TTA-hankkeen esittely

Pirjo-Leena Forsström
TTA-hankkeen pääsihteeri

Dataintensiivinen tiede

Tyypillistä:

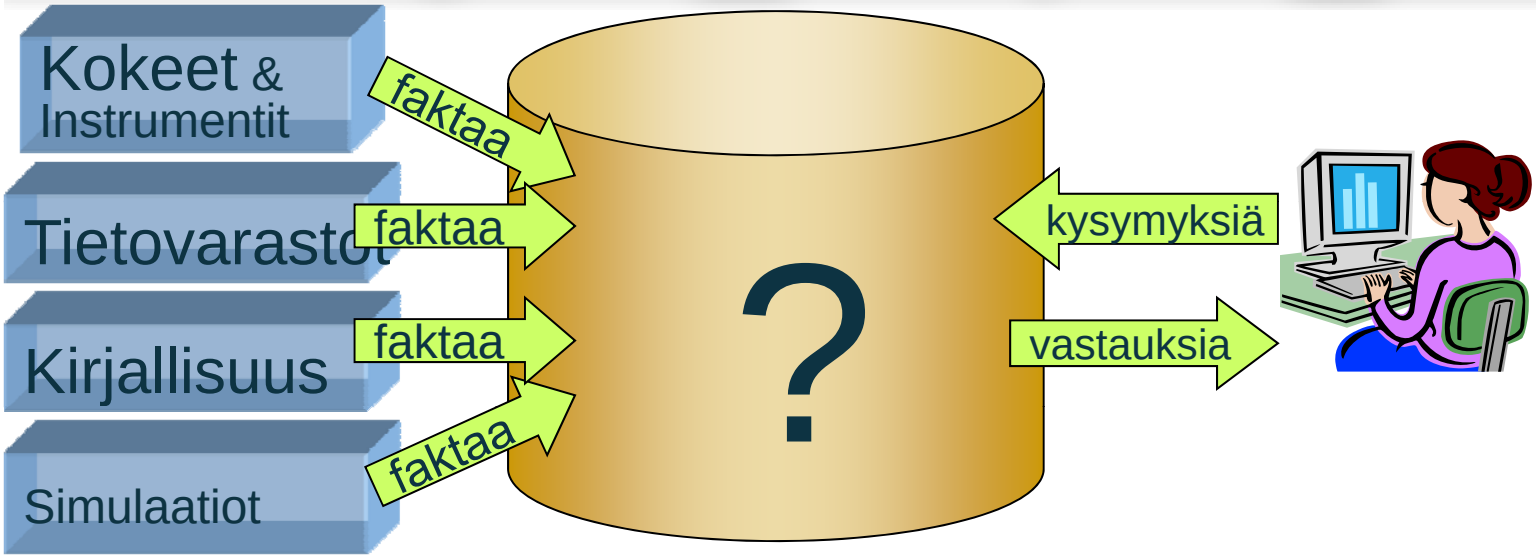
- Kasvava datan ja lähteiden määrä
- Datan käsittelyn kompleksisuus
- Datan suuri dynaamisuus
- Datan suuri käyttötarve
- Tutkijan ja datan monimutkainen vuorovaikutus

Tärkeimmät haasteet:

- Eksponentiaalisesti kasvavien datajoukkojen hallinta ja käsittely
- Analyysisyklin huomattava nopeutuminen
- Datalähteiden yhdistely

Kuva: wikipedia PD Image resources

DigiTutkijan ongelma



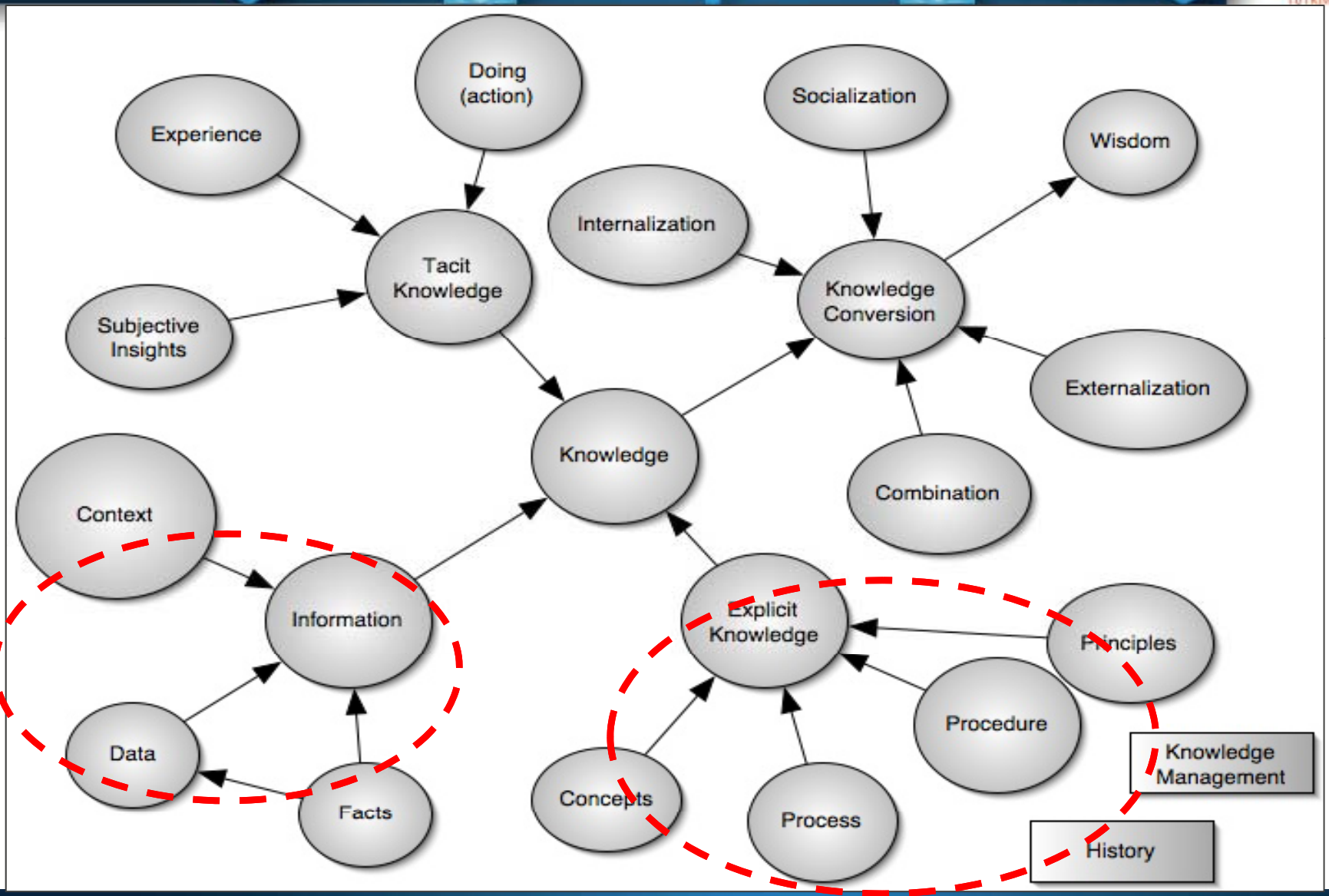
- Datan kokoaminen eri lähteistä
- Miten organisoida tieto?
- Miten säilöä vastaukset?
- Rinnakkaiselo ja yhteistyö muiden tutkijoiden kanssa?
- Louhinnan mahdollisuudet?
- Kyselyjen ja visualisoinnin työkalut?
- Tuki ja koulutus

Suorituskyky

- Vastaukset kyselyihin riittävän nopeasti
- Erittäin suurten massojen analysointi ja hallinta

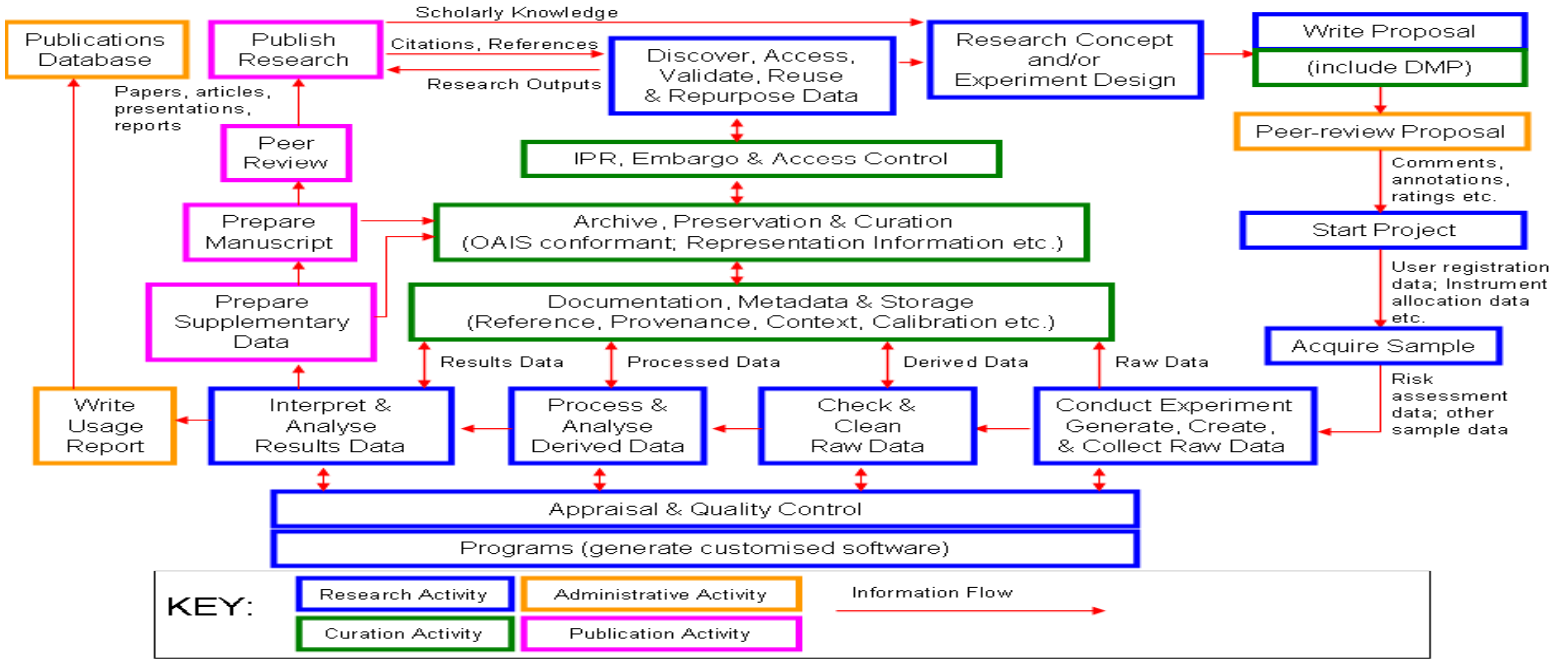
Kuva: wikipedia PD Image resources

Tiedon mallittaminen

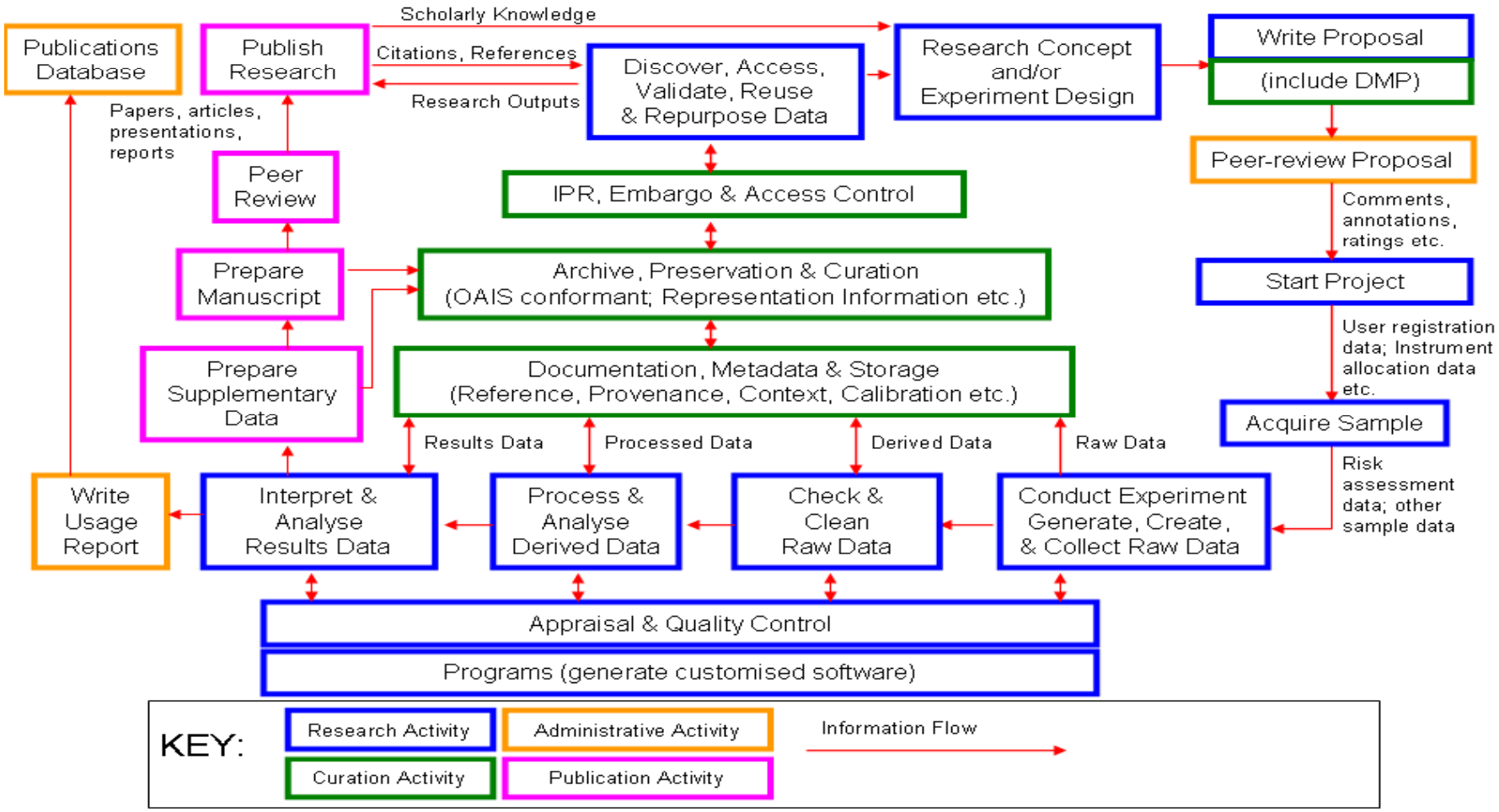


Mitä uutta dataintensiivisessä?

- Tuottamisen ja analyysin automaatio
- Läpitunkeva läsnäolo
- uudelleenkäyttö ja yhdistely
- Koko, Nopeus, Määrä

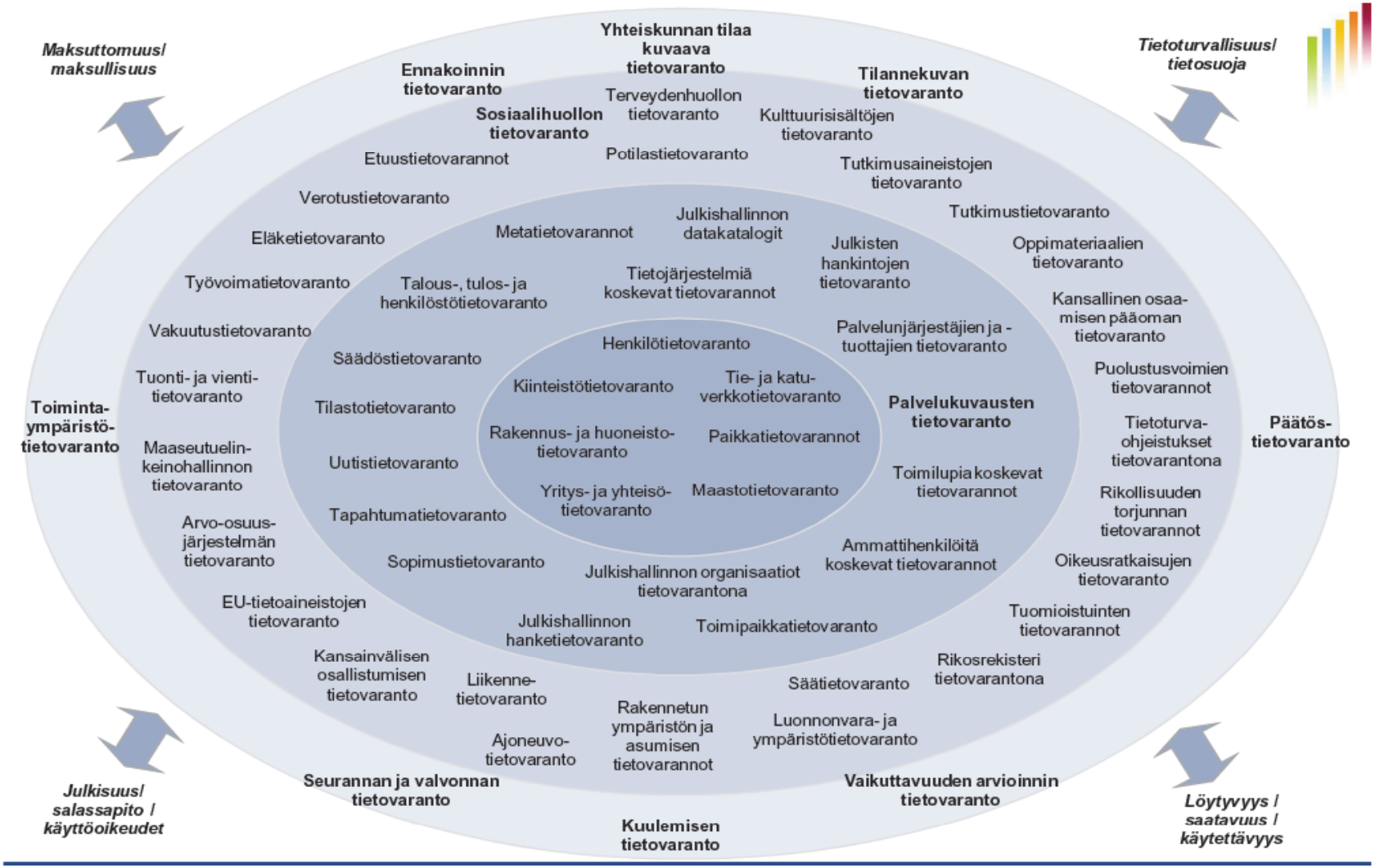


Tutkimustiedon prosessi

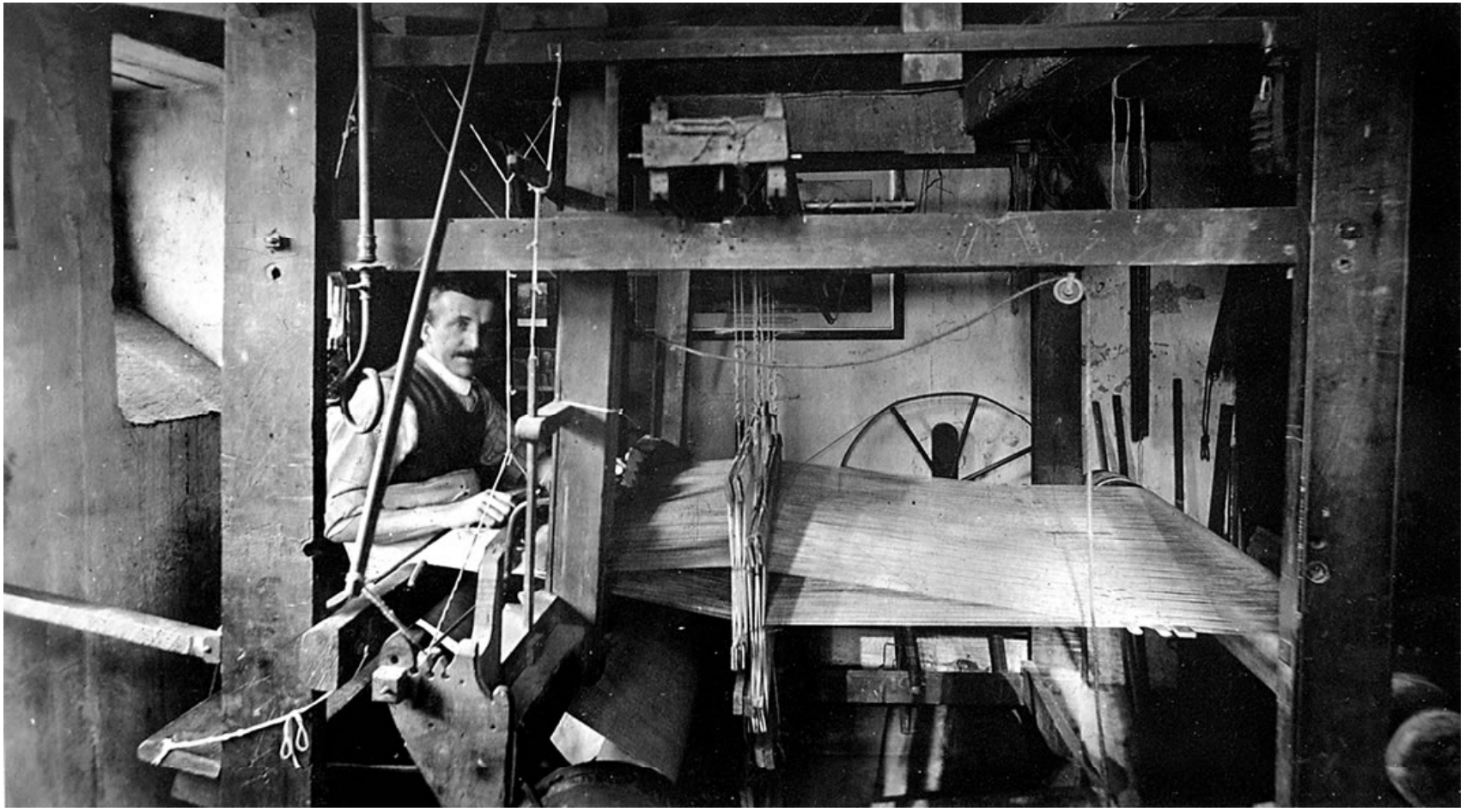


Lähde: JISC

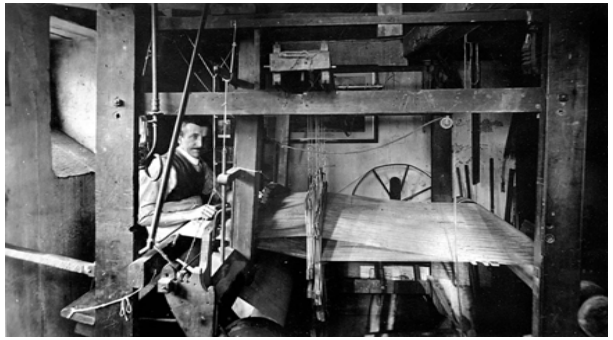
Yhä enemmän tietoa



Tutkimustiedon infrastruktuuri



Tutkimustiedon infrastruktuurin



- koko tutkimusprosessia
- tietoaineistojen siirtoa tieteenalojen välillä
- avoimen datan tuottoa
- Työvuo-pohjaista toimintaa
- julkaisujen ja tietoaineistojen linkitystä
- tiedepolitiikan kehikkoa

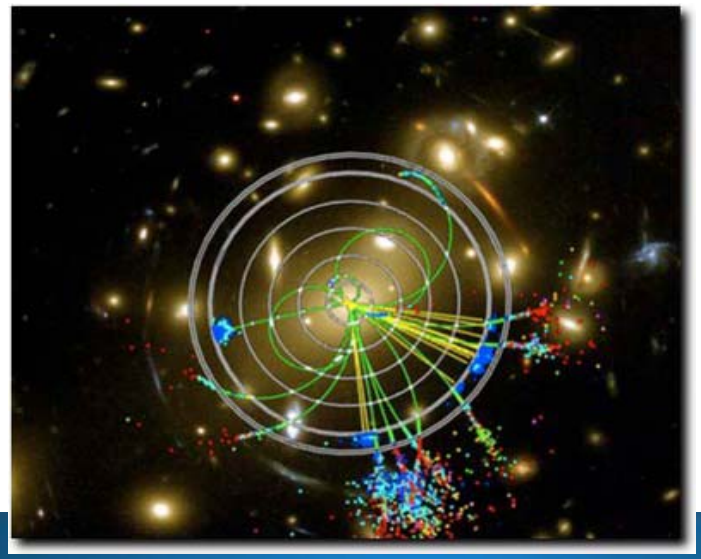
...jos tutkimustieto ei säily

- Tutkimusaineistot ovat korvaamaton tutkimus- ja kehitystyössä
- Systemaattinen hallinta on tärkeää koska:
  - Tukee monia tutkimustoiminnan vaiheita
  - Liittyy moneen osa-alueeseen (esim. temaattiset keskittymät, kokoelmat...)
  - Antaa uskottavuutta ja luo luottamusta tutkijaan, organisaatioon, toimijoihin

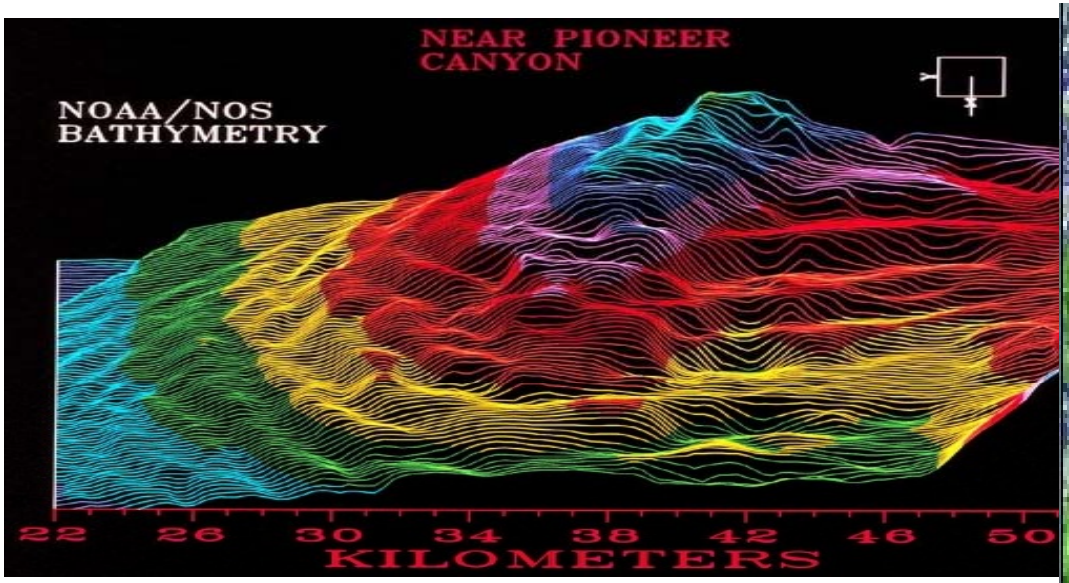
Dataintensiivinen tiede

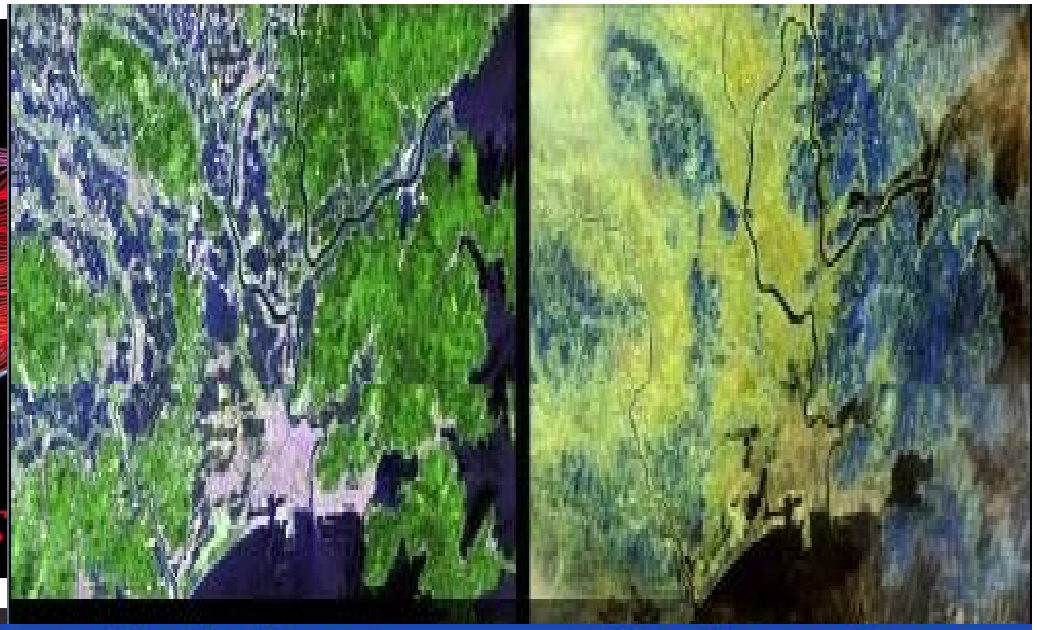
Tärkeimmät haasteet:

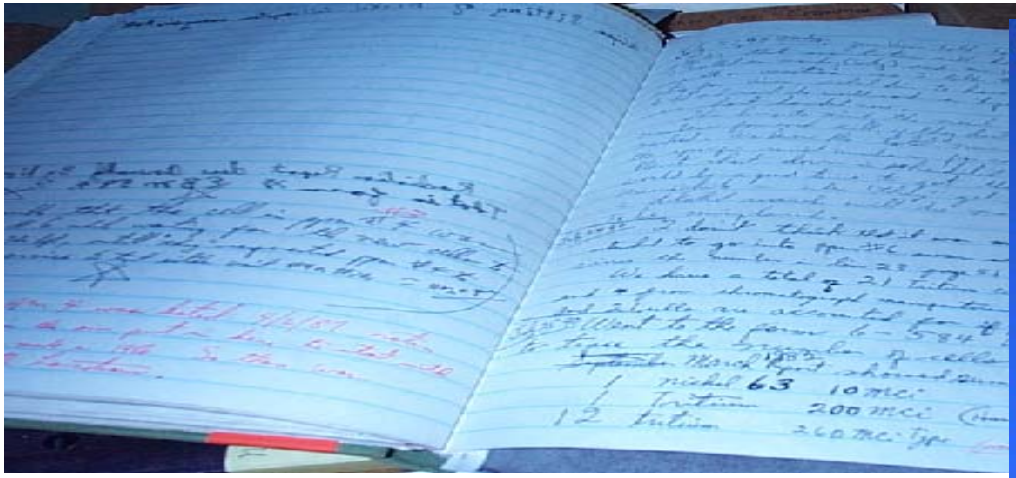
- Eksponentiaalisesti kasvavien datajoukkojen hallinta ja käsittely
- Analyysisyklin huomattava

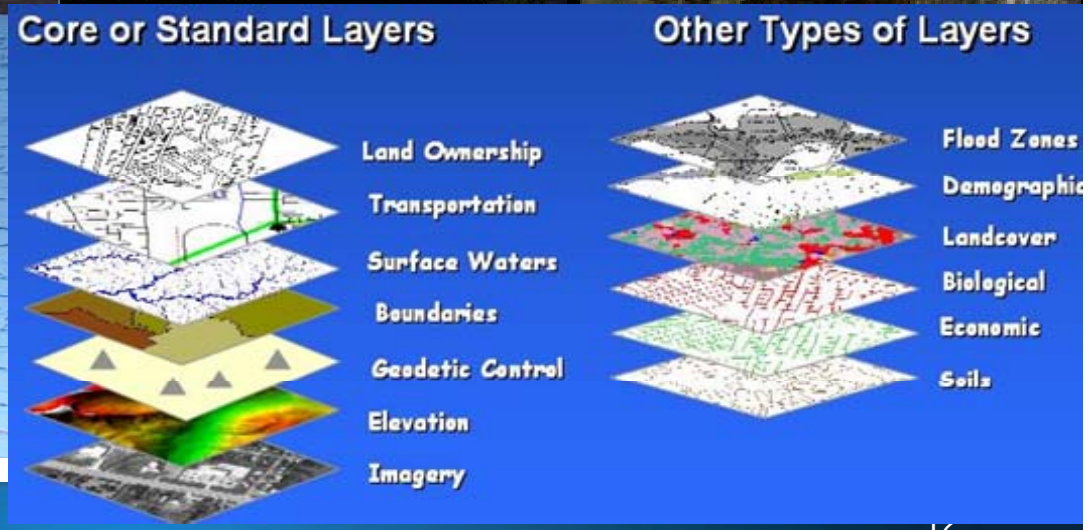


Monessa muodossa









Monella medialla

Kysymyksiä

1. Teknologiset:
   a) Datan mallinnuksen haasteet
   b) Datan hallinnan haasteet
   c) Työkalujen haasteet
2. Organisatoriset
3. Systeemiset:
   a) Virtuaaliset tutkimusympäristöt
   b) Yhteentoimivuus, heterogeenisuuden ongelmat
   c) Välitysohjelmistot
   d) Infrastruktuuripalvelut
4. Uudet paradigmat
5. Policy –haasteet: tieto päätöksenteon tukena

Digitaalisen tiedon turvaaminen

Pitkäaikaissäilyttäminen:

Tiedon elinkelpoisuuden turvaaminen jatkuvalla muutoksella

Jakelu ja uudelleenkäyttö:

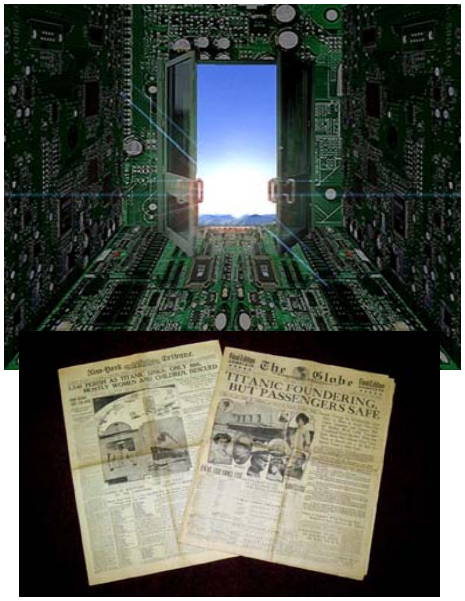
Tiedon hyödyntämisen mahdollistaminen ja edistäminen, tietotuotantoon jo tehtyjen investointien hyödyntäminen

Digitaalinen pitkäaikaissäilyttäminen

•prosessi, jossa digitaalinen kokonaisuus irroitetaan luontiympäristöstään, ja sen olemassaolo varmistetaan säilytysympäristössä autenttisuus ja eheys säilyttäen

•Kriittiset asiat:
- •Autenttisuuden ja eheyden varmistaminen
- •Teknologisten riskien hallitseminen
- •Kustannusten hallitseminen

•Päämääränä infrastruktuuririippumattomuus, jotta voidaan käyttää mitä tahansa tallennusratkaisua



HAASTEET:

Mitä pitää säilyttää?
- Materiaali joka pitää ”pelastaa”
- Materiaali jota arvioimme tarvitsevamme pitkällä aikavälillä

Kuinka se tulisi säilyttää?
- Formaatit
- Tallennusmediat
- Kuka on vastuussa, kuka tekee?

Kuka maksaa?
- Sisällön tuottajat?
- Laitokset ja instituutiot?
- Käyttäjät?

Haasteet digitaalisessa säilyttämisessä

- Mitä pitää säilyttää?
  - Materiaali joka pitää ”pelastaa”
  - Materiaali jota arvioimme tarvitsevamme pitkällä aikavälillä
- Kuinka se tulisi säilyttää?
  - Formaatit
  - Tallennusmediat
  - Kuka on vastuussa, kuka tekee mitäkin?
- Kuka maksaa?
  - Sisällön tuottajat?
  - Laitokset ja instituutiot?
  - Käyttäjät?
- Kuka pääsee aineistoihin?

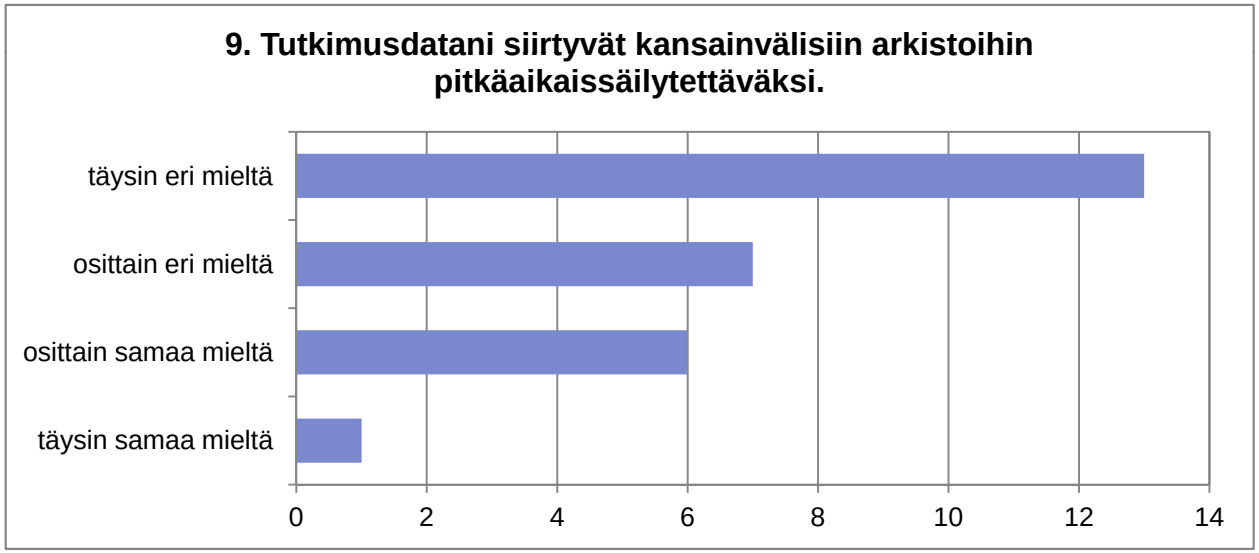
Print media provides easy access for long periods of time but is hard to data-mine

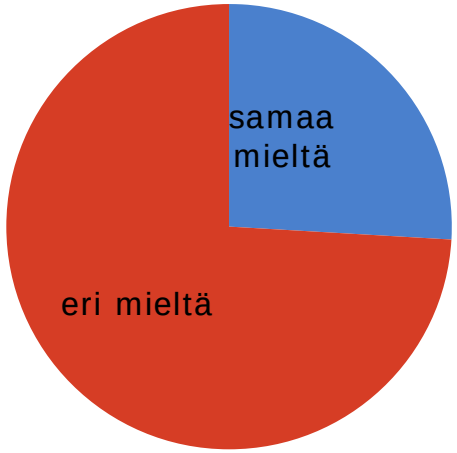
Digital media is easier to data-mine but requires management of evolution of media and resource planning over time

Monen tieteenalan viesti

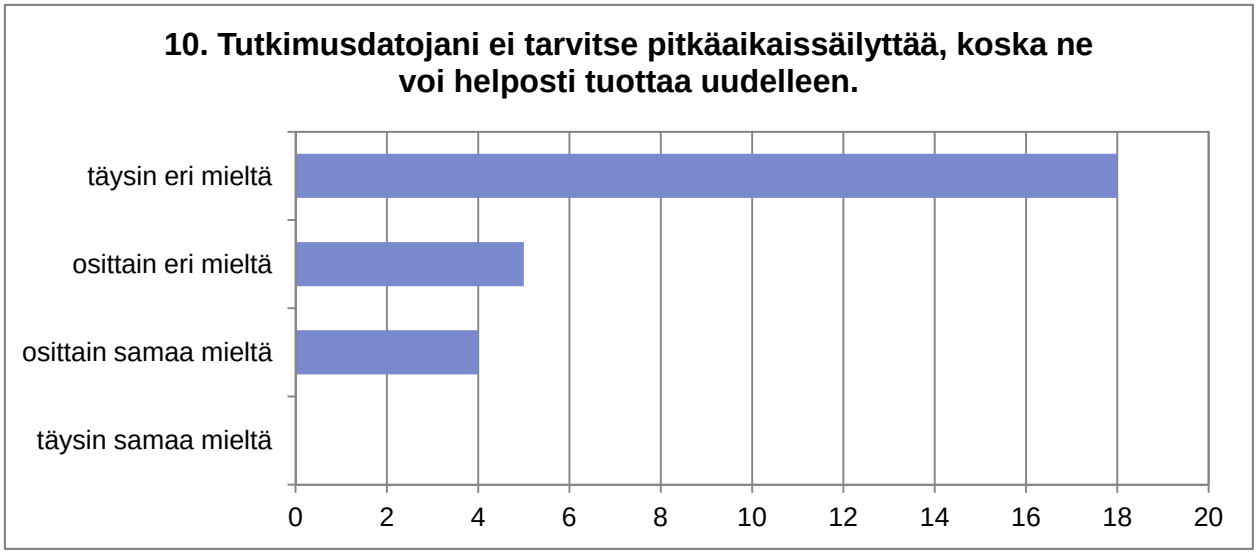
- Säilytykseen tarvitaan kiireellisiä toimenpiteitä
- Säilytyttäväksi suositellaan datasettejä joka mahdollistavat uudelleenkäytön
- Omistajuuden ja kuratoinnin vastuiden tulee olla selkeästi määritetty
- Tarvitaan kaikkien toimijoiden yhteistyötä
- Ohjausmalli ei saa olla liian raskas

9. Tutkimusdatani siirtyvät kansainvälisiin arkistoihin pitkäaikaissäilytettäväksi





10. Tutkimusdatojani ei tarvitse pitkäaikaissäilyttää, koska ne voi helposti tuottaa uudelleen



Kansainvälisiä säilytyspaikkoja

- CERN
- EBI ja EMBL
- World Data System

Miten hukata digitaalista informaatiota, lyhyt oppimäärä:

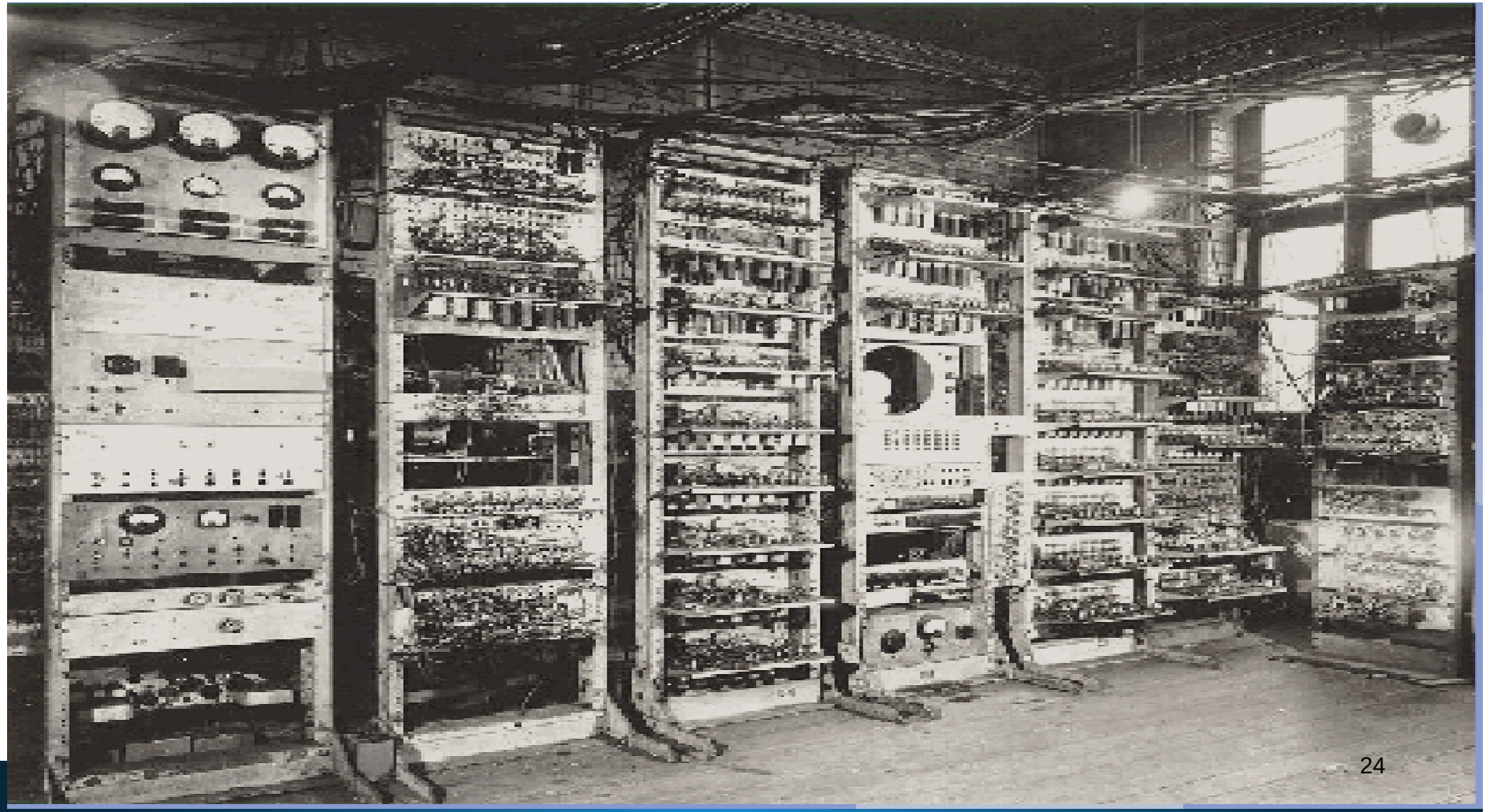
Se on hukattu kun

- Et voi lukea sitä
- Et voi tulkita sitä oikein
- Et voi varmistaa sen alkuperää
- Et löydä sitä

Säilytysvaatimukset ovat täällä..

- 7 year Retention
  - Payroll
  - Inventories
  - Expense Reports
  - Safety Records
  - IRS Audits
  - Payables
  - Tax Work Papers
  - Receivables
  - Retirement
  - History
  - Audited Finance Stmts
- 15 year Retention
- 40+ year Retention
  - Medical Records
  - Accident Reports
  - Pension Contributions
- Permanent Retention
  - Contracts & Leases
  - Legal Correspondence
  - Insurance Reports
  - Patent Docs
  - Property Appraisals
  - Training Manuals
- U.S. Army
  - Support 2.4 Million Soldiers
  - Retain Life of Service + (62 years)
- European Union – Dir. 2006/24/EC
  - Retain all electronic communications info (VoIP) including phone #, user IDs, equipment etc.
  - Retain 6 month to 2 years
  - Starting 3/2006 – Delon until 3/2009

...mutta valmiit ratkaisut eivät ole



Säilyttäminen on monimutkaista

- Teknologiset haasteet
  - Fyysinen ymmärrettävyys (=koneelle)
  - Looginen ymmärrettävyys
  - Suurten digitaalisten arkistojen migraatio
  - Erilaisten tiedostomuotojen/rakenteiden emulaatio
  - Historiallisten sovellusten ja ratkaisujen ylläpito
  - Muutoksilta suojaaminen
  - Vahingoittumisen tai hukkaamisen estäminen
  - Fyysinen ja looginen turvallisuus
  - Automatisointi
  - Tuhoaminen ja uudelleensijoitus
  - Hakeminen ja löytäminen
  - Testaus ja auditointi

Muita haasteita

- Luokittelu ja valinta
- Riittävän metatiedon kerääminen
- Käytäntöjen standardointi
- Arkistojen arvon löytäminen ja määrittäminen
- Ammatillinen meriitti ja tunnustus
- Yhteistyö vaatimuksien määrittelyssä (=valinnat)

Pitkäaikaissäilytysympäristön pitää huolehtia näistä:

- Autenttisuus
- Eheys
- Teknologioiden muutokset
- Riskien hallinta
- Säilytysmetatiedon hallinta
- Ratkaisun skaalautuvuus

Pitkäaikaissäilytyksen riskit

- **Tekniset riskit**
  - Mm. tallennusmedian tai laitteiston rikkoutuminen, ohjelmistovirheet
- **Inhimilliset riskit**
  - Ylläpitäjän virhe
  - Hyökkäys järjestelmää vastaan (organisaation ulkoa tai sisältä)
- **Hallinnon riskit:**
  - Huonot valinnat ja päätökset
- **Erityisesti pitkäaikaissäilytykseen liittyvät riskit**
  - Tiedon ymmärrettävyyden häviäminen
  - Piilevät virheet ("bit rot")
- **Katastrofit**
  - Luonnonilmiöt: tulva, tulipalo ym.
  - Yhteiskunnalliset: Sotatila, vararikko ym.

Conclusions

- silent corruptions are a fact of life
  - first step towards a solution is detection
  - complete elimination seems impossible
- existing datasets are at the mercy of Murphy
- effort has to start now (if not started already)
- correction will cost time AND money
  - multiple cost-schemes exist:
    - trade time and storage space (à la Google)
    - trade time and CPU power (correction codes)
  - the best protection is probably a combination of redundancy and correction codes

Pitkäaikaissäilytyksen tekniseltä infrastruktuurilta vaadittavat ominaisuudet

- Järjestelmässä ei saa olla yksittäistä kohtaa, jonka pettäessä koko järjestelmä pettää (single point of failure)
- Järjestelmän täytyy toimia, vaikka mediat, ohjelmistot ja laitteistot vaihtuvat tasaisena virtana
  - Järjestelmän täytyy tukea erilaisuutta ja välttää lukkiutumista tiettyjen laite- tai ohjelmistotoimittajien ratkaisuihin
- Järjestelmän täytyy jatkuvasti seurata tietopakettien eheyttä
- Suurinta osaa järjestelmässä olevaa aineistoa käytetään harvoin, ja se aiheuttaa suuren riskin piilevien virheiden kerääntymiseen
- Järjestelmän tulee olla ylläpidettävissä pääosin katkotta

Pitkäaikaissäilytyksen projektit hyötyvät näistä:

- Selkeä organisoituminen ja vastuutus
- Standardien noudattaminen
- Kansainvälinen yhteistyö
- Teknologiainfrastruktuurin syvällinen tuntemus
- Exit-strategia!

Riskien minimointi

- Keinoja riskien minimoimiseksi
  - Useampi kuin yksi tallennusmediatyyppi (esim. kiintolevy, nauha, optiset levyt, kertakirjoitettavat mediat)
  - Automaattinen informaation eheyden tarkistus ja virheiden korjaus
  - Avoimet rajapinnat, avoimen lähdekoodin ohjelmistot
  - Turvalliset verkko- ja tilaratkaisut
  - Järjestelmän toiminnan valvonta, lokitietojen keruu ja analysointi
  - Kopioiden hajautus maantieteellisesti eri paikkoihin, eri kopioiden tallennusprosessien riittävä itsenäisyys
  - Järkevä rahankäyttö
- PAS-organisaation on laadittava riskianalyysi ja riskienhallintasuunnitelma
  - Riskianalyysia on myös määräajoin päivitettävä

Standardit, lait ja normit

- Kansainväliset standardit
  - Open Archival Information System (OAIS) -viitemalli
  - Trustworthy Repositories Audit & Certification (TRAC)
  - Muita mm. MoReq2, METS, ISO-laatustandardit
- Kotimaiset lait ja normit
  - Julkisuuslaki, henkilö- ja tietosuojalait, tekijänoikeuslaki, valmiuslaki
  - VAHTI-tietoturvaohjeet ja -määräykset
- Muut hyvät toteutuskäytännöt, mm. tietoturvan osalta
  - Paljon materiaalia tarjolla, kannattaa valikoida ja hyödyntää!

Miten kustannuksia voi minimoida?

- Prosessien automatisointi
  - Metadatan tuotanto
  - Luettelointitoiminto
- Ajan säästöt: perusteellinen käsittelytyö alkuvaiheessa säästää aikaa myöhemmin
  - Prosessien standardointi
  - Tarkka suunnittelu
  - Yhteistyö ja tiedon jakaminen: oppimiskäyrän huomioiminen

Et voi astua samaan virtaan kahdesti

- Teknologia muuttuu
  - Migroi tiedostomuodot
  - Virkistä data
  - Reagoi muutoksiin laite- ja ohjelmistotasolla
  - ➔Oppimiskäyrät, eivät useinkaan näy riskeissä ja kustannuksissa
- Anna teknologian auttaa – automatisoi kaikki minkä voit
- Ajan säästöt: perusteellinen käsittelytyö alkuvaiheessa säästää aikaa myöhemmin
- Säilytyksellä ja kuratoinnilla on elinkaari
- Jatkuvuus tarkoittaa ettemme tiedä lopputulosta – open end

»Ποταμοῖς τοῖς αὐτοῖς ἐμβαίνομέν τε καὶ οὐκ ἐμβαίνομεν, εἶμέν τε καὶ οὐκ εἶμεν.
Me sekä astumme että emme astu samoihin jokiin, olemme ja emme ole.»
Herakleitos

Kiitos!

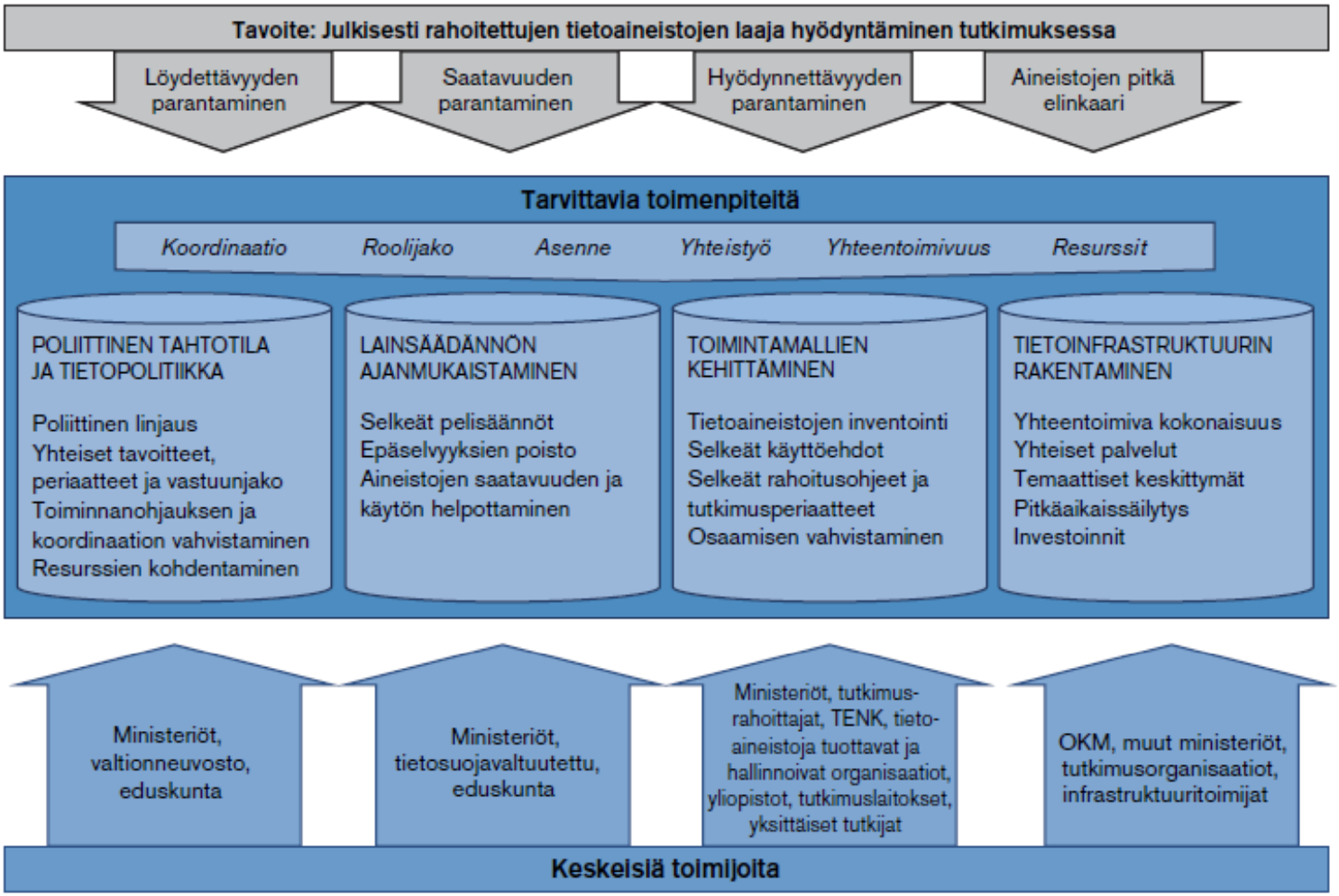
Kuva: wikipedia PD Image resources

Tutkimusaineistojen pitkäaikaissäilytys

- Tutkimustiedon **kumuloituminen** on tutkimuksen ja tutkimusorganisaatioiden toiminnan ydin
  - Nykyisin ei ole käytössä hallittua ja toimivaa tapaa huolehtia digitaalisesta tiedosta pitkällä aikavälillä
- Digitaalisten aineistojen pitkäaikaissäilytys tarkoittaa digitaalisen informaation **luotettavaa säilyttämistä kymmenien tai jopa satojen vuosien ajan**
  - Laitteet, ohjelmistot ja tiedostomuodot vanhenevat, mutta tästä huolimatta informaation täytyy säilyä ymmärrettävänä.
- Tutkimusaineistojen pitkäaikaissäilytyksessä tehdään tiivistä **yhteistyötä** Kansallisen Digitaalisen Kirjaston (KDK) kanssa

Tutkimuksen tietoaineistopyramidi

Toimenpiteitä tietoaineistojen hyödyntämisen parantamiseksi



TTA-VISIO

- *Suomessa on selkeä tietopolitiikka, jota yhteiset sähköiset palvelut tukevat.*
- *Julkisen sektorin hallinnoimat tietoaineistot sekä julkisen tutkimusrahoituksen tuella syntyneet tietoaineistot ovat lainsäädännön ja yhtenäisten käyttöehtojen ohjaamina pääsääntöisesti maksutta yhteiskunnan hyödynnettävissä.*
- *Pitkäjänteinen tietoinfrastruktuurin kehittämisen ja ylläpidon rahoitusjärjestelmä takaa, että olemassa olevat ja uudet tietoaineistot kuvaillaan ja ne ovat helposti löydettävissä ja otettavissa käyttöön tietoverkossa olevien palveluiden avulla.*
- *Kannustava ja oikeudenmukainen meriittijärjestelmä varmistaa uusien, laadukkaiden tietoaineistojen liittämisen tietoinfrastruktuuriin.*

Määritelmiä

- **Tutkimuksen tietoaineistolla tarkoitetaan tässä työssä julkisin varoin tuotettuja sähköisiä tietoaineistoja ja -varantoja. Tietoaineistoihin kuuluvat sekä tutkimuksen tuottamat että tutkimuksen hyödynnettävissä olevat aineistot.**

- **Tutkimuksen tietoinfrastruktuurilla tarkoitetaan aineistojen sijaintiin, rakenteeseen, organisointiin, hakemistoihin ja luetteloihin, omistajuuteen, saatavuuteen, varmistukseen, tietoturvaan ja tietovarastoihin liittyvät keskeiset resurssit ja kyvykkyydet sekä niiden elinkaaren hallinta.**

- **TTA-palveluilla tarkoitetaan niitä tutkimuksen tietoinfrastruktuurin palveluita, jotka toteutetaan TTA-hankkeessa. Joulukuuhun 2012 mennessä tällaisiksi palveluiksi on sovittu tallennuspalvelu IDA, metatietopalvelu KATA sekä pitkäaikaissäilytys PAS. Jatkossa uusista palveluista sopiminen tapahtuu valitun hallintomallin puitteissa**

- Käytännön tavoitteena on:
  - tutkimuksen tietoinfrastruktuuri palveluineen toimii saumattomasti yhteen muun kansallisen tietoinfrastruktuurin kanssa ja tarjoaa tutkimuksen tietoaineistojen säilyttämisen ja hyödyntämisen yhteiset palvelut.
  - Tietoinfrastruktuurin rakentaminen, kehittäminen ja ylläpito on pitkäjänteistä ja takaa, että tietoaineistot kuvataan ja tuodaan tietoinfrastruktuuripalvelujen piiriin.
  - Eri toimijoiden välinen roolijako on selkeä.
  - Kaikista tutkimuksen kannalta merkittävistä tietoaineistoista on tuotettu tarvittavat metatiedot ja kuvaukset on koottu niin, että tietoaineistot on helposti löydettävissä.
  - Tietoaineistot säilytetään pysyvästi ja niille on annettu pysyvä tunnus.
  - Tietoaineistot ovat palveluiden avulla helposti löydettävissä, saatavissa ja käytettävissä.
  - Tutkimusorganisaatiot ja yhteenliittymät vastaavat tietoaineistojen metatietotyöhön tarvittavista työkaluista ja järjestelmistä.

TTA-hankkeen hyödyt

- Keskitettyä tietoa tutkimuksen tietoaineistoista: helpompi löytää, helpompi käyttää
- Yhtenäisempiä käytäntöjä aineistojen hallintaan
- Yhteentoimivuuden lisääntyminen: metatietomalli, rajapinnat
- Monipuolisen palvelukokonaisuuden kehitys
- Tietoaineistojen säilymisen turvaaminen

Tier 1 – International data services

Tier 2 – National data services

Tier 3 – Institutions (Universities & Institutes)

Tier 4 – “Small science” researchers & research groups

Tutkimuksen tietoaineistot TTA

TTA-hanke edistää tietoaineistojen **kuvausten yhtenäistämistä, säilytystä ja käyttöä**.

TTA-hankkeessa tuotetaan tätä tukemaan:

- *tutkimuksen tietoaineistojen prosessikartta*
- *tietoinfrastruktuurin palvelukokonaisuuden suunnittelu*
- *metatietomalli tutkimustiedon hallinnalle*
- *yhteinen metatietokatalogi/hakupalvelu tutkimusaineistoille*
- *tallennuspalvelu tutkimuksen tietoaineistoille*
- *yhteisen pitkäaikaissäilytysratkaisun valmistelu*

TTA-palvelukokonaisuus edistää tutkimuksen tietoaineistojen kansallista, eurooppalaista ja kansainvälistä **yhteentoimivuutta.**