Pitkäaikaissäilytyksen haasteet hiukkasfysiikassa

Kati Lassila-Perini

CMS-kokeen tutkimusaineiston pitkäaikaissäilytyksen ja avoimen saatavuuden koordinaattori

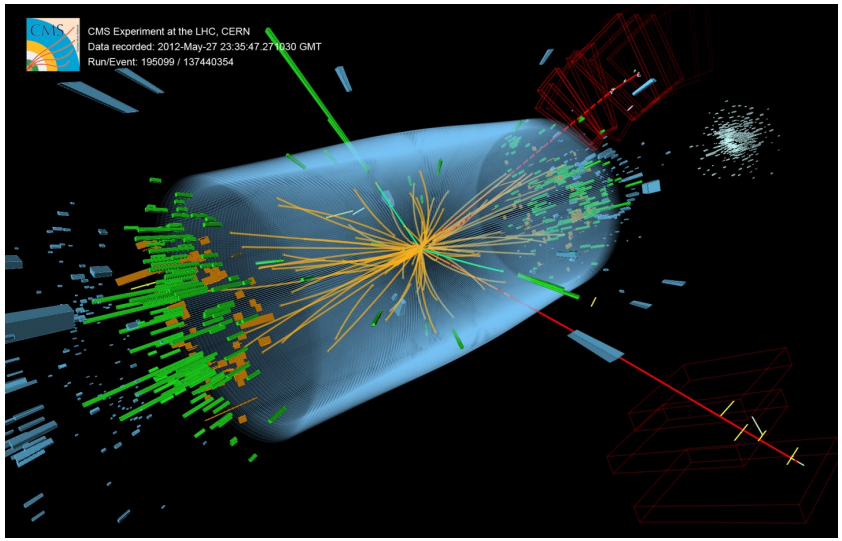
Fysiikan tutkimuslaitos

- Minkälaista on hiukkasfysiikassa käytettävä tutkimusaineisto?
- Miten tutkimusaineistoa käytetään?
- Mitä erityispiirteitä hiukkasfysiikan tutkimusaineistolla on?

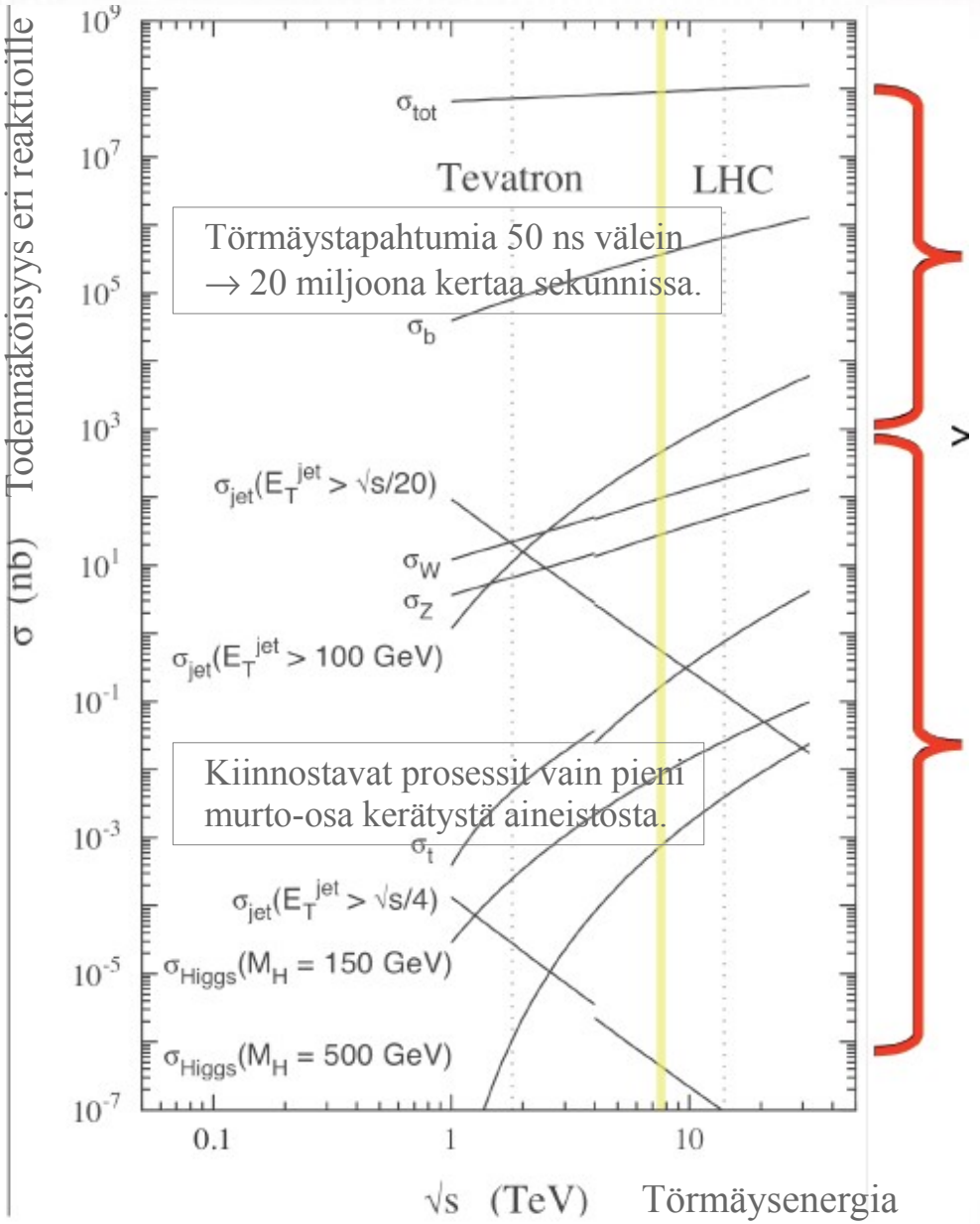
- Hiukkasfysiikan pitkäaikaissäilytyksen haasteet
- CMS-kokeen julkinen tietoaineisto TTA-AVOIN-hankkeen pilottina – avoimen tiedon hyödyntäminen pitkäaikaissäilytyksen *primus motor*ina.

Minkälainen aineisto?

- Hiukkasfysiikan kokeet keräävät dataa hiukkastörmäyksistä:



  - raakadata: impulsseja, koordinaatteja, aikoja – ei sinällään käyttökelpoista
  - *rekonstruoitu* data: törmäyksessä syntyneet hiukkaset, niiden energia ja liikemäärä kussakin kokeessa omassa, mutta yhtenäisessä formaatissa.
    - Tulokset saadaan analysoimalla rekonstruoitua dataa kokeen omalla analyysiohjelmistolla. Käytetään myös suuria määriä *simuloitua* dataa tulosten vertailuun.



Online-valinta tiedonkeruun aikana: talteen yksi 10^5 törmäystapahtumasta

Tallennettu aineisto

Offline-valinnat analyysin aikana: esim. Higgsin hiukkasen etsintä – valitaan yksi 10^5-10^6 tallennetusta törmäystapahtumasta

Miten aineistoa käytetään?

- Tutkimusaineiston kulku esim. CMS-kokeessa:
  - tallennettava raakadata suodatetaan *online*-valinnalla, kerätään ja tallennetaan CERNin T0-laskentakeskukseen, kalibroidaan ja rekonstruoidaan analysoitavaan muotoon
  - rekonstruoitu data lähetetään T1-laskentakeskuksiin, joilla on tallennusvastuu ja joissa
    - datasta suodatetaan eri analyysiryhmien tarvitsemaa aineistoa
    - data rekonstruoidaan uudestaan tarvittaessa (esim. uudet, paremmat kalibraatiot)
  - analyysi tehdään T2-keskuksissa, jotka saavat datan T1-keskuksista.
- Tutkija lähettää analyysiohjelmansa verkon kautta siihen T2-keskukseen, jossa tarvittava aineisto on tarjolla.
- Yksittäinen tutkimus (data-analyysi) kestää useasta kuukaudesta yli vuoteen.

Aineiston erityispiirteet

- Kokeet ovat suuria ja maailmanlaajuisia, esim. CMS:
  - lähes 3000 tutkijaa
  - 38 maata
  - lähes 200 tutkimuslaitosta

Map showing CMS collaborator countries around the world

  - laskenta- ja toimintamallin on taattava aineiston nopea saatavuus kokeen tutkijoille ympäri maailmaa.
- Volyymit ovat suuria, esim. CMS:n resurssit v. 2012:
  - raakadata, useita prosessointeja, simuloitu data

T0 Disk [TB]	1000
T0 Tape [TB]	23000
T1 Disk [TB]	22000
T1 Tape [TB]	45000
T2 Disk [TB]	26000

- Aineisto on yhtenäisessä formaatissa.
- Aineiston tulkintaan tarvitaan analyysiohjelmistoja, käyttö vaatii erityisosaamista.
- Aineiston on oltava helposti saatavilla koko yksittäisen analyysin ajan.
- Aineistoa kerätään lähes kaksi vuosikymmentä.

Aineiston erityispiirteet pitkäaikaissäilytyksen kannalta

:

- Toimivat, tehokkaat rakenteet tietoaineiston jakeluun jokapäiväisessä käytössä.
  - datakatalogit, tietoaineiston siirtotoiminnot.
- Laskentakeskukset, jotka ovat sitoutuneet suurten tietovolyymien keskipitkän aikavälin säilytykseen.
- Tietoaineisto homogeenista (sama tiedostomuoto käytössä myös eri kokeissa).
- Ei luottamuksellista tai vaarallista tietoa.

:

- Arkistoitavan aineiston käsittely kilpailee resursseista uuden aineiston kanssa.
- Tietoaineisto on monimutkaista ja sen tulkitsemiseen tarvitaan erityisosaamista (aineiston sisältö, analyysiohjelmistot, kokeellisen hiukkasfysiikan tuntemus).
- Tarvittavan tietotaidon dokumentointi keskittyy tämän päivän problematiikkaan, ei arkistoitavan aineiston käytön ohjeistukseen.

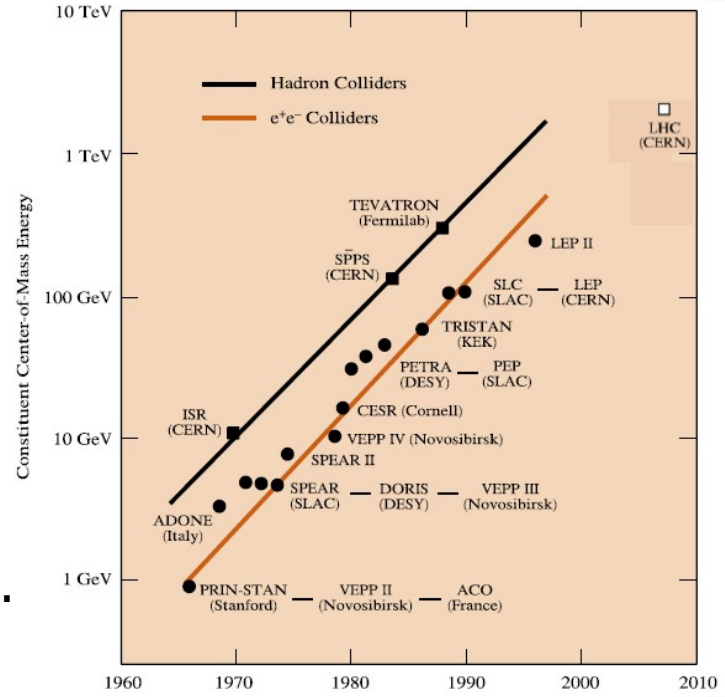
Tutkimusaineistojen tasot

- **Data ≠ tietoa:** pitkäaikaissäilytyksen yhteydessä hiukkasfysiikassa määritellään digitaalinen tutkimusaineisto seuraavasti:
- **Taso 1:**
  - Tutkimustulokset *open access* -julkaisuna ja niihin liittyvät lisämateriaalit, esim. kuvaajat ja taulukot numeerisessa muodossa.
- **Taso 2:**
  - Yksinkertaistetut pienet tietoaineistot *outreach*-toimintaan ja opetukseen.
- **Taso 3:**
  - Rekonstruoitu tietoaineisto ja **ohjelmistot, joita voi käyttää niiden analysointiin sekä käyttöön tarvittava dokumentaatio**.
- **Taso 4:**
  - Raakadata ja **ohjelmistot, joilla aineisto voidaan rekonstruoida analysoitavaan muotoon sekä käyttöön tarvittava dokumentaatio.**

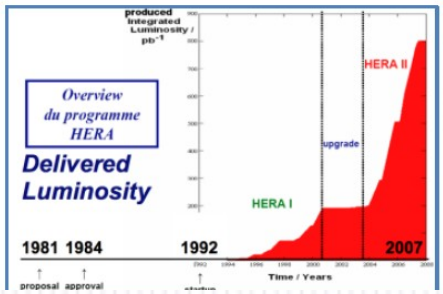
Pitkäaikaissäilytyksen lyhyt historia

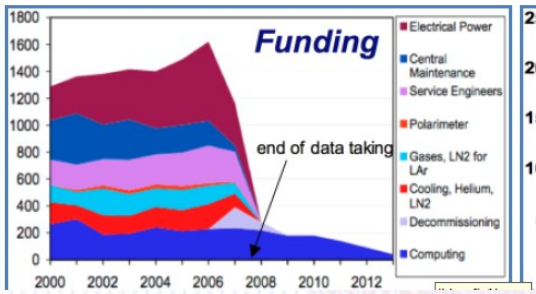
- Viimeisen 50 v. aikana hiukkastörmäyttimien törmäysenergia on kasvanut
  - monet aineistot ovat ainutkertaisia.
- Tutkijat pitävät aineiston pitkäaikaissäilytystä *hyvin* tai *ratkaisevan tärkeänä* (70%).

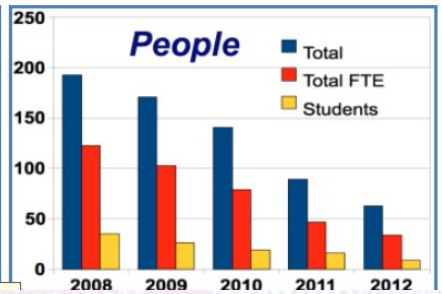
  (PARSE-Insight -tutkimus)
- Tähän asti pitkäaikaissäilytys on ollut suunnitelematonta ja yritykset usein yksittäisiä aloitteita.



- Ongelmana vähenevät resurssit (rahoitus ja tietotaito) tiedonkeruun loputtua.
  - esim. HERA-koe Saksan DESY-tutkimuslaitoksessa:







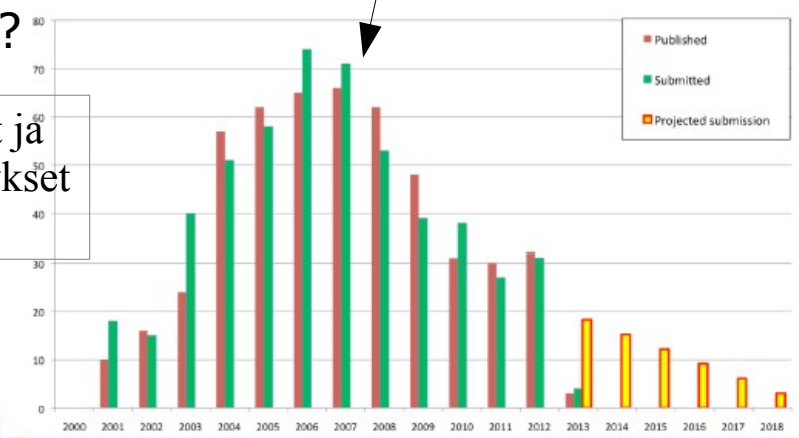
Faktat tänään

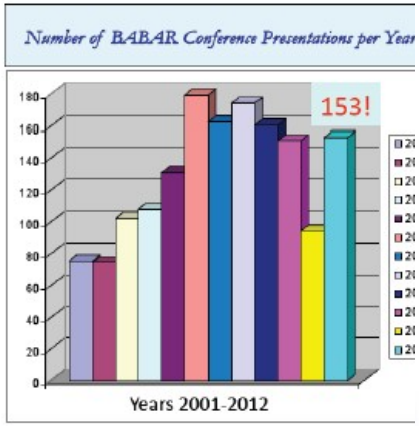
- Tiedonkeruun päättyessä kerätyn tietoaineiston volyymi on valtava – se « pienenee » muutamassa vuodessa → suurin **vaikeus ei ole datan säilytys**.

Törmäytin	Laboratorio	Päättyi v.	Volyymi
LEP	CERN (CH)	2000	~0.4 PB
HERA	DESY (D)	2007	1 PB
BaBar	SLAC (USA)	2008	2 PB
Belle	KEK (Japani)	2010	4 PB
Tevatron	FNAL (USA)	2011	18 TB

- Julkaisutoiminta motivoi aineiston pitämisen **käytettävänä** useita vuosia – mutta entä sen jälkeen?

BaBar: julkaisut ja konferenssiesitykset vuosittain





Pitkäaikaissäilytys nyt

- Pitkäaikaissäilytyksen haasteisiin on nyt herätty hiukkasfysiikan tutkimuksessa – Data Preservation in High Energy Physics (DPHEP) -työryhmä vuodesta 2008.
  - Työpajoja, joissa on ideoitu ja esitelty eri kokeiden ja laboratorioiden projekteja, jaettu tietoa ja kokemuksia.
  - Blueprint-raportti v. 2012 nykytilasta ja yhteisistä projekteista, suosituksia siitä, miten edetä.
  - DPHEP-työryhmä muuttuu kansainväliseksi kollaboraatioksi vuonna 2013, tavoitteena yhteinen visio ja konkreettisia toimia sen saavuttamiseksi.
  - Visio vuodelle 2020:

> The goal is that by 2020 **all existing efforts** on Long-Term Preservation in HEP will be **fully harmonized, both internally and also with other disciplines**.
> The target is that **all *archived* data** – such as that described in the DPHEP Blueprint, including LHC data – **remains fully usable** by well-defined *designated* communities.
> The best practices, tools and services should all be well run-in, fully documented and sustainable. Any exceptions to the above should be clearly described.
>
> Jamie Shiers – DPHEP project manager

Pitkäaikaissäilytys: uudet tuulet

- Hiukkasfysiikan yhteisö tukee pitkäaikaissäilytystä edistäviä toimia ja sen tärkeyttä aletaan korostaa:
  - CERN council: « Update of the European Strategy for Particle Physics »
    - ... pidettävä yllä ja kehitettävä tietoaineistojen säilytykseen tarvittavia rakenteita...
  - ICFA (International Committee for Future Accelerators): « ICFA Statement on Long-Term Data Preservation »
    - ... tukee DPHEP-työryhmän toimia ja kannustaa laboratorioita, tutkimuslaitoksia ja kokeita liittymään uuteen DPHEP-kollaboraatioon...
    - ... kiinnittää huomiota lähitulevaisuuden puuttuviin resursseihin ja sen mahdollisiin seuraamuksiin pidemmällä tulevaisuudessa...
    - ... panee merkille mahdollisuuden kansainväliseen yhteistyöhön eri tieteenalojen välillä ja kannustaa DPHEP-kollaboraatiota jatkamaan tarmokkaasti toimiaan näiden mahdollisuuksien toteuttamiseksi...
    - ... panee merkille pitkäaikaissäilytyksen arvon tutkimusaineistojen tieteellisen potentiaalin hyödyntämiseksi, mukaanlukien julkisten aineistojen opetus- ja outreach-käyttö...

Yhteistyöfoorumit

- Pitkäaikaissäilytyksen tärkeyttä korostavat myös esim.:
  - Euroopan komissio:

    « Open Infrastructure for Open Science – Horizon 2020 consultation report »

    « Research Data e-Infrastructures: Framework for Action in H2020 »

    - *e-Infrastructure fiche 03: storing, managing and preserving research data* – miten säilyttää korvaamattomat tutkimusaineistot ja taata niiden saatavuus myös tulevaisuudessa?
    - ... palvelukeskeinen tietoinfrastruktuuri ... eurooppalaisten IT-keskusten kehittämä ja kansallisten, kansainvälisten ja temaattisten IT-keskusten verkoston täydentämä Tier-0-infrastruktuuri tietoaineistojen säilyttämiseksi ja niiden käytön mahdollistavan tiedon kokoamiseksi...
  - Research Data Alliance
    - kansainvälinen liittymä tietoinfrastruktuurien kehittämiseksi, toimii työryhmissä, joista yhdeksi ehdotettu « Preservation e-Infrastructure »
- Hiukkasfysiikan yhteisö haluaa olla näissä aktiivisena toimijana
  - yhteistyö kanavoituu pääosin DPHEP-kollaboraation kautta.

CMS-koe: periaatepäätös

- CMS-koe on ensimmäisenä LHC-kokeista tehnyt periaatepäätöksen tietoaineiston **pitkäaikaissäilytyksestä** ja sen osittaisesta **avaamisesta julkiseen käyttöön**.
- Motivaatio sekä omalta tutkimusyhteisöltä että rahoittajatahoilta:
  - Haluamme pitää LHC-kokeiden tuottaman aineiston käytettävänä pitkään, koko kokeen elinajan ja myös tiedonkeruun päättymisen jälkeen.
    - LHC-kokeiden koko, maantieteellinen ulottuvuus ja pitkä kesto ovat sanelleet laskenta- ja toimintamallin, joka on takaa aineiston säilytyksen aikavälillä « joitakin vuosia » ja on yhteensopiva pitkäaikaissäilytyksen vaatimusten kanssa.
    - Aloitimme suunnittelun ajoissa: meillä on hyvät mahdollisuudet muokata toimintatapoja niin, että pitkäaikaissäilytys onnistuu.
  - Rahoittajatahot ovat kiinnostuneet tietoaineistojen pitkäaikaissäilytyksestä ja sen julkisesta saatavuudesta.
    - Esim. USAssa NSF-rahoitus vain esityksille, joissa nämä asiat on selvitetty.

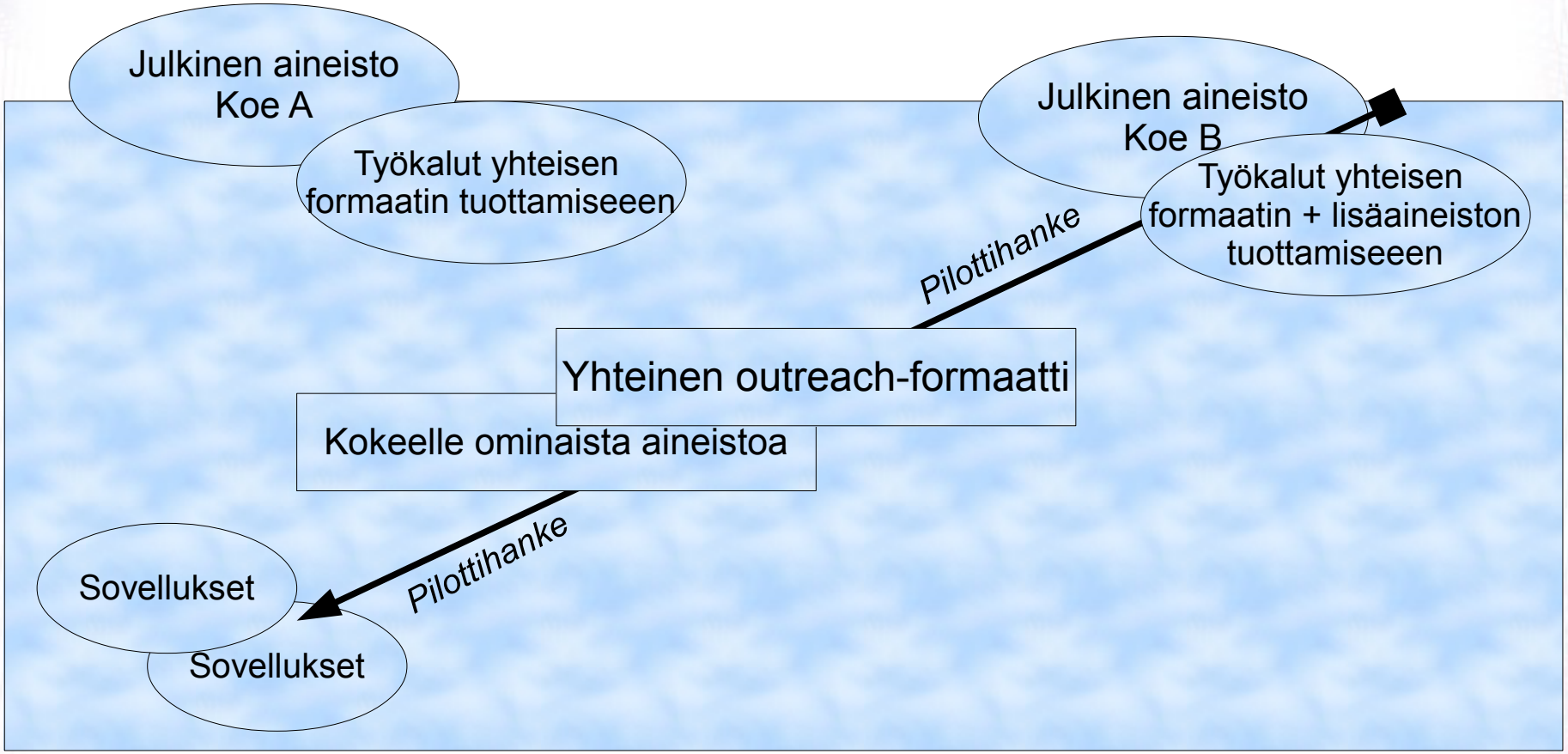
Miten julkinen aineisto tukee pitkäaikaissäilytystä?

- CMS julkistaa tutkimukseen sopivaa (rekonstruoitua) tietoaineistoa sekä tutkimukseen tarvittavat analyysiohjelmistot ja dokumentaation.
- Aineiston julkistaminen vastaa moniin pitkäaikaissäilytyksen käytännön haasteisiin:
  - Puuttuva konkreettinen *use case* – tutkijayhteisö paneutuu uusimpaan aineistoon, ja alkuvaiheen arvokas aineisto unohtuu.
    - Aineiston julkaisu viiveellä (esim. 3 v) asettaa määräajan sille, milloin jo aktiivisesta käytöstä poistuneen aineiston pitää olla saatavilla ja dokumentoituina.
  - Suuri osa aineiston tulkintaan tarvittavasta tietotaidosta on tällä hetkellä itsestään selvää, mutta unohtuu nopeasti.
    - Julkaistavaan aineistoon on toimitettava selkeät käyttöohjeet, mitkä voivat olla hyödyksi myös kokeeseen osallistuville tutkijoille.
  - Lyhyellä tähtäimellä « kotitekoinen » tietoinfrastruktuuri saattaa olla helpompi toteuttaa (kokeen sisäiset tietokannat, työkalut).
    - Aineiston julkinen saatavuus motivoi käyttämään tieteenalan tai koko tieteellisen yhteisön käyttämiä yhteisiä ratkaisuja.

TTA-hankkeen pilotti: CMS-kokeen julkisen aineiston opetuskäyttö

- CMS-kokeen periaatepäätös antaa mahdollisuuden aineiston hyödyntämiseen tutkimuksen lisäksi mm. opetuskäytössä. CMS-aineisto voi toimia pilottina TTA-hankkeessa avoimen tiedon hyödyntämiseen.
- Ensimmäisenä kohdepiirinä opetusesimerkit lukio-opetuksessa:
  - Esimerkkien sisältö ei välttämättä rajoitu jo sinällään kiinnostavaan aihepiiriinsä: aineistoa voi fysiikan periaatteiden opettamisen lisäksi käyttää valaisemaan tilastollista data-analyysia, numeerisia menetelmiä, datan hallintaa ja data visualisointia.
- Ensimmäinen, vuonna 2010 kerätty tietoaineistoerä suunnitellaan julkaistavaksi LHC-kiihdyttimen käyttökatkon aikana 2013-2014.
- Pitkän aikavälin tavoitteena on mahdollistaa hiukkasfysiikan perustutkimuksen temaattinen keskittymä, jossa eri hiukkasfysiikan kokeet tuottavat tietoaineistoa, jolle voidaan käyttää yhteisiä rajapintapalveluja tietoaineistojen hyödyntämiseksi.
- Tämän projektin myötä Suomella olisi erinomainen mahdollisuus profiloitua avoimen tiedon edelläkävijänä.

Pilottihanke



- Tavoitteena ei ole niinkään pelkkä opetussovellus, vaan avointa tietoa hyödyntävä palveluympäristö, joka on avoin, helppokäyttöinen, laajennettavissa ja pidemmälle kehitettävissä.

Tiivistelmä

- Pitkäaikaissäilytyksen tärkeys ja sen riittävän aikaisen suunnittelun tarve on nyt ymmärretty hiukkasfysiikan yhteisössä.
  - Suurin haaste ei ole data vaan sen käytettävyys.
- Hiukkasfysiikan tutkimusaineistossa on monia piirteitä, jotka – jos niitä pystytään hyödyntämään – tekevät pitkäaikaissäilytyksen mahdolliseksi.
- Emme selviä tästä yksin – kansainvälinen yhteistyö muiden tieteenalojen kanssa on välttämätöntä.
  - DPHEP toimii hiukkasfysiikan yhteisön äänitorvena.
- CMS-koe on tehnyt periaatepäätöksen tutkimusaineiston pitkäaikaissäilytyksestä ja ja sen avaamisesta julkiseen käyttöön.
  - Olemme pitkän tien alussa, mutta jo nyt avautuneet yhteistyö-mahdollisuudet osoittavat, että valitsimme oikean reitin.