CSC

ICT Solutions for Brilliant Minds

PAS-yhteistyöryhmä
28.3.2023

Kokouksen avaus

- Todetaan läsnäolijat
- Esityslistan hyväksyminen
- Edellisen kokouksen muistio

PAS-sanasto

PAS-sanaston työpaja

- PAS-sanastoa digitalpreservation.fi-sivustolla päivitetään
- YTR sopi joulukuun kokouksessa sanastotyöryhmän perustamisesta

- Sanastotyöryhmä kokoontui 9.3.2023
- Osallistujia hyödyntäviltä organisaatioilta 13 kpl
  - Kiitos kaikille osallistujille!
- Työpaja kävi läpi sanastotyön periaatteita ja käsitteli PAS-sanaston päivittämistä

Työpajan tulokset

- Hyväksyttiin PAS-sanaston periaatteet
- Käytiin läpi päivitettyä sanastoa
- Sovittiin, että sanasto on kokonaisuudessaan kommentoitavissa 31.3 saakka
- Sanasto julkaistaan päivitettynä huhtikuussa
- Työryhmä kokoontuu jatkossa aina tarvittaessa

PAS-sanaston periaatteet

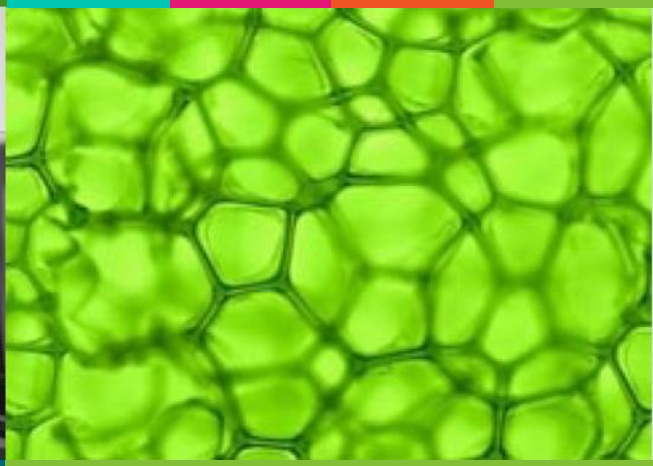
- Tavoite: PAS-sanasto sisältää termit ja käsitteet selityksineen siinä muodossa kuin niitä käytetään OKM:n omistamien PAS-palveluiden dokumentaatiossa ja ohjeistuksessa
- PAS-sanaston periaatteet:
  a) määritellään termit, jotka ovat käytössä tavoitteen laajuisesti
  b) määritellään termit siinä muodossa mitä niillä tarkoitetaan PAS-palvelujen kontekstin sisällä
  c) ei määritellä termejä, jotka kuuluvat yleissivistykseen
  d) sanaston käyttö ei edellytä toisen sanaston tuntemista
  e) sanasto ei saa olla ristiriidassa muiden sidosryhmien sanastojen kanssa

Sanasto kommentoitavana 31.3 saakka

- https://docs.google.com/spreadsheets/d/1shWhbW29VQ-Rb8xbccTcAq_gJI8UHE8AUhXSMUCAveI/edit?usp=sharing
- Kiitokset kaikille jotka ovat jo sanastoa kommentoineet

- Päivitetty sanasto julkaistaan huhtikuussa, mutta työ ei lopu siihen
  - Sanastoa päivitetään jatkossa tarpeen niin vaatiessa

Loogisen säilyttämisen edellytykset



Looginen taso ja bittien säilyttäminen

Loogisen säilyttämisen edellytykset

- Loogisen säilyttämisen vaatimuksiin liittyen tarvitaan hyödyntävien organisaatioiden näkemyksiä
- On tullut eri suunnilta toiveita siitä, että jokin aineisto halutaan siirtää bittitason säilytykseen
- Tyypillisesti syitä on kaksi:
  - Tiedostomuoto ei ole PAS-kelpoinen
  - Tiedosto on teknisesti viallinen

Nykytilanne bittitason säilytyksessä

- Aiemmin minkä tahansa muodon pystyi toimittamaan PAS-palveluun säilytyskelpoisen muodon rinnalla bittitason säilytykseen
- Myöhemmin toimintaa muutettiin niin, että tiedostojen analysointi antoi tiedostolle tason
  - Tietyille muodoille tiedosto saa tulla bittitason säilytykseen automaattisesti
  - Tasot: säilytyskelpoinen, siirtokelpoinen, bittitason kolme eri tilannetta
- Nyt vain niitä tiedostoja pystyy lähettämään bittitason säilytykseen, mitkä analysoinnissa luokitellaan bittitason säilytykseen
  - seurataan muotojen kirjoa
  - sovitaan tiedostomuodoista
  - tekninen tunnistaminen pitää olla mahdollista

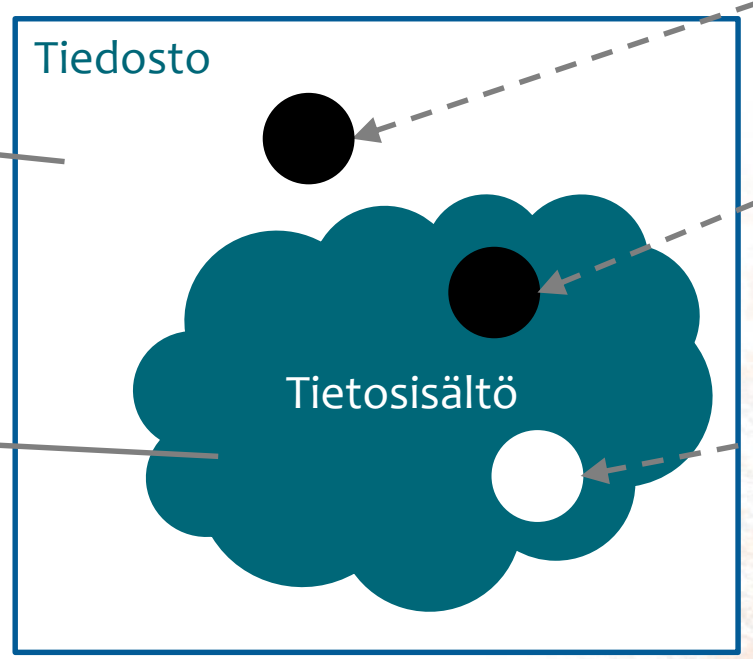
Loogisen säilyttämisen vastuu

- Valittu säilyttämisen taso käytännössä vaikuttaa vastuujakoon
  - Bittitason säilytykseen tulevan aineiston loogisen säilyttämisen vastuu on hyödyntävällä organisaatiolla
  - PAS-palveluiden arkkitehtuuriperiaate 7: *PAS-palvelut vastaavat säilytettävien aineistojen loogisesta ja bittitason säilyttämisestä sekä loogisen säilyttämisen tehtävien jaon sopimisesta hyödyntävien organisaatioiden kanssa.*
- Jos tiedosto on teknisesti rikki tai tuntematonta muotoa
  - Migraatiota ei välttämättä pystytä toteuttamaan normaalilla/tunnetulla migraatioprosessilla
  - Tiedoston kanssa voidaan joutua umpikujaan, jossa avaaminen on mahdotonta
  - Vaikka tiedostot saadaan tänä päivänä auki, näin ei välttämättä ole tulevaisuudessa
- Ohjelmistotuen kadotessa myöskään ehjiä tiedostoja ei välttämättä enää saa auki ilman loogista säilyttämistä

Virhe tiedostossa

Tietokoneen tarvitsema hyödyntäjälle näkymätön sisältö
- Yleensä liittyy tiedostomuotoon

Kohdeyleisön hyödynnettävissä oleva pitkäaikaissäilytettävä tietosisältö



Virhe tiedostossa, mutta ei tietosisällössä

Virhe tiedostossa ja tietosisällössä

Virhe tietosisällössä, mutta ei tiedostossa
- Semanttisen säilyttämisen asia
- Ei ole validaattorin mielestä virhe

Esimerkkejä eri virhetilanteista

- Tiedosto ja tietosisältö viallinen
  - Tiedoston siirto on jäänyt kesken
  - Videotiedoston kuvaframe ei ole tallentunut tiedostoon kokonaan (tavuja puuttuu)
- Tiedosto viallinen, mutta ei tietosisältö
  - EXIF-metatiedoista puuttuu pakollinen EXIF-määrityksen versionumero
  - Syntaksivirhe XML/HTML-tiedostossa
- Tietosisältö viallinen, mutta ei tiedosto
  - *Semanttisen tason asia: Validaattori ei ilmoita, ei ole tiedostovirhe*
  - Kirjoitusvirhe kirjan tekstissä
  - Videokuvassa näkyvät sisällölliset häiriöt (kohina, värienvaihtelu)
  - Valokuvassa osa kohteesta jäänyt kuvan ulkopuolelle

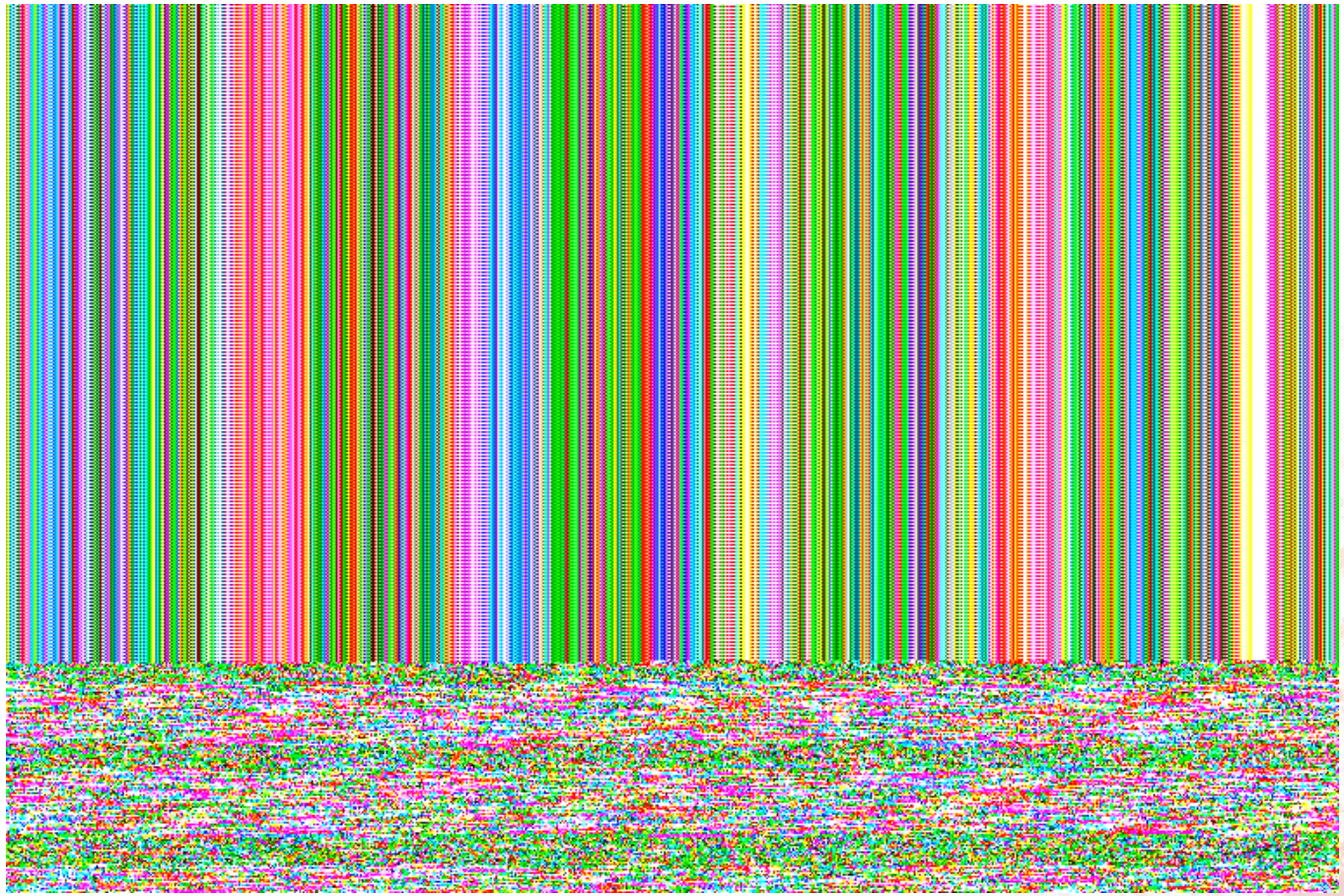
Esimerkki

Alkuperäinen kuva

Viallinen tiedosto avattuna:
Vain yksi bitti on muuttunut

…molemmilla sama tiedostokoko…

Esimerkki viallisesta kuvasta



Source:
Ikou, "A corrupted PCD photo file, converted to png.", Wikimedia Commons, Creative Commons CC0 1.0 Universal Public Domain Dedication.

Tarkistussumma ja tietosisältö

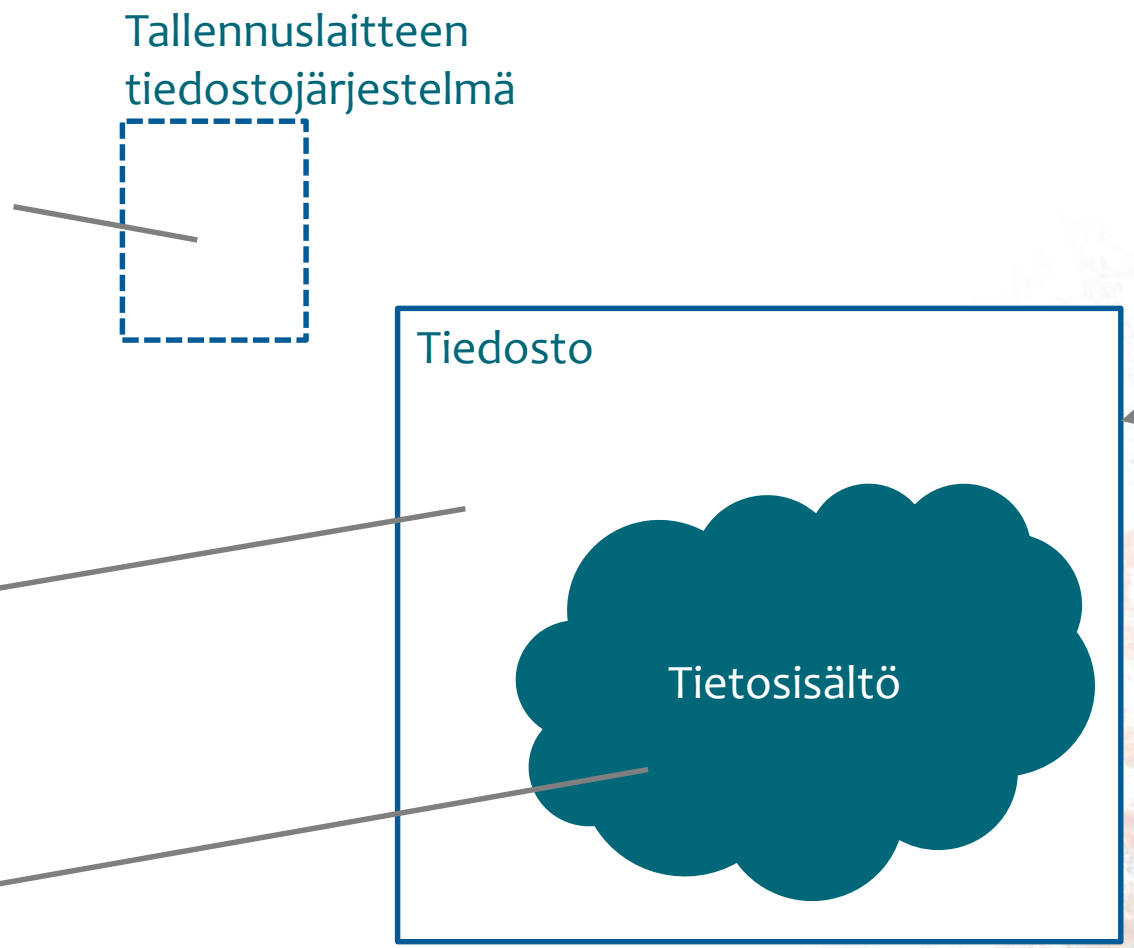
Tallennuslaitteen järjestelmän ylläpitämät tiedostoon sisältymättömät tiedostoa koskevat tiedot, kuten

- tiedostopolku ja –nimi
- tiedoston aikaleimat
- tiedoston luku- ja kirjoitusoikeudet

Tietokoneen tarvitsema hyödyntäjälle näkymätön sisältö

- Yleensä liittyy tiedostomuotoon

Kohdeyleisön hyödynnettävissä oleva pitkäaikaissäilytettävä tietosisältö



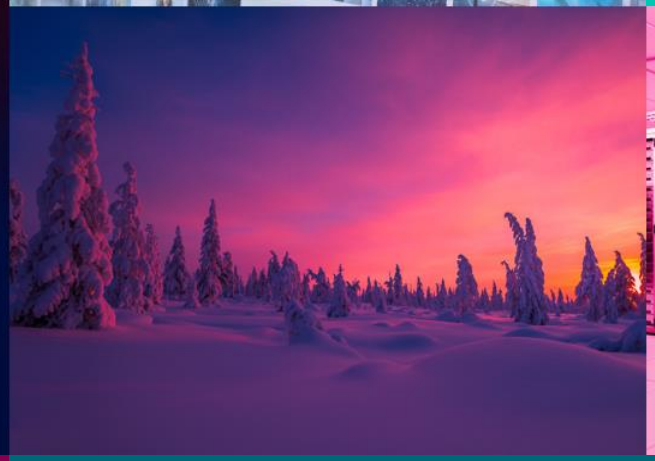
Tarkistussumma lasketaan koko tiedostosta. Summa muuttuu minkä tahansa tiedostossa olevan tiedon muuttuessa.

- Summa voi muuttua, vaikka säilytettävään tietosisältöön ei tulisi muutoksia.
- Summa ei muutu tallennuslaitteen järjestelmän tietojen muuttuessa (esim. tiedostonimi).

Työpaja

- Keskustellaan:
  - Millä periaatteilla tiedostoja voidaan ottaa bittitason säilytykseen?
  - Kuinka loogisen tason säilyttäminen on tällöin mahdollista?
  - Millä tavalla PAS-palvelu voi tukea organisaatioita niin, että loogisen säilyttämisen edellytykset täyttyvät?
- Erityisesti ei-tuettujen tiedostomuotojen ja viallisten tiedostojen osalta

PAS-palveluiden ohjelmistokerros



PAS-palveluiden ohjelmistokerros

- Viimeinen Python2 päivitys oli maaliskuussa
- AlmaLinux siirtymä hieman aikataulusta jäljessä
- Paketointikomponentin uudistus etenee

PAS-palveluiden uudet ominaisuudet Q1/2023

- ARC/WARC-migraation kehitystyö
  - Lähellä koemigraatiovaihetta
- Usean nauhakirjaston tuki
- Kehitystyö nauhamigraatiolle aloitettu
- PAS-määrittelyiden vuositarkastuksen muutokset
- PREMIS-kirjasto: tuki tiedoston luoneen sovelluksen kirjaamiseksi PREMIS-objektiin
- Räätälöityyn paketointityökaluun muutoksia
  - validointiin ja raportointiin
  - tunnisteiden käsittelyyn
  - tarkistussummien tulkitsemiseen
- Uudemmat ALTO-versiot XML-katalogiin
- FFMpeg-ohjelman käytön korjaus tiedostojen analysointityökalussa
- Bugikorjauksia ja muita pienempiä kehitystöitä
- SAPA-kehitystyö omassa projektiryhmässään
- Fairdata PAS
  - Uusi paketointityökalu edistynyt
  - PAS-määrittelyiden vuositarkastuksen muutokset
  - Kuvailussa annetun tapahtumahistorian käsittely
  - Selkeytystä hallintaliittymään
  - RHEL 9 / AlmaLinux 9 –työ on edistynyt
  - Bugikorjauksia ja muita kehitystöitä

Sopimukset – KP-PAS (1/2)

(8.3.2023)

Organisaatio	Kapasiteetti (Tt)	Aineistoa (Tt)	Täyttöaste	Säilytyspaketteja
Celia	110	85,10	77,36 %	31 312
Kavi	2 400	1 434,85	59,79 %	3 712
Kansallisarkisto Kansallisarkiston vastaanottamat alkujaan digitaaliset valtionhallinnon asiakirjalliset aineistot	41	1,23	3,00 %	1 706
Kansallisarkisto VAPA-järjestelmään siirretyt tietoaineistot	1	0,14	14,00 %	389
Kansallisarkisto Kansallisarkiston massadigitointi-hankkeen aineistot	114	80,39	70,52 %	67 908
Kansallisarkisto Kansallisarkiston digitaaliarkistosta siirrettävät aineistot ja takautuvan digitoinnin aineistot	805	289,80	36,00 %	568 764
Kansallisarkisto Kansallisarkiston yksinomaan digitaalisessa muodossa olevat yksityisarkistoaineistot	27	0	0,00 %	0

Sopimukset – KP-PAS (2/2)

(8.3.2023)

Organisaatio	Kapasiteetti (Tt)	Aineistoa (Tt)	Täyttöaste	Säilytyspaketteja
Kansallisgalleria	20	6,32	31,60 %	478
Kansalliskirjasto Kulttuuriaineistolain nojalla kerätyt aineistot	355	198,72	55,98 %	1 932 724
Kansalliskirjasto Kansalliskirjaston digitoimat kulttuuriperintöaineistot	175	14,73	8,42 %	3 892
KOTUS	60	9,83	16,38 %	407
Museovirasto	1	0,53	53,00 %	40 186
Musiikkiarkisto	70	0	0,00 %	0
Svenska litteratursällskapet i Finland	50	1,14	2,28 %	166
Yhteiskuntatieteellinen tietoarkisto	1	0,06	6,00 %	8 929
Yhteensä	**4 230,00**	**2 122,84**	**50,19 %**	**2 660 573**

Säilytyspäätökset – FD-PAS (1/2)
(8.3.2023)

Organisaatio	Kapasiteetti (Tt)	Aineistoa (Tt)	Täyttöaste	Säilytyspaketteja
Geologian Tutkimuskeskus GTK:n tomografialaitteen tuottamat tietoaineistot	12	6,39	53,25 %	131
Helsingin yliopisto Helsingin yliopiston SMEAR-aineistojen valikoima meteorologisia - ja ilmanlaatumittauksia	2	0,01	0,50 %	13
Helsingin yliopisto M. cinxia and C. melitaearum in the Åland metapopulation system	2	0,00	0,05 %	1
Helsingin yliopisto FIRE (The Finnish Reflection Experiment)	1	0,00	0 %	0
Helsingin yliopisto Luomuksen aineistot	150	0,00	0 %	0

Säilytyspäätökset – FD-PAS (2/2)
(8.3.2023)

Organisaatio	Kapasiteetti (Tt)	Aineistoa (Tt)	Täyttöaste	Säilytyspaketteja
Itä-Suomen yliopisto SENSOTRA	1	0	0 %	0
Jyväskylän yliopiston kiihdytinlaboratorio 250-Nobeliumin hajoamisspektroskopia	1	0	0 %	0
Oulun yliopisto, Sodankylän geofysikaalinen observatorio Havaintoaineistot	30	0	0 %	0
Tampereen yliopisto Yhteiskuntatieteiden tiedekunnan Kansanperinteen arkiston A-K-kokoelma	2	0	0 %	0
Turun yliopisto Historian, kulttuurin ja taiteiden tutkimuksen arkiston aineistot (HKT-arkisto)	20	0,25	0,01 %	1
Yhteensä	**221**	**6,40**	**2,90 %**	**146**

CSC:n PAS-työn uudelleenorganisointi

- CSC:n sisäisistä organisaatiomuutoksista johtuen, Kimmo jää sivummalle PAS-yhteistyöryhmästä
  - …mutta ei kuitenkaan katoa mihinkään sen kauemmaksi
- Jatkossa…
  - Heikki toimii puheenjohtajana
  - Mikko Vatanen toimii sihteerinä
  - Johan ja Juha jatkavat ”pysyvinä asiantuntijoina”

Ilmoitusasiat

- Päivitetyt määrittelyt julkaistu
- PAS-seminaari 24.-25.4.2023
- Blogi: Tutkimusaineistojen säilyvyyden hallinta
- iPres konferenssiin lähetty ehdolle julkaisu teemana PAS-palveluiden hiilijalanjälki

Kokouksen päättäminen

- Muut asiat
- Seuraava kokous
  - 15.6.2023 klo 12:30-15:00
  - Mahdollisesti hybridikokous
- Kokouksen päättäminen