Sisällönanalyysi ja sisällönkuvailu bibliografisen valvonnan välineinä

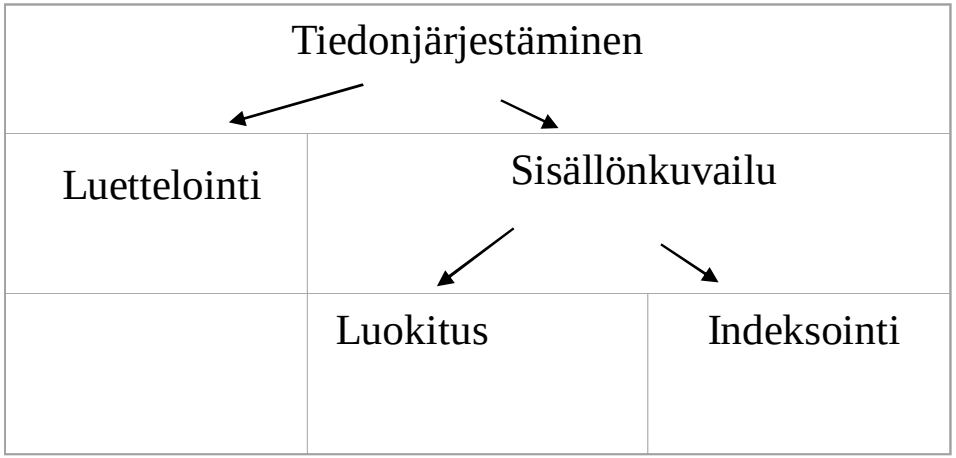
Eloseminaari sisällönkuvailusta, Oulun yliopiston kirjasto

25.8.2010 klo 10.30-11.15

Jarmo Saarti

jarmo.saarti@uef.fi

Tiedonjärjestämisessä käytettävät välineet



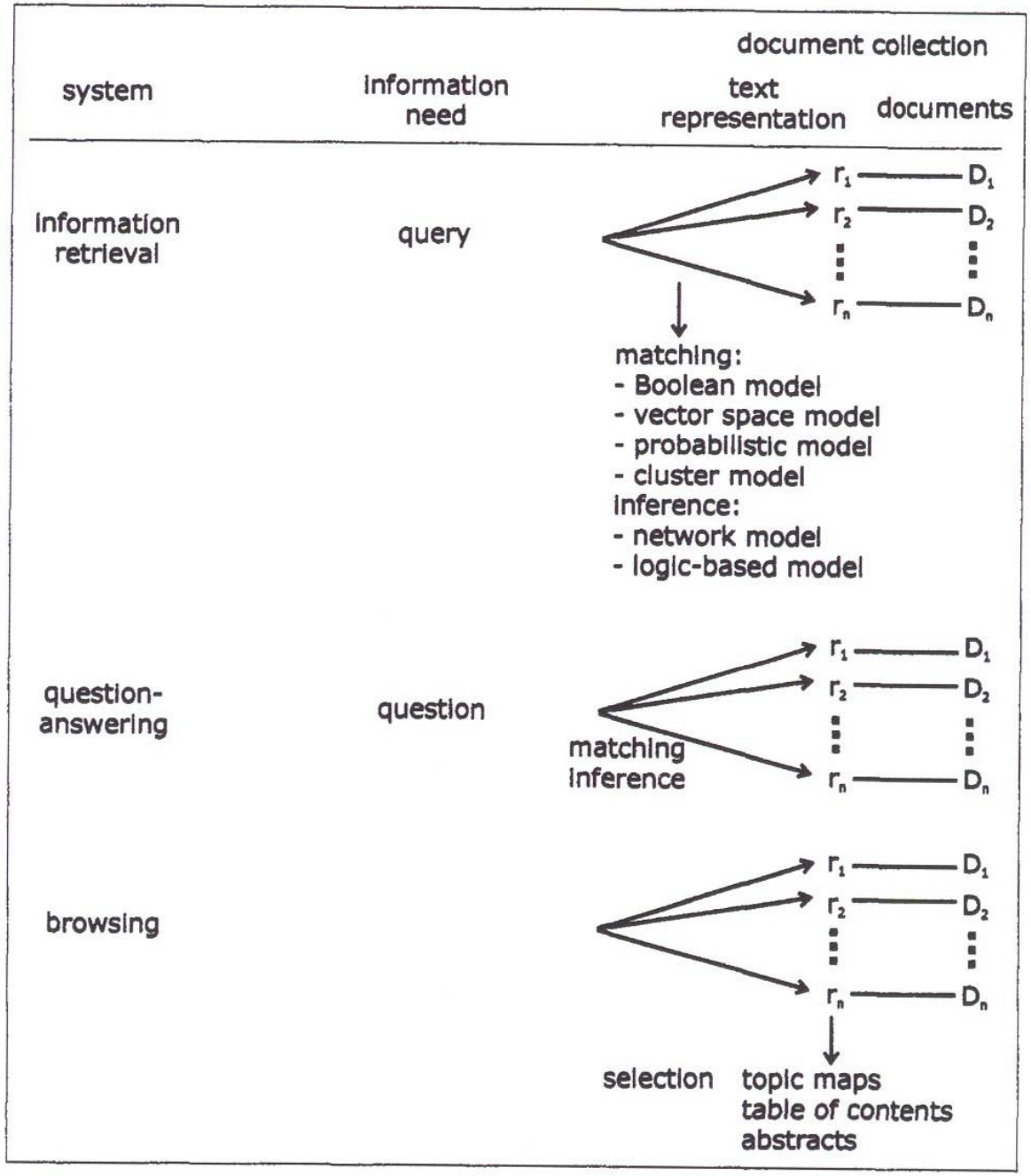


Figure 2. Actualization of an information need.

Tiedon tallennus- ja hakujärjestelmän luominen

- käyttäjäanalyysi - mille käyttäjäkunnalle se luodaan
- ympäristöanalyysi - millaiseen käyttöympäristöön se luodaan
- funktioanalyysi - mikä on järjestelmän käyttötarkoitus
- aineistoanalyysi - millaista aineistoa järjestelmällä kuvaillaan
- projektisuunnitelma - kuka tekee, missä, millä välineillä ja kuka maksaa kulut ja vastaa ylläpidosta
- kustannus/hyötyanalyysi - kannattaako järjestelmään rakentaa ja millä ehdoilla

Luettelointi, sisällönkuvailu, luokitus, indeksointi, tiivistelmät

- luettelointi = dokumentin eri aspektien identifioimista. Luettelon tehtävänä on vastata käyttäjän kysymyksiin koskien tiettyä aihetta, tietyn tekijän teoksia tai tiettyä dokumenttia. Lisäksi luettelosta on käytävä ilmi, mistä etsitty dokumentti on saatavilla.
- sisällönkuvailu = dokumentin sisällön tiivistetty kuvailu tiedonhakua ja tiedonvälitystä varten. Se perustuu sisällön analyysiin = sisällön erittely ja jäsentely.
- luokitus = toisiinsa liittyvien sisältöjen systemaattista järjestämistä (luokituskielen avulla, jonka rakenne esitetään luokituskaavana)
- indeksointi = sisällönkuvailu asiasanojen tai avainsanojen avulla
- tiivistelmä (abstrakti, referaatti, lyhennelmä) = suppea, itsenäinen esitys dokumentin sisällöstä.

Informaatiorakenteet informaatiotutkimuksessa

- BIBLIOMETRIA: dokumenttien keskinäisten suhteiden muodostamat rakenteet, dokumenttien tuotanto ja käyttö
- DOKUMENTAATIOKIELET (dokumenttien rakenteita kuvaavat rakenteet + käyttäjien tavat lähestyä, tiedonhankintakäyttäytyminen, kognitiiviset tottumukset)
- IR-JÄRJESTELMÄT (dokumenttien rakenteita kuvaavien ja käyttäjien tarpeita kuvaavien rakenteiden rakenteina)
- BIBLIOGRAFIA (dokumenttien kuvaus bibliografisella tasolla)
- AUTOMAATTINEN INDEKSOINTI (dokumenttien rakenteet perusteina valittaessa dokumenttien sisältöä parhaiten kuvaavia sanoja)
- DOKUMENTTIEN TUOTTAMINEN (mm. hyperteksti, vrt. dokumentaatiotutkimus ja dokumentaatiomuodot / Lund)

Missä suhteessa dokumenttien sisältöjä luokitetaan (eritellään)

- Opinalojen (disipliinien) suhteen
- Mihin oppiaineeseen / tieteenalaan kirja kuuluu?, kirja kuuluu filosofiaan / sosiologiaan jne.
- Muut dokumentin piirteet (fyysinen / sisällöllinen muoto, kieli jne.)
- Käsiteltyjen ilmiöiden ja ’tapahtumien’
- Mistä ilmiöstä kirja kertoo?, kirja kertoo lapsista ja heidän kehityksestään
- Aiheet, jotka eivät ’alistu’ tapahtuman tai edes ilmiön muotoon (korkeasti abstraktit käsitteet dokumentin aiheina, esim. filosofiassa tai sosiologiassakin)
- Request warrant: minkä oletetaan kiinnostavan käyttäjiä, ilmentävän heidän näkökulmaansa?

Vastauksia ongelmiin

- Dokumenttien automaattinen käsittely
  - automaattinen luokitus ja indeksointi
  - bibliometria, viiteindeksointi
  - hakualgoritmit
- Standardisointi
  - luettelointisäännöstöt
  - kontrolloidut sanastot
- Ennustettavuus
  - Haun ja dokumenttien yhteen sovittelun mahdollisuudet
  - suhteita ilmaisevat sisällönkuvailun järjestelmät (luokituskaavat, tesaurustyyppiset sanastot)
  - käyttäjäliittymät, tiedonhakujärjestelmät, hakuohjelmat, tietuerakenteet
- Selailtavuus

Standardien hyödyntämisketju

Kuva 6. Standardien hyödyntämisketju.

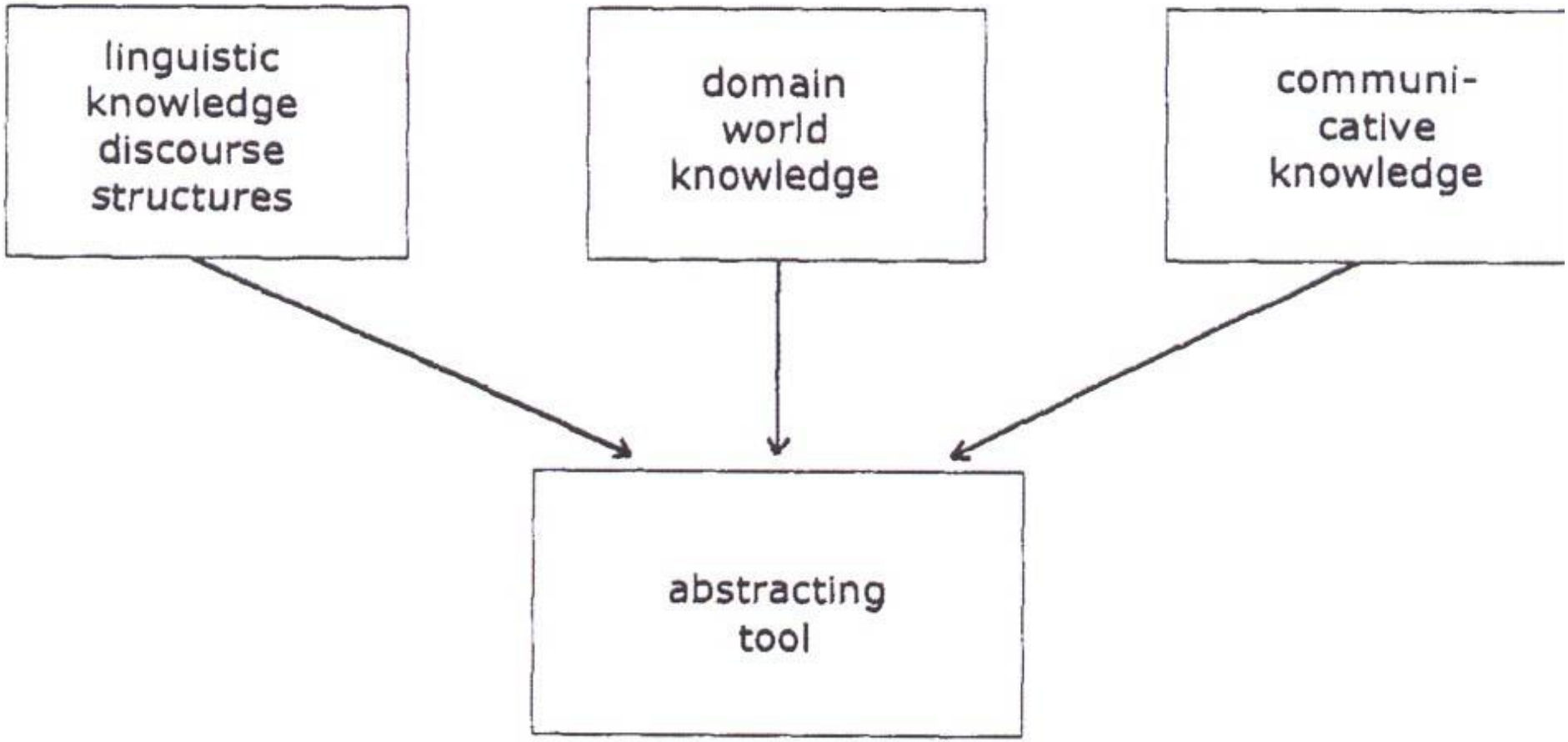


Figure 7. Important knowledge sources in automatic abstracting (cf. Sparck Jones, 1993).

Ulottuvuudet ja perusteet, joihin kuvaus ja organisointi perustuvat

- Warrant ja aboutness ja dokumenttien kuvaamisen tematiikka
- WARRANT (peruste):
- Mikä oikeuttaa erityisen luokan, asiasanan tms. olemassaolon järjestelmässä?
- ABOUTNESS (ei kunnollista suomennosta)
- ”what the book is about”,
- miten ja mitä siitä pitäisi ilmaista tiedon tallennuksessa ja haussa?
- Warrant, dokumentaatiokielen perusta
- Aboutness, dokumentaatiokielen käytön perusta
- Yhdessä huomautuksia siitä, miten dokumenttien sisältöä pitäisi kuvata!

Aboutness problematiikka

- aboutness-käsitte liittyy siihen, miten luokittelijat ja indeksoijat päättävät, mikä on dokumentin sisältö eli mistä se kertoo (what it is ‘about’).
- teema = mitä dokumentti käsittelee
- reema = mitä uutta dokumentti sanoo teemastaan

Relevanssin käsitteestä

- käsitteen määrittely hankalaa - sen ala on sumea
- tautologia: tavoitteena etsiä relevanttia informaatiota
- rinnastettu yhteenkuuluvuuteen, vastaavuuteen, aiheenmukaisuuteen, osuvuuteen, hyödyllisyyteen, käyttökelpoisuuteen
- = suhde informaation, sen käyttäjän ja käyttötilanteen välillä.

Relevanssi

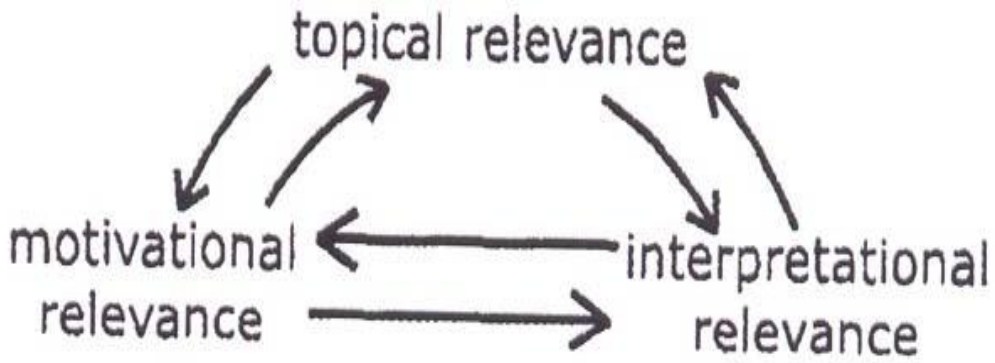


Figure 2. Relationships between topical relevance, motivational relevance, and interpretational relevance (cf. Saracevic, 1975).

Aihe vs. käyttäjärelevanssi

- aiher. = sanojen täsmäyttämistä dokumenteissa ja kyselyissä
- käyttäjär. = käyttäjän arvio dokumentin käyttökelpoisuudesta

Dokumenttien sisällönkuvailu

- sisällönkuvailu = dokumentin sisällön tiivistetty kuvailu tiedonhakua ja tiedonvälitystä varten.
- se perustuu sisällön analyysiin = sisällön erittely ja jäsentely.

Modernin luokitusteorian ydin (D. Austin, 1970)

- Mahdollisuus jakaa (kompleksit) aiheet / käsitteet osiin
- Aiheiden/käsitteiden kokoaminen uudelleen määrämuotoisiksi kokonaisuuksiksi ”mainintajärjestykseksi kutsuttua päätöksentekomallia noudattaen”
- Fasetoitu l. analyyttis-synteettinen luokitus

Sisällönkuvailujärjestelmän käyttäminen

- tutustu käyttämääsi sisällönkuvailujärjestelmään – lue esipuhe, sisäistä logiikka, jonka mukaan se on rakennettu
- katso miten tiedontallennus- ja hakujärjestelmä, jota käytät hyödyntää käyttämääsi sisällönkuvailujärjestelmää
- tutustu johonkin (hyvään) tietokantaan, jossa sisällönkuvailujärjestelmää on käytetty
- mieti, miten käännät tällä järjestelmällä teosten sisällön niiksi asioiksi, joita asiakkaasi hakee
- opeta asiakkaasi käyttämään järjestelmää tai tee hyvät opasteet

Sisällönkuvailu tekeminen

- tutustu teokseen ja mieti sen ydinsisältö:
  - nimeke, sisällysluettelo, kansitekstit, johdanto, päätäntö, hakemisto
- mihin kokonaisuuksiin teoksen ydinsisältö jakaantuu
- miten kertoisit teoksen ydinsisällön itsellesi/asiakkaalle
- miten suhteutat teokset muihin oman kokoelman teoksiin – mitä samanlaista, mitä erilaista
- miten käännät teoksen käyttämiesi sisällönkuvailujärjestelmien kielelle

Sisällönkuvailun laatuvaatimukset

1. **Exhaustivity (tyhjentävyys)** refers to the degree to which all the concepts and notions included in the text are recognized in its description, including the central topics and the ones treated only briefly.
2. **Specificity (spesifisyys)** refers to the degree of generalization of the representation. (“**Kymmenen laki”.**)
3. **Correctness (virheettömyys)** is important. Indexing and abstracting are susceptible to two kinds of errors: errors of omission and errors of commission. The former refers to a content description that should be assigned, but is omitted. The latter refers to a content description that should not be assigned, but is nevertheless attributed. Omitting a correct description and assigning a broader, narrower, or related description is a special kind of error that is at once an error of omission and commission. Correctness compares the actual text representation with the ideal one.
4. **Consistency (konsistenssi)** compares representations that are made of the same source

Tiivistelmät

- tiivistelmä (abstrakti, referaatti, lyhennelmä) on suppea ja itsenäinen kirjallinen esitys dokumentin sisällöstä.
- 1. indikatiivisia tiivistelmiä, joiden tarkoituksena on johdattaa alkuperäisen dokumentin luo
- 2. informatiivisia tiivistelmiä, joiden tarkoituksena on antaa mahdollisimman paljon tietoa dokumentin sisällöstä
- 3. evaluatiivisia eli kriittisiä tiivistelmiä, joiden tarkoituksena on arvioida ja arvostella alkuperäistä dokumenttia

Rakenteelliset tiivistelmät

- abstraktiin voidaan määritellä rakenne, jolloin tuloksena saadaan fasetoitu rakenne tiivistelmään tiedonhakua tehostamaan
- esim. Emerald-kustantajan tutkimusartikkelin tiivistelmän rakenne
  - purpose
  - desing/methodology/approach
  - findings
  - research limitations/implications
  - practical implications
  - originality/value
  - keywords
  - paper type

Automaattinen kuvailu

1. Sanastoanalyysi – yksittäisten sanojen tunnistaminen tekstistä
2. Turhien sanojen karsiminen – alueelle epäspesifit termit ja kieliopilliset termit stoplistoja käyttäen
3. Sanojen palauttaminen perusmuotoon (mahdollisesti)
4. Indeksitermien muodostaminen lauseiksi (mahdollisesti)
5. Vapaatekstitermien korvaaminen tesaurustermeillä (mahdollisesti)
6. Painojen laskeminen kullekin termille

Luokitus

- luokitus on samanlaisten asioiden yhdistämistä ja erilaisten erottamista
- dokumenttien sisältöjen jäsentämistä ja niiden suhteuttamista toisiin dokumentteihin
- luokitus on kokonaisuuksien järjestelmällistä hallintaa

Luokitustyypit laajuuden mukaan

- yleisluokitukset
- alakohtaiset luokitukset
- erikoisluokitukset

Luokitustyypit funktion mukaan

- luokitus aineiston järjestäjänä, esim. hyllyluokitus
- luokitus sisällönkuvaajana ja toisaalta tiedon tallennuksen ja haun välineenä

Luokituskaavatyypit

- hierarkkiset, enumeratiiviset luokitukset
- fasetoidut luokitukset
- hybridit eli edellisten sekamuodot

Luokituskaava

- jakautuu yleensä vähintään luokitustauluihin ja hakemistoon
- luokitustauluissa luokat ovat sisällön mukaisessa, systemaattisessa järjestyksessä
- notaatio (luokkamerkki, luokkasymboli)

Luokituskaavojen analysointi

- ideataso = luodaan taulukot eli systematiikka
- notaatiotaso = luodaan luokituksen notaatio, sen merkit ja kielioppi
- verbaalitaso = valitaan luokituksessa käytettävät luokkien nimet, joita käytetään mm. taulukoissa ja hakemistoissa

Indeksointi

- johdettu/mekaaninen indeksointi (derived i.), jossa indeksointi suoritetaan mekaanisilla tai matemaattistilastollisilla menetelmillä tai algoritmeilla
- intellektuaalinen/määrittelevä indeksointi (intellectual/assigned i.), jossa indeksoija kuvailee dokumentit asiasanoilla

Intellektuaalinen/määrittelevä indeksointi (assigned indexing)

- indeksointi tehdään määrittelemällä dokumentin kuvailutietoihin asiasanoja
- indeksoija (=ihminen) tekee tulkinnat ja valinnat – ammattilainen tai dokumentin tekijä
- apuna asiasanastot

Termityypit

- termityypit ja niiden valinta:
- entiteetit
- aktiviteetit
- abstraktit seikat
- ominaisuudet

Termien väliset suhteet

- hierarkkinen suhde (suku-laji;kokonaisuus-osa)
- määrittävä/assosiatiivinen suhde (koordinaatio, sukulaisuus, samanaikainen, perusta/seuraus,väline, materiaali, samanlaisuus)

Asiasanastot

- asiasanojen ja ohjaustermien luettelo
- yleiset a., alakohtaiset a., aakkoselliset a., hierarkkiset a.
- asiasanasto liittyy usein johonkin aihepiiriin, dokumenttikokoelmaan tai tiedonhakujärjestelmään

Tesaurukset

- asiasanasto, jossa asiasanojen yhteydessä ilmaistaan niiden suhteet muihin asiasanoihin ja ohjaustermeihin

Liha-alan sanasto

YLEISET SANAT JA TERMIT 10000-

10000
teuraseläin
teurastettavaksi tarkoitettu, yleensä lihotettu eläin, joka teurastetaan siitä saatavan lihan ja elinten vuoksi
en slaughter animal
sv slaktdjur (n)
de Schlachttier (n)

10001
elopaino
eläimen paino (kg) elävänä
en live weight, green weight
sv levande vikt
de Lebendgewicht (n)

10002
ante mortem
kuolemaa/teurastusta ennen. Ko. latinankielinen ilmaisu on yleisesti käytössä teurastamoteollisuudessa puhuttaessa teurastusta edeltävistä tapahtumista.
myös teurastusta edeltävä
en ante mortem, pre-slaughter
sv ante mortem, före slakten
de ante mortem,
vor der Schlachtung (f)

10002
teurastusta edeltävä
ks. ante mortem

10101
teurastuksen jälkeinen
ks. post mortem

10103
tajuton
ympäristöään ja omia subjektiivisia kokemuksiaan tiedostamaton. Tajuttomalla on hermojen ja aivojen yhteistoiminta syystä tai toisesta häiriintynyt. Tajuttomuutta on monenasteista, täydellistä tajuttomuutta kutsutaan koomaksi.
en unconscious, comatose, stunned
sv medvetslös, bedövad
de bewußtlos, betäubt

10110
lattiateurastus
teurastus, joka tapahtuu pääosin lattiatasossa eikä teurastettava eläin kulje kattoraiteella teurastuksen aikana
en slaughter on the floor
sv golvslakt
de Schlachtung (f), liegende, stationäre Schlachtung (f)

10120
ratateurastus
teurastus, joka tapahtuu teuraseläimen edetessä kattorataa pitkin työvaiheesta toiseen
en slaughter on rail
sv hängande slakt, banslakt
de Schlachtung (f), hängende Bandschlachtung (f)

en casualty slaughter, quarantine slaughter
sv karantänslakt
de Sanitätschlachtung (f), Karantenschlacht (f)

10160
rituaaliteurastus
mm. juutalaisten (kosher-sektaus) ja muslimien (halal) harjoittama teurastustapa, jossa eläimen kaulavaltimot katkaistaan eläintä tainnuttamatta
en ritual slaughter
sv ritualslakt
de Ritualschlachtung (f)

11000
ruho
kuollut eläin/teuraseläin, jonka kaulasuonet on katkaistu, veri valutettu sekä vuota (sialla karvat ja orvaskesi) ja sisäelimet sekä sioilla korvat, naudoilla, lampailla ja hevosilla lisäksi pää, sorkat ja häntä on poistettu
myös teurasruho
en carcass, carcase
sv slaktkropp, slaktfall (n)
de Schlachtkörper (m)

11000
teurasruho
ks. ruho

11001
teuraspaino

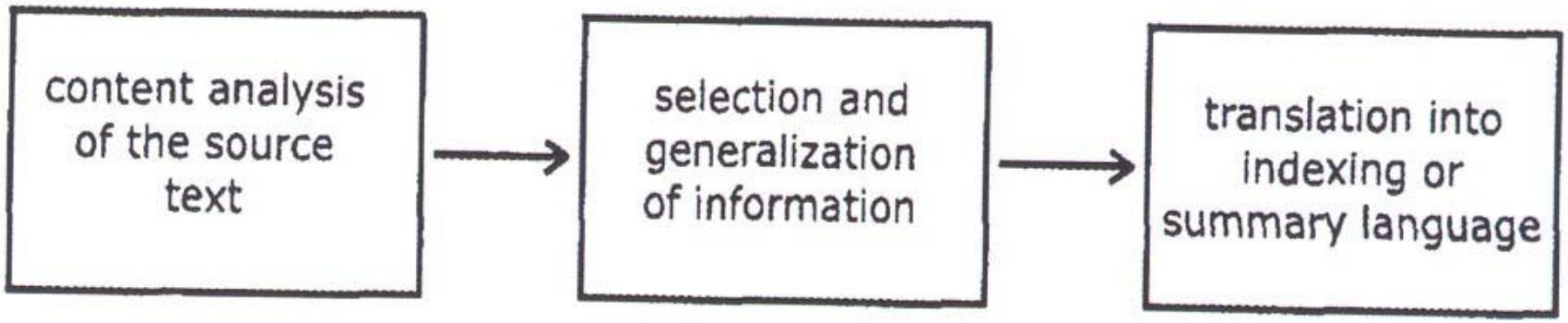


Figure 1. Intellectual indexing and abstracting.

Asiasanastojen suunnittelu ja kokoaminen

- yleissanasto vs. erikoissanasto
- alueen ja sen keskeisten aspektien määrittäminen
- auktoriteettien määrittäminen
- termien valinta
- termien suhteuttaminen toisiinsa
- sanaston kokoaminen
- sanaston koekäyttö
- sanaston julkaiseminen
- ylläpito
- ks. esim. Hoidokki http://www.shks.fi/hoidokki_hoitotyon_asiasanasto-/

Sisällönkuvailun prosessi ja tulos

- teoslähtöisyys vs. asiakaslähtöisyys
- valmiiden tietueiden noutaminen
- näiden täydentäminen
- sisällönkuvailun toimivuuden testaus

Sisällönkuvailun hyödyntäminen tiedonhaussa

- tiedonhaun kohdentaminen olemassa oleviin haettaviin kenttiin
- eri kenttien funktio tiedonhaussa
- vapaatekstihaku vs. kohdennettu haku
- käytettyjen kuvailuvälineiden (mm. luokitusten, asiasanastojen) hyödyntäminen tiedonhaun suunnittelussa

Sisällönkuvailun hyödyntäminen t&h-järjestelmien kehittämisessä

- dokumentaatiokielen integroiminen järjestelmään
- opasteiden teko
- luonnollisen kielen hyödyntäminen hakukäyttöliittymässä
- dokumentaatiokielen ja sen rakenteiden ”piilottaminen” hakujärjestelmään
- dokumentaatiokielen rakenteiden mallintaminen hakujärjestelmään
- spesifisyyden ja tyhjentävyyden optimointi

Asiakkaiden tekemä sisällönkuvailu

- Yleistä verkkopalveluluissa (tagging termi)
- Vapaa vs. kontrolloitu sisällönkuvailu
- Tiedonhaussa eri datatyyppien erottaminen

Automaattisen indeksoinnin menetelmät

- dokumenttien tekstien indeksointi tilastollismatemaattisilla menetelmillä
- ongelmana: miten dokumentin ydinsisältö pystytään löytämään
  - vrt. esim. koko teoksen tai sen tiivistelmän indeksointi
- näitä menetelmiä kehitteillä myös kuville ja äänelle
- lisäksi voidaan hyödyntää erilaisia bibliometrisia menetelmiä myös sisällönanalyysissä/tiedonhaussa

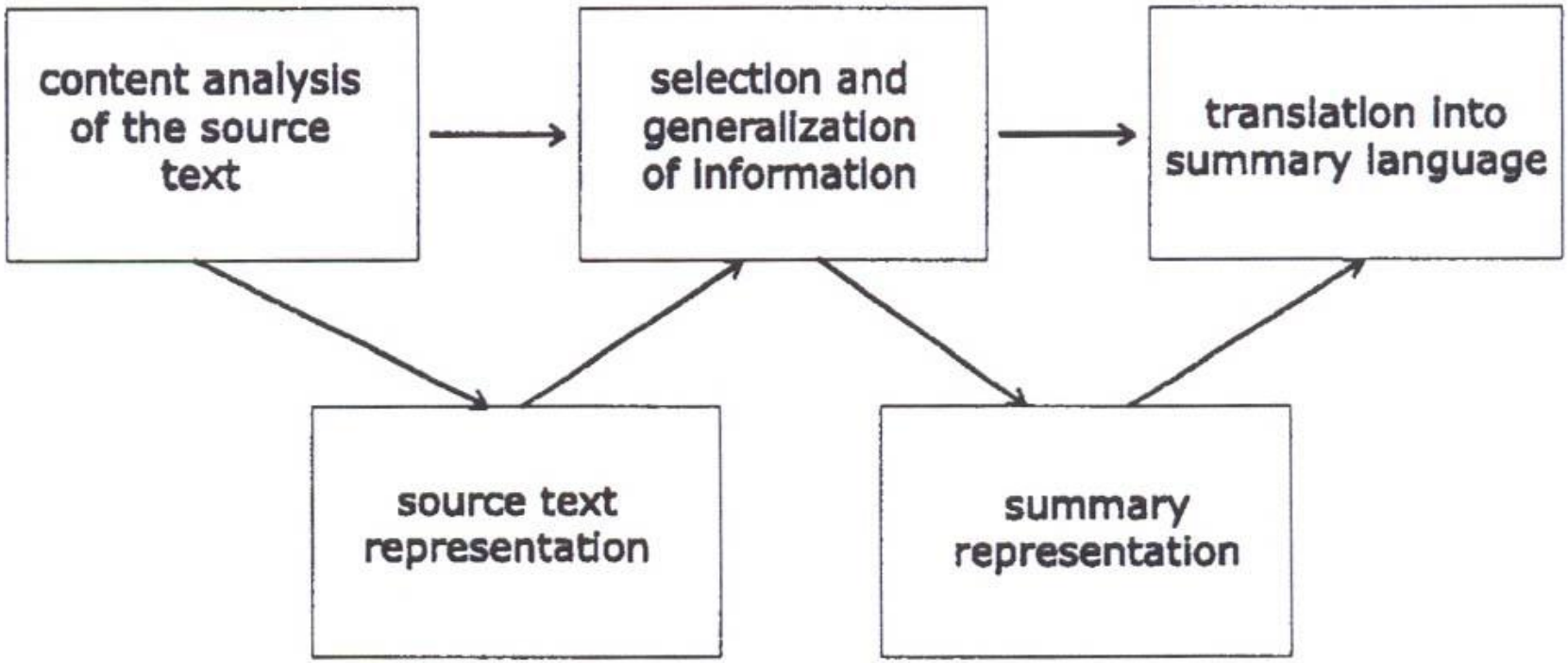


Figure 1. The process of automatic abstracting.

Tiedon tallennuksen organisointi kansallisesti ja kansainvälisesti

- kansalliskirjastojen tietokannat
- kaupalliset toimijat, mm. kustantajat
- kansainväliset standardit
  - tiedontallennukseen liittyvät
  - sisällönkuvailuun liittyvät
- ongelmia:
  - kulttuurien väliset erot
  - kielten väliset erot
  - alojen väliset erot

Kiitos!

Kysymykset nyt

tai

jarmo.saarti@uef.fi