

CLARIN

Common Language Resource and Technologies Infrastructure

Kimmo Koskenniemi

*Member of the Executive Board of CLARIN
Leader of the Work Package 7 (for IPR issues)*

Eri CLARINit

- **EU-CLARIN**, joka määrittelee puitteet
- Kansalliset **CLARIN-verkostot**, jotka koordinoivat sen maan **CLARIN-keskuksia**
- **FIN-CLARIN** on Suomen CLARIN-verkosto ja CSC sen ainoa CLARIN-keskus
- CLARINIin liittyvät yksittäiset **hankkeet**, jotka tuottavat sisältöä (aineistoja, menetelmiä, ohjelmia)

EU-CLARIN

- **Puitteet ja säännöt** yhteensopivalle verkostolle
- **Standardit** aineistojen muodoille, työkaluohjelmille, metadatoille ym.
- **Sopimusmallit** aineistojen saamiseksi, käyttäjien lisenssit ym.
- Yhteisesti sovittu **autentikointi** ja **auktorisointi**
- CLARIN-keskusten väliset **sopimukset**, ERIC tms.

Kansalliset CLARINit

- Kussakin maassa tulee järjestää kansallisilla resursseilla oma **CLARIN-yhteensopiva** infra.
- Saksassa D-SPIN, Suomessa FIN-CLARIN jne.
- Organisoivat maansa eri tahot toimimaan yhdessä.
- Koostuvat CLARIN-yhteensopivista keskuksista (1 tai useammasta) sekä aineistoja tuottavista ja käyttävistä tahoista
- Kansalliset keskukset noudattavat EU-CLARINin standardeja, tietojen siirtoa, autentikointia, sopimuksia ja lisenssejä.

Kauppa/kirjasto/CLARIN

- CLARIN ei ole kauppa, jollaisesta **kuka tahansa** saa ostaa tavaran.
- CLARIN ei ole (vain) kirjasto, josta **kaikki** asiakkaat saavat lainata ja lukea **kaikkia** teoksia
- CLARIN tarjoaa tutkijoille (tarvittaessa säänneltyyn) pääsyn aineistoihin ja niihin perustuviin palveluihin (osa-aineistojen valikointi, tilastoja, trendejä, hakuja, esimerkkejä)

FIN-CLARIN

- HY, TaY, JY, JoY, OY, Kotus ja CSC
- Aiesopimus: pyrkimys siihen, että osapuolet saattaisivat aineistojaan CLARINin mukaisesti mahdollisimman laajalti käytettäviksi
- Aineistot ja palvelut talletetaan CSC:hen, joka liittyy EU-CLARINIin

Aineistot, kokoaminen ja säilytys

- **Tekstejä, sanakirjoja** (XML-muotoisia, osa kieliopillisilla koodeilla täydennettyjä), 10-100 TB.
- Digitoitua **puhetta**, keskustelua videoituna, PB.
- Kerätty ja digitoitu aiemmin, **saatettava** CLARINin **vakiomuotoon**. Vajavaiset sopimukset saattavat estää laajemman käytön.
- Uusissa hankkeissa tehtävä **suoraan** CLARINin muotoon ja sen **sopimusmalleilla**.
- Säilytetään CLARIN-keskuksissa, esim. CSC:ssä.

Sisältö tuotetaan hankkeina

- **Olemassa olevia** aineistoja pyritään kansallisten verkostojen toimesta **siirtämään** CLARINiin, kuten CSC:n Kielipankin aineistoja (sanomalehtiä, kirjoja), murrekokoelmia, nauhoitearkistoja, sanakirjoja
- **Uusien** aineistojen kerääminen, toimittaminen, konversiot ja annotoinnit **hankkeina**, joiden tulokset suoraan CLARINiin (esim. SA)
- Kieliteknologisten **menetelmien** ja ohjelmien kehittäminen (ehkä 100 kielelle) (EU, SA)

Jakeluratkaisut sekä oikeuksien ja maksujen hallinta

- Jakelu verkon välityksellä.
- Toistaiseksi **ei** ajateltu **maksuja** käyttäjille suoraan, vaan kulut jaettaisiin ja maksettaisiin keskitetysti.
- Osa aineistoista **avoimia**, kaikki voivat käyttää.
- Osalle voidaan **automaattisesti** myöntää **lupa** digitaalisesti allekirjoitettua sitoumusta vastaan esim. yliopistojen henkilökunnalle ja opiskelijoille.
- Osalle aineistoista täytyy esim. eettisten syiden takia **anoa ja perustella** lupaa yksilöllisesti.

Autentikointi ja auktorisointi

- Aineistolle tulee olla taho, joka voi antaa **luvan**.
- Lupa voi olla **ryhmille** (esim. yliopistojen henkilökunnalle) tai **yksilöllisesti** haettu, perusteltu ja myönnetty.
- Käyttäjien tulee olla **luotettavasti** tunnistettuja (HAKA). Pelkkä tekniikka ei riitä (Shibboleth, SAML 2.0). Yhteiset sovitut käytännöt välttämättömiä.
- Autentikointi on EU-CLARINille iso haaste.

Ongelmat ja haasteet

- **Vanhimpia** tekstejä voidaan tallettaa ja käyttää ongelmitta.
- Uudemmat ovat tekijänoikeuslain takia luvanvaraisia. Yliopistot eivät voi niitä digitoida eivätkä käyttää tutkimuksessa ilman nimenomaista lupaa (mikä uusimmille on toki jokaiselle erikseen hankittavissa).
- **1900-luvun** alkupuolen aineistoille hyvin vaikeata saada lupia. Kustantajaa ja tekijän perikuntaa ei ehkä helppo tavoittaa.
- Jotkut aineistot sisältävät tunnistettavien henkilöiden puhetta, minkä vuoksi tutkimus vain **perustellusta syystä** ja erityisin sitoumuksin voi olla sallittua. Hoidettavissa CLARINissa.
- Jotkut aineistot on **luovutettu ehdoin**, että vain alan tutkijat saavat käyttää. Hoidettavissa CLARINissa.

Ongelmat ja haasteet (jatkoa)

- Tekijänoikeuslaki ei Suomessa salli sellaistakaan tutkimusta, joka ei vaaranna tekijän ja kustantajan olennaisia intressejä.
- EU:n pitäisi **muuttaa** lainsäädäntöä, pakollinen ns. poikkeus tutkimusta varten. (Aloite tekeillä)
- SA:n ja muiden **rahoittajien** pitäisi edellyttää, että julkisella rahoituksella syntyvät digitaaliset kieliaineistot tehdään CLARIN-yhteensopiviksi mahdollisimman vapain ehdoin ja talletetaan CLARINIin.

Muita CLARINin haasteita

- Yhteistoiminta ja työn koordinointi **digitaalisten kirjastojen hankkeiden** kanssa. (Kirjallisuus on kieltä, CLARIN tarjoaa lisää käyttömahdollisuuksia.)
- Yhteistyön kehittäminen **kustantajien** ja kirjailijoiden, kääntäjien ja toimittajien järjestöjen kanssa. (CLARINin puitteissa voisi digitoida ja tarjota kustantajalle hyödyntämismahdollisuutta).
- **Googlen** digitointihankkeen merkitys?
- **Kaupallisten** periaatteiden mukaantulo CLARINIin?
Vrt. venäläinen Integrum-palvelu.