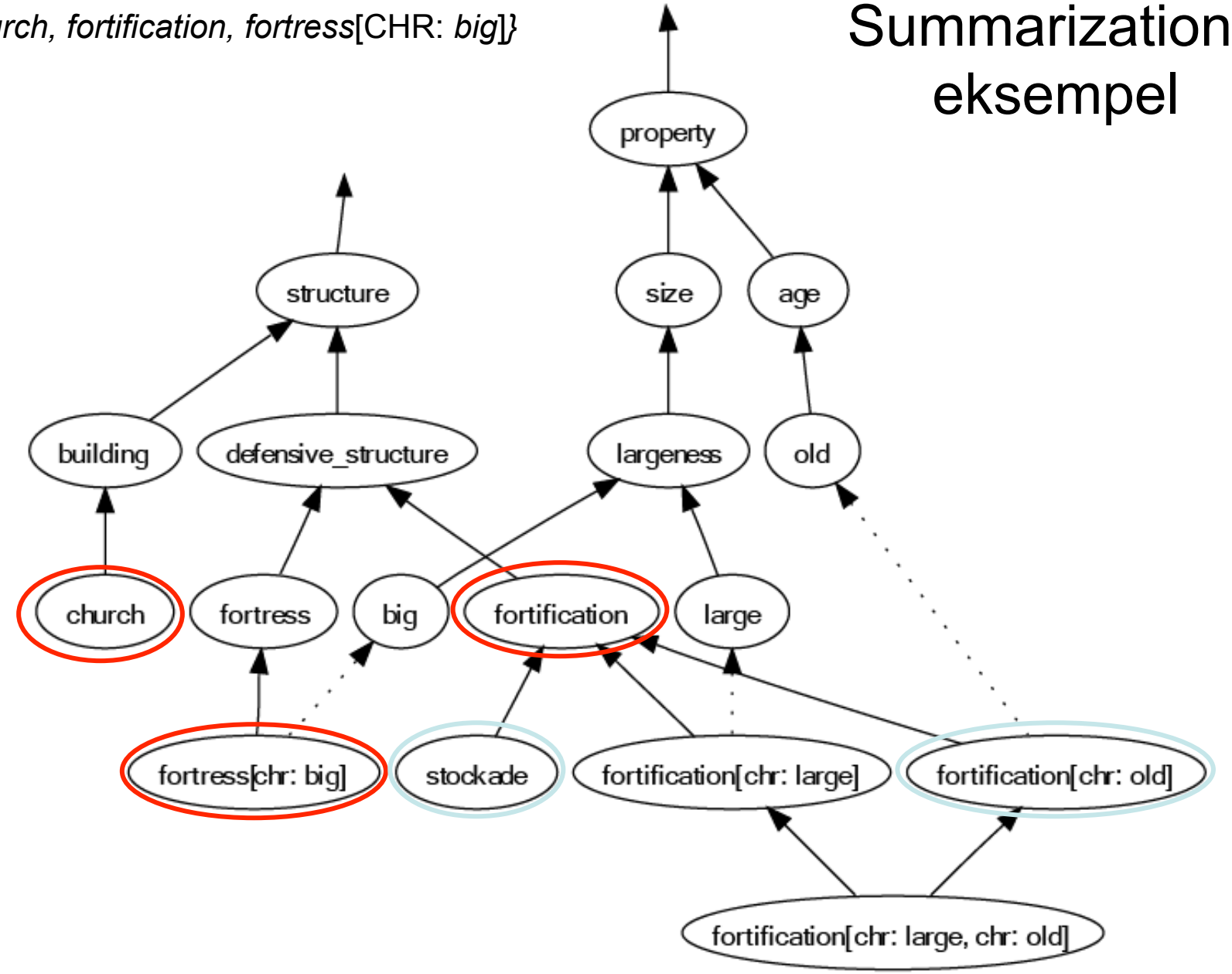
Summarization, eksempel



Summarization, eksempel

C = *{church, fortification*[CHR: *old*], *stockade, fortress*[CHR: *big*]*}*

δ(*C*) = *{church, fortification, fortress*[CHR: *big*]*}*

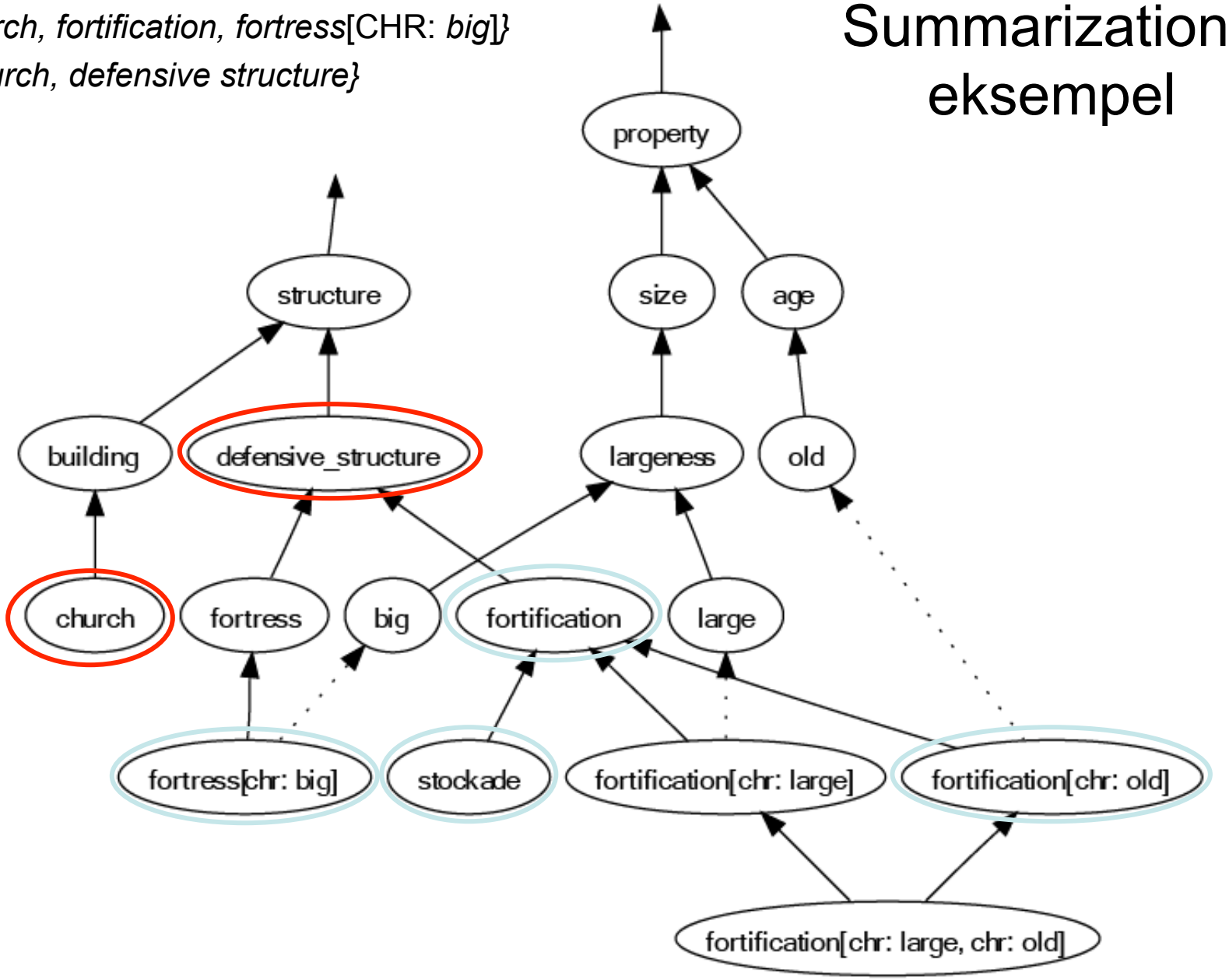


Summarization, eksempel

C = *{church, fortification*[CHR: *old*], *stockade, fortress*[CHR: *big*]*}*

$\delta(C)$ = *{church, fortification, fortress*[CHR: *big*]*}*

$\delta^2(C)$ = *{church, defensive structure}*



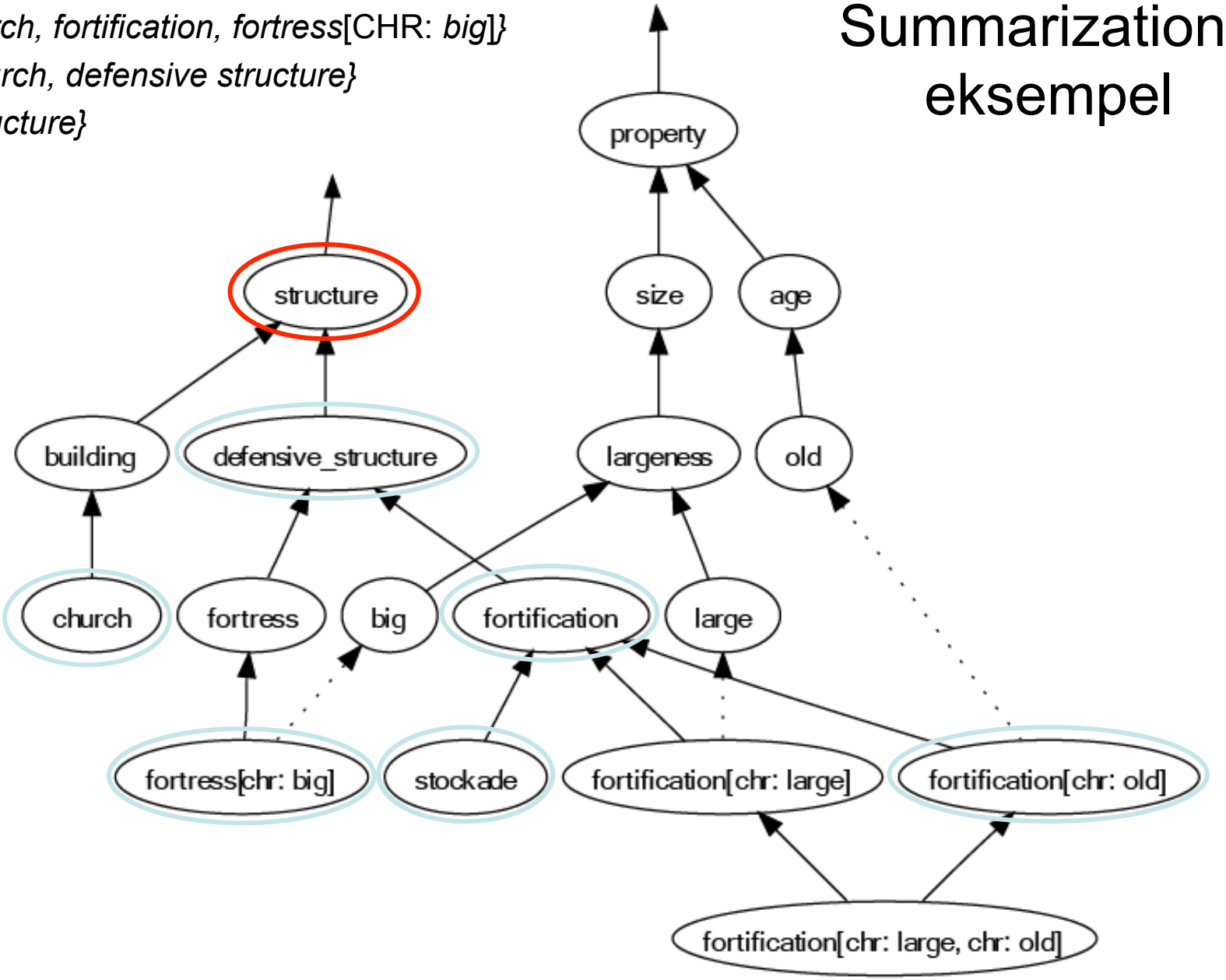
Summarization, eksempel

C = *{church, fortification*[CHR: *old*], *stockade, fortress*[CHR: *big*]*}*

$\delta(C)$ = *{church, fortification, fortress*[CHR: *big*]*}*

$\delta^2(C)$ = *{church, defensive structure}*

$\delta^3(C)$ = *{structure}*



Brug af ressourcer

- metadata (indholdsbeskrivelse), som
  - simple ordindeks
  - kontrollerede emneord
  - begrebs-indeks (ontologi/taksonomi-baseret)
  - klassifikation
  - facet-specifikationer
- suppleret med de ressourcer meta-data refererer til, som
  - kontrollerede emneords-systematik
  - ontologi/taksonomi
  - klassifikationsystemer
  - journalplan
- kan danne grundlag for
  - bedre og mere nuanceret søgning
- samt for alternative ”forespørgselssvar”
  - zoom og summarization (opsummering af svar)
  - clustering (samling af dokumenter i grupper)
  - automatisk klassifikation
  - grafisk visualisering (præsentation af dokumenter, såsom svar, fx igennem graf over systematikken)
  - …

Brug af ressourcer

- udover det
  - simple ordindeks, der dannes for at understøtte fritekstsøgning
- kan man nå langt med automatiske metoder til at danne
  - semantisk indeksering
  - semantisk analyse af tekstindhold
  - “kunstig systematik” såsom termnet, termhierarkier, afledte ontologier
  - meta-data/indholdsbeskrivelse med reference til “kunstige systematikker”
  - meta-data/indholdsbeskrivelse med reference til “rigtige systematikker”, som kontrollerede emneord, ontologier/ taksonomier, mm
- men …

Meta-data?

  - der er skabt manuelt
  - og skabt udfra fra et kendskab til indhold i forbindelse med registrering/ oprettelse
  - udgør typisk meget værdifuld “viden”
  - har ofte en meget stor betydning for mulighederne for genfinding og
  - kan normalt ikke erstattes/genskabes ved automatiske metoder

- og vil dette ændre sig?
  - (efterhånden som automatiske metoder til indeksering, indholdsbeskrivelse og ressource-opbygning bliver mere raffinerede i fremtiden)
  - næppe
- men derimod åbnes der nye spændende muligheder med den helt naturlige **synergi imellem**
  - den pålidelige “viden” som **manuelt tilføjede meta-data** udgør og
  - disse **automatiske metoder**