

# Syntysähköisten asiakirjojen digitointi



Markus Merenmies / Kansallisarkisto



**KANSALLISARKISTO**  
RIKSARKIVET

# OSA I : Johdanto ja viitekehys



**KANSALLISARKISTO**  
RIKSARKIVET

# Lähtökohtia

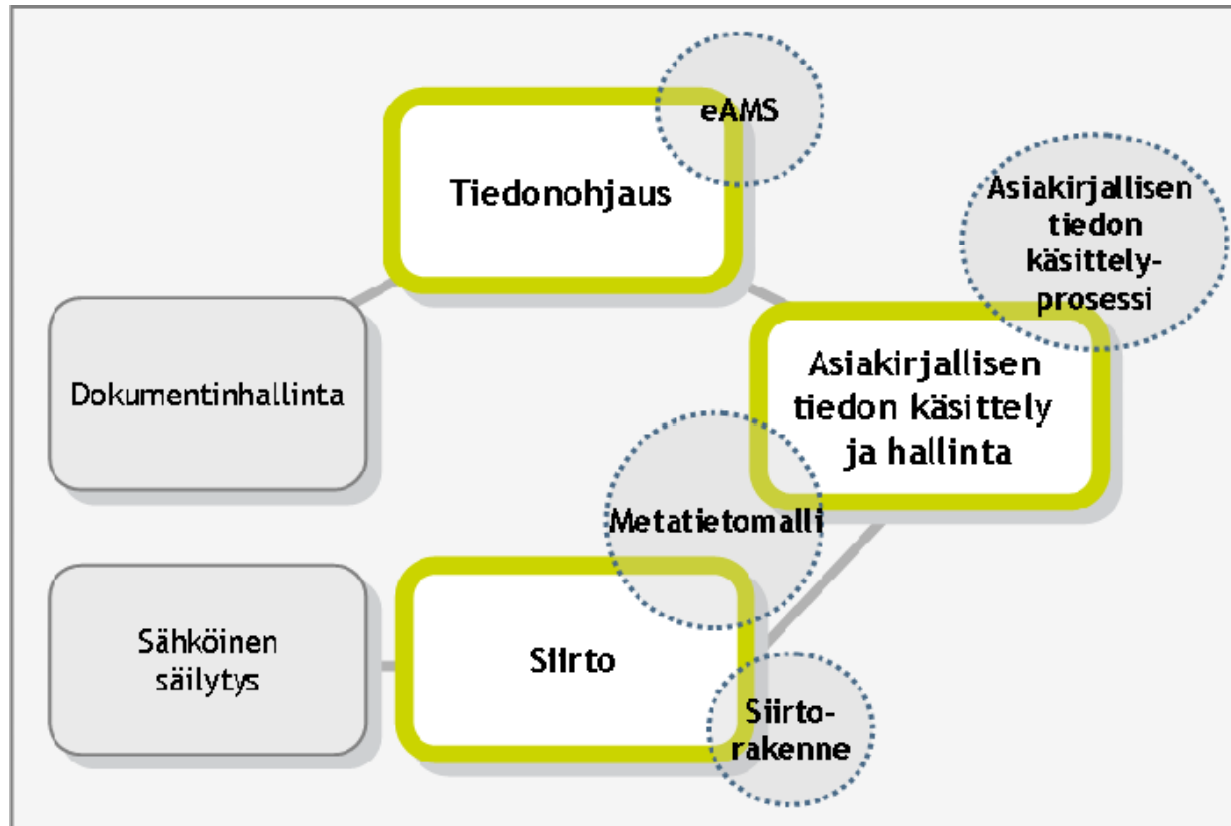
---

- Viranomaisten syntysähköisten asiakirjojen säilyttäminen yksinomaan sähköisessä muodossa edellyttää arkistolaitoksen lupaa
- Arkistolaitos on esittänyt vaatimukset asiakirjan hyväksytyistä formaateista sekä metatiedoista (Sähke)

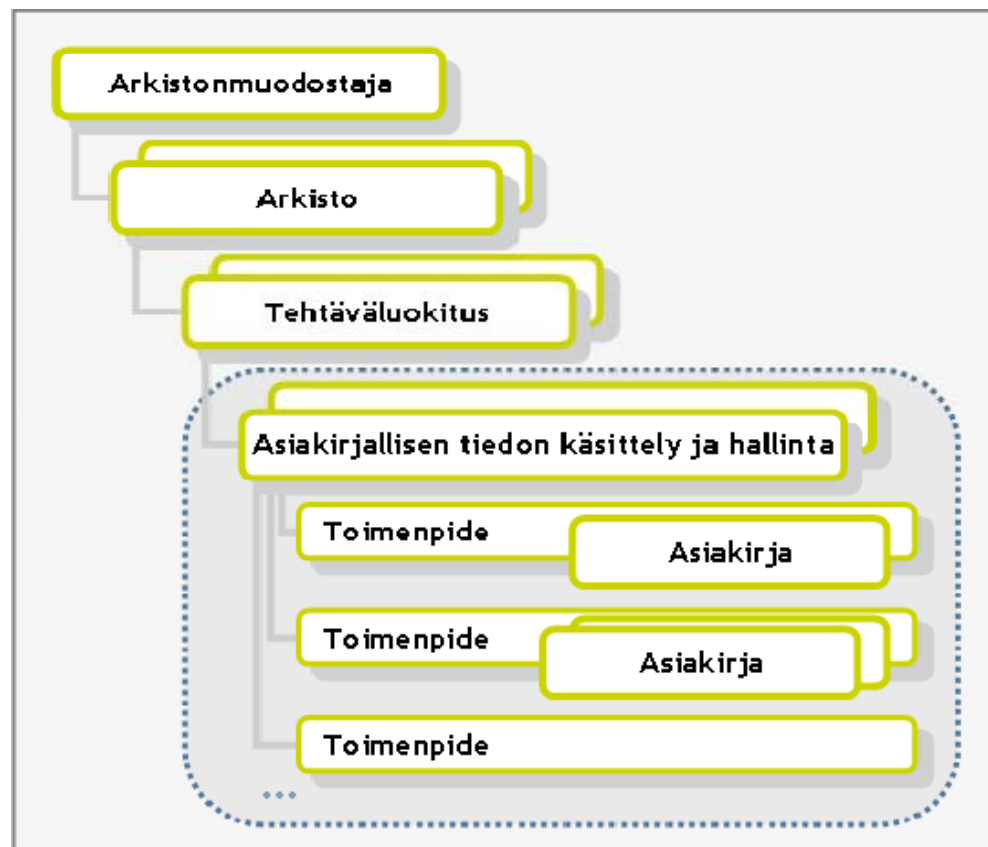


**KANSALLISARKISTO**  
RIKSARKIVET

# Viitekehys



# Sähke2-hierarkia...



ARCHIVVM



KANSALLISARKISTO  
RIKSARKIVET

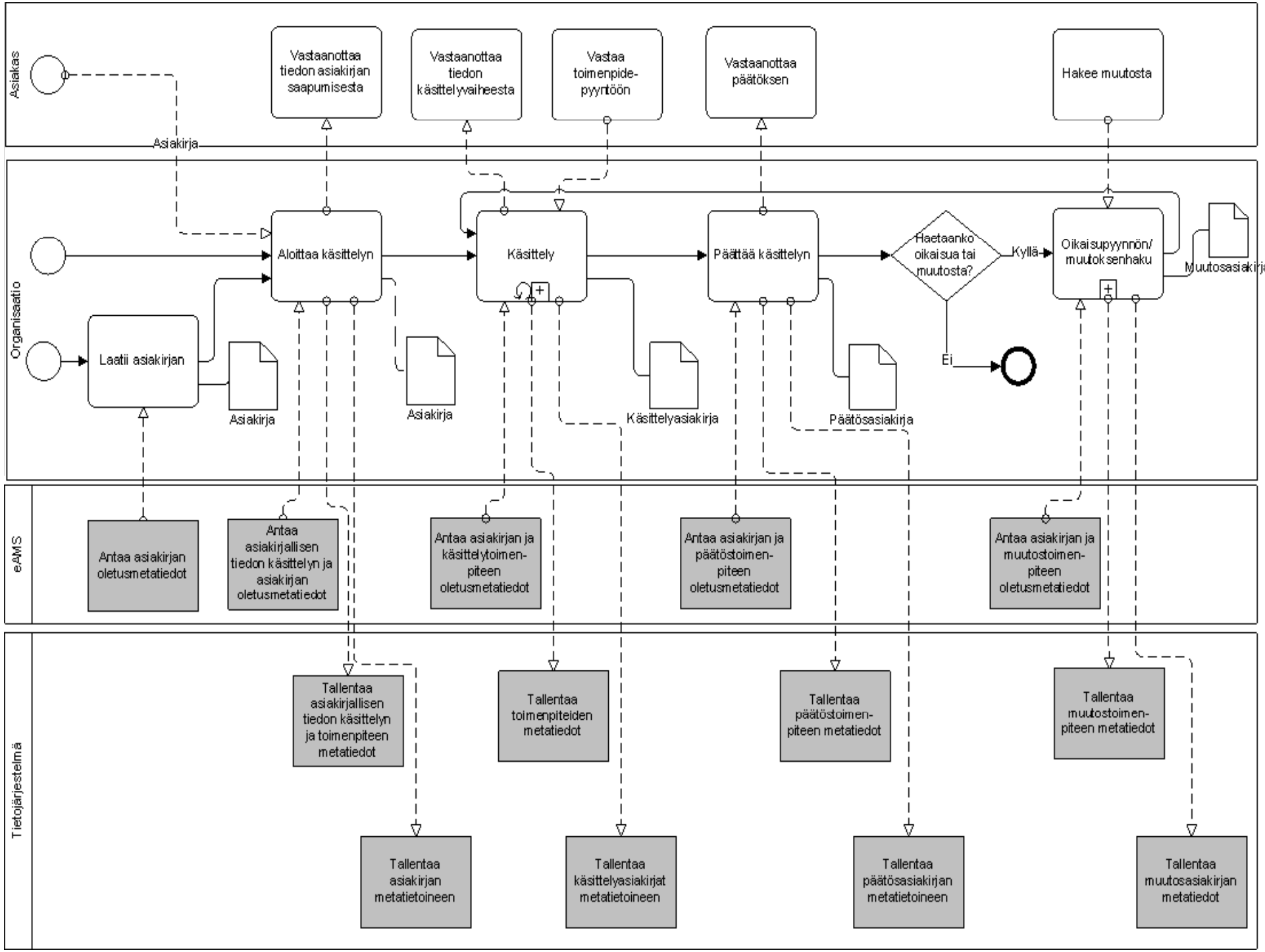
# Prosessin päävaiheet

---

- **Aineiston muodostaminen**
  - Formaattikorversiot
  - Järjestäminen paketeiksi
  - Yksilöinti
- **Siirron valmistelu**
  - Paketin allekirjoittaminen
  - Salaus ja/tai pakkaus
- **Siirto**
  - Off- tai On-line



**KANSALLISARKISTO**  
RIKSARKIVET



# OSA II : Metatiedot



**KANSALLISARKISTO**  
RIKSARKIVET

# Sähke-metatiedot

---

- Kuvailevan ja hallinnollisen metatiedon minimi
- Määrittää luovutuspaketin käsitteet
- Operatiiviset tarpeet voivat laajentaa → custom metadata



**KANSALLISARKISTO**  
RIKSARKIVET

# Organisaatio laatii asiakirjan

- **Tietojärjestelmän tuottamat metatiedot:**
  - Identifiointitunnus (2.3)
  - Kieli (2.4)
  - Laadittu (2.2.1)
  - Toimija (2.14)
  - Versio (4.11)
- **eAMS:n tuottamat metatiedot:**
  - Käyttörajoitus (2.6)
  - Säilytysaika (2.11)
  - Tehtävä (2.13)
  - Tila (2.12)
  - Tyyppi (2.15)
- **Käyttäjän tuottamat metatiedot:**
  - Nimeke (2.7)



# Asiakirjan saapuminen, alkuperäisyyden ja eheyden toteaminen

---

- Merkinnät asiakirjan vastaanottamisen sekä alkuperäisyyden ja eheyden toteamisesta tallennetaan asiakirjan metatietoihin:
  - Asiakirjan alkuperäisyys ja eheys todettu (4.2)
  - Asiakirjan sähköinen allekirjoitus (4.3)
  - Toimija (2.14)
  - Vastaanotettu (2.2.6)



**KANSALLISARKISTO**  
RIKSARKIVET

# OSA III : Tiedostot ja formaatit



**KANSALLISARKISTO**  
RIKSARKIVET

# Hyvän formaatin kriteerit

---

- **Avoimuus:** Tietämys ja dokumentaatio formaatista on helposti saatavilla
- **Hyväksyntä:** Formaatin muodollinen muistiorganisaatioiden hyväksyntä
- **Vakaus ja yhteensopivuus:** taakse- ja eteenpäin. Suojautuminen "murentumiselta". Formaatin ylläpito ja päivittäminen
- **Riippumattomuus:** sidonnaisuus toimijoihin tai muihin teknisiin ratkaisuihin.
- **Standardointi:** Muodollinen standardointiprosessi



**KANSALLISARKISTO**  
RIKSARKIVET

# Käytettävät formaatit

Formaatti	Nimi siirto-tiedostossa	Versioiden nimi siirto-tiedostossa	Pakkaus	Kuvaus
Unicode UTF-8 -tekstitiedosto	UTF-8	N/A	N/A	Vapaa, formatoimaton teksti, merkistö kattaa kaikki maailman kielet
TIFF Rev 5, Rev 6	TIFF	5,6	CCIT Group 3 (CCIT3), CCIT Group 4 (CCIT4)	Bittikarttakuva
PDF/A, ISO 19005-1:2005, IDT tai uudempi	PDF/A	1a, 1b, tulevaisuudessa myös muita	N/A	Muotoiltu asiakirja, voi sisältää myös kuvia
PCM WAV, 16 bit, 44.1 kHz	WAV	16 bit 44.1 kHz	N/A	Ääni
MPEG-1, Audio Layer 3, 128 kbit/s	MPEG1	AL3 128 kbit/s	N/A	Video
MPEG-2, 50 Mbit/s, 4:2:2	MPEG2	50 Mbit/s 4:2:2	N/A	Korkealaatuinen video



ARCHIVVM



KANSALLISARKISTO  
RIKSARKIVET

# Suosituksset

Asiakirjan luonne	Suosittelut formaatti	Muut mahdolliset formaatit
Teksti, johon ei sisälly kuvia tai muotoilua, esimerkiksi sähköpostit tai yksinkertaiset muistiot	UTF-8	PDF/A
Semanttiset XML-formaatit, esim. CDA	UTF-8	PDF/A
Muotoiltu teksti, johon voi sisältyä kuvia, esimerkiksi useimmat toimistosovelluksilla (Microsoft Office, OpenOffice yms) tuotetut asiakirjat	PDF/A	TIFF
Skannatut paperiasiakirjat	TIFF	
Kuvat	TIFF	
Ääni	WAV	
Video	MPEG1	MPEG2



ARCHIVVM



KANSALLISARKISTO  
RIKSARKIVET

# Miten toimia formaattien kanssa

---

- Sähke2 hyväksyy natiivin rinnakkaismuodon
- Konversio sp-muotoon mahdollisimman aikaisin
- PDF/A ja TIFF
- DOCX ja ODF ovat vielä avoimia kysymyksiä.



**KANSALLISARKISTO**  
RIKSARKIVET

# Mahdolliset ongelmat

Formaatti	Tyypillisiä virhetilanteita
UTF-8	<ul style="list-style-type: none"><li>• Virheellinen merkistömuunnos, joka tekee tiedostosta lukukelvottoman tai rikkoo osan merkeistä</li><li>• rivin- tai kappaleenvaihtojen katoaminen tai lisääntyminen</li></ul>
TIFF	<ul style="list-style-type: none"><li>• Puutteellinen resoluutio: teksti puuroutuu niin pahasti, että asiakirja ei ole luettavissa.</li><li>• Värihin liittyvät virheet: värejä katoaa tai ne vääristyvät niin paljon, että luettavuus kärsii.</li><li>• Asemointiin liittyvät virheet: elementti peittää toisen; elementti ylittää sivun reunan ja osa jää kuvasta pois.</li><li>• Tarpeettoman korkea resoluutio tai värisyys, mikä kasvattaa tiedostot tarpeettoman suuriksi.</li></ul>
PDF/A	<ul style="list-style-type: none"><li>• Asemointiin liittyvät virheet: elementti peittää toisen; elementti ylittää sivun reunan ja osa jää kuvasta pois.</li></ul>
WAV	<ul style="list-style-type: none"><li>• Tiedoston kokoon liittyvät virheet (kirjoitus katkeaa ennen aikojaan).</li><li>• Koodaukseen liittyvät virheet (kuva tai ääni sekoaa).</li></ul>
MPEG-1	<ul style="list-style-type: none"><li>• Tiedoston kokoon liittyvät virheet (kirjoitus katkeaa ennen aikojaan).</li><li>• Koodaukseen liittyvät virheet (kuva tai ääni sekoaa).</li></ul>
MPEG-2	<ul style="list-style-type: none"><li>• Tiedoston kokoon liittyvät virheet (kirjoitus katkeaa ennen aikojaan).</li><li>• Koodaukseen liittyvät virheet (kuva tai ääni sekoaa).</li></ul>



ARCHIVVM



KANSALLISARKISTO  
RIKSARKIVET

# Esimerkki

## **<s2:Document>**

**<s2:NativeId>**urn:oid:1.2.246.10.2048198.10.1.11.2009.89.1**</s2:NativeId>**

**<s2:UseType>**Arkisto**</s2:UseType>**

**<s2:File>**

**<s2:Name>**R0900089.pdf**</s2:Name>**

**<s2:Path>**2008245/pdf/R0900089.pdf**</s2:Path>**

**</s2:File>**

**<s2:Format>**

**<s2:Name>**PDF/A**</s2:Name>**

**<s2:Version>**1b**</s2:Version>**

**</s2:Format>**

**<s2:HashAlgorithm>**sha1**</s2:HashAlgorithm>**

**<s2:HashValue>**pC4LUmP5L6DRL4gUYTx2y82JanG=**</s2:HashValue>**

**<s2:Encryption>**Ei salattu**</s2:Encryption>**

**</s2:Document>**



ARCHIVVM



KANSALLISARKISTO  
RIKSARKIVET

# Esimerkki 2

---

```
<s2:Restriction>
  <s2:PublicityClass>Salassa pidettävä</s2:PublicityClass>
  <s2:SecurityPeriod>20</s2:SecurityPeriod>
  <s2:SecurityPeriodEnd>2029-02-05</s2:SecurityPeriodEnd>
  <s2:SecurityReason>Kalastuslaki,</s2:SecurityReason>
  <s2:SecurityClass>Ei turvallisuusluokiteltu</s2:SecurityClass>
  <s2:PersonalData>Sisältää henkilötietoja</s2:PersonalData>
  <s2:Owner>Rauno Murju</s2:Owner>
  <s2:AccessRight>
    <s2:Name>Rauno Murju</s2:Name>
    <s2:Role>henkilöstöpäällikkö</s2:Role>
    <s2:AccessRightDescription>2</s2:AccessRightDescription>
  </s2:AccessRight>
</s2:Restriction>
```



# OSA IV : Ohjeistus



**KANSALLISARKISTO**  
RIKSARKIVET

# Miksi ohje? – eli lähtötilanne

---

- Sähke2-kuvaa miten tulisi toimia
- XML-skeema kuvaa luovutuksen rakenteen
- Mutta, puuttui ohje miten toiminta välittyy skeemaan.



**KANSALLISARKISTO**  
RIKSARKIVET

# Ohjeen tarkoitus

---

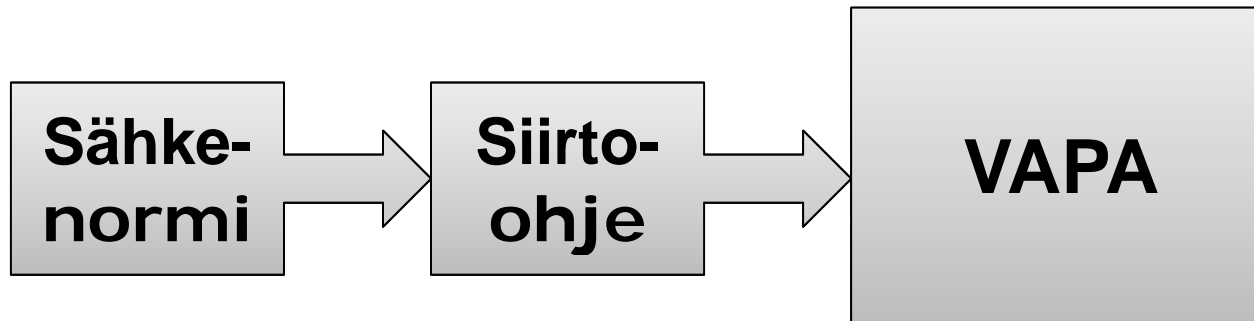
- Täydentää Sähke2-määräystä
- Kohdennettu materiaalia tuottaville organisaatiolle ja luonteeltaan tekninen
- Vastaa kysymykseen: "Miten luovutuspaketti (SIP) saadaan siirtokuntoon?"
- Vähentää epävarmuutta, tulkinnanvaraisuutta ja virheitä



**KANSALLISARKISTO**  
RIKSARKIVET



ARCHIVVM



**KANSALLISARKISTO**  
RIKSARKIVET

# Mitä kuvataan

---

- Eheystartkisteiden käyttö ja siirron allekirjoittaminen
- Aineiston identifioinnin periaatteet
- Konversiot säilytysmuotoon
- S2 mallin vastineet XML-skeemassa
- Luovutuspaketin siirtorakenne
- Aineistokokonaisuuksien lohkominen
- Siirron periaatteet



**KANSALLISARKISTO**  
RIKSARKIVET