CLARINO

Common Language Resources and Technology Infrastructure – Norway

Koenraad De Smedt

Koordinator

Oslo, 4. juni 2012

Outline

Outline

EC survey on scientific information in the digital age

- Er en EU-strategi ønskelig for tilgang til og bevaring av forskningsdata?
  83% JA
- Er det uproblematisk å få tilgang til forskningsdata?
  87% NEI
- Er det tilstrekkelig støtte for forskningsinfrastruktur?
  NEI
- Får forskere tilstrekkelig uttelling for å gjøre data tilgengelig?
  80% NEI
- Bør forskningsdata være offentlig og gratis tilgjengelig på nettet?
  90% JA

2011, http://link.uib.no/?3JcAK

Swedish National Data Service

Strategy for Sharing and Mediating Data. Practices of Open Access to and Reuse of Research Data – The State of the Art in Sweden 2009

- Viktige grunner for ikke å gjenbruke digitale data er usikkerhet mht. kvalitet (62%), etiske aspekter (57%), tekniske problemer (53%) og juridiske aspekter (49%).
- 53% av forskerne mente det var særdeles viktig å dele data, men bare 26% hadde planer om å dele sine *egne* data.

Verdenen før CLARINO

- En forsker googler en språkressurs, f.eks. et tekstkorpus, en digitalisert grunnbok, osv.
- Hun finner en peker som kanskje består, kanskje er død
- Ressursen kan ikke lastes ned og er ikke dokumentert
- Ressursen er søkbar men man må be om å få brukernavn og passord
- Søkeresultat kommer på en html-side
- Det finnes ingen opplagte måter å analysere resultatene videre
- Forskeren må skriver egne script for å jobbe med resultatene
- Resultatet blir oppsummert i en publikasjon men blir ikke arkivert som data
- Hverken det opprinnelige korpuset, de ekstaherte eller de analyserte data er siterbare

Hva er problemet?

Digitale forskningsdata

- er ikke søkbare i noen samlet katalog
- er ikke tilstrekkelig synlige
- har uklare bruksbetingelser
- har en høy terskel for å tas i bruk
- er ikke alltid kompatible med verktøy/plattformer
- er ofte laget for ett formål og er ikke gjenbrukbare
- blir ikke behandlet som publikasjoner
- har ingen permanent adresse og er ikke direkte siterbare
- er ikke alltid koblet til forskerne som laget dem
- kan forsvinne eller bli ubrukelige

CLARINs mål

"CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities."

`http://www.clarin.eu`

Konsolidering, samling, tilgjengeliggjøring, analyse og bedre utnyttelse av språkressurser som ellers kan gå tapt.

CLARINs visjon

- En forsker logger inn ved sin egen institusjon (f.eks. via Feide) og kommer inn på CLARIN-portalen
- Hun ser i en katalog, søker i metadata eller søker i innhold
- Hun lager sin egen virtuelle samling av ressurser fra flere databaser i inn- og utland
- Hun signerer lisensavtaler digitalt der nødvendig
- Hun spesifiserer en arbeidsflyt for prosessering av den virtuelle samlingen ved hjelp av analyseverktøy gjennom webtjenester og fjernbruk av tungregning
- Resultatene blir lagret i et eget arbeidsområde
- Data og metadata lastes opp i en database og den virtuelle samlingen lagres for fremtidig bruk ved hjelp av PIDer

Språkets mangfold i digital form

- Primære språkdata, rå eller annotert: korpus, arkiver, osv.
- Sekundære språkdata og verktøy: ordbøker, nøkkelord, grammatikker, frekvenslister, osv.
- Språkverktøy: maskinoversettelse, automatisk sammendrag, indeksering, osv.
- Eksperimentelle data: øyebevegelser, reaksjonstider, spektogrammer, EEG, fMRI, osv.

Hva er digital forskningsinfrastruktur?

- Et permanent og lett tilgjengelig tilbud av digitale informasjonskilder og kunnskapstjenester
- En fusjon og videreutvikling av digitale biblioteker og institusjonelle arkiver
- eVitenskap er en ny dimensjon i vitenskapelig praksis: kunnskapsutvikling gjennom kobling av store datasett og avansert modellering

Outline

CLARIN: LAG (LAYERS)

1. Koordinasjon og styring (ERIC)
2. Infrastruktur (langsiktig nasjonalt ansvar) — CLARINO legger grunnlaget
3. Innhold inkl. digitalisering og tilrettelegging (prosjekter støttet av Forskningsrådet, institusjonene, Nasjonalbiblioteket mm.)

Nasjonalt nettverk av noder

Type A. Nasjonalbiblioteket (+Uninett)

Type B. Tekstlaboratoriet (UiO);
EDD (UiO);
LAP/IFI (UiO);
Bergen: LaMoRe+UBB(UiB)+NHH+UniComputing

Type C. NTNU;
UIT

Ansvarsfordeling for tjenester

- *Infrastrukturtjenester* (Nasjonalbiblioteket, Uninett). Nasjonal katalog over språkressurser, langtidslagring, HPC-tilgang, autentisering og autorisering, PID-tjeneste.
- *Språkdatatjenester* (UiO, UiB, Uni Research, NHH). Korpusanalyse, terminologiportal, elektronisk utgaveplattform.
- *Språkteknologitjenester* (UiO, UiB, Uni Research). Språkanalyseportal (LAP), verktøy og prossesseringskjeder.
- *Levering av data og metadata* (alle noder).

Andre infrastrukturer

INESS, MENOTEC, ...

Bidrar med tilgjengeliggjøring av sine data av egne midler

Forhold til bibliotek

"Language infrastructures represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architecture"

Institusjonelle strategier for:

- åpne institusjonelle arkiver for deponering og arkivering av forskingsdata
- kobling av publikasjoner til data
  data → publikasjon → data
- uttelling for publisering av data
- håndtering av IPR, lisensiering av data til gjenbruk
- PID for sitering av data

Bibliotek som eHumanities-miljø: Perseus

Plato, *Republic*

Search

("Agamemnon", "Hom. Od. 9.1", "denarius")
All Search Options [view abbreviations]

Home | Collections/Texts | Research | Grants | Open Source | About | Help

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position. Hide browse bar

book:
section:

This text is part of:
Greek and Roman Materials
Greek Prose
Greek Texts
Plato
Plato, Republic

View text chunked by:
book : page
book : section
page
section

Table of Contents:
book 1
section 327a
section 327b
section 327c
section 328a
section 328b

Plat. Rep. 1.327a

Click on a word to bring up parses, dictionary entries, and frequency statistics

[327α]

Σωκράτης
κατέβην χθὲς εἰς Πειραιᾶ μετὰ Γλαύκωνος τοῦ Ἀρίστωνος προσευξόμενός τε τῇ θεῷ καὶ ἅμα τὴν ἑορτὴν βουλόμενος θεάσασθαι τίνα τρόπον ποιήσουσιν ἅτε νῦν πρῶτον ἄγοντες. καλὴ μὲν οὖν μοι καὶ ἡ τῶν ἐπιχωρίων πομπὴ ἔδοξεν εἶναι, οὐ μέντοι ἧττον ἐφαίνετο πρέπειν ἢν οἱ Θρᾷκες ἔπεμπον.

Plato. Platonis Opera, ed. John Burnet. Oxford University Press. 1903.

The Annenberg CPB/Project provided support for entering this text.

XML

An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.

Notes (James Adam) focus hide

327A - 328B Socrates describes how he visited the Piraeus in company with Glauco, and was induced by Polemarchus and others to defer his return to Athens.

κατέβην κτλ. Dionys. Hal. de comp. verb. p. 208 (Reiske) ὁ δὲ Πλάτων, τοὺς ἑαυτοῦ διαλόγους κτενίζων καὶ βοστρυχίζων, καὶ πάντα τρόπον ἀναπλέκων, οὐ διέλιπεν ὀγδοήκοντα γεγονὼς ἔτη. πᾶσι γὰρ δή που τοῖς φιλολόγοις γνώριμα τὰ περὶ τῆς φιλοπονίας τἀνδρὸς ἱστορούμενα, τά τ' ἄλλα, καὶ δὴ καὶ τὰ περὶ τὴν δέλτον ἣν τελευτήσαντος αὐτοῦ λέγουσιν εὑρεθῆναι ποικίλως μετακειμένην τὴν ἀρχὴν τῆς πολιτείας ἔχουσαν τήνδε "κατέβην χθὲς εἰς Πειραιᾶ μετὰ Γλαύκωνος τοῦ Ἀρίστωνος." See also Quint. VIII 6. 64, and Diog. Laert. III 37. The latter gives as his authorities Euphorion and Panaetius. As Cicero was tolerably familiar with the writings of Panaetius, it is possible that he too has the same story in view in de Sen. V 13, where he says of Plato ""scribens est mortuus."" The anecdote may well be true, but does not of course justify any inference as to the date of composition of the Republic. See Introd. § 4.

τῇ θεῷ. What goddess? Bendis or Athena? The festival is the Bendideia (354 A) and it is perhaps safest to acquiesce in the usual view that Bendis is here meant. "Alii Minervam intelligunt, quae vulgo ἡ θεὸς appellabatur; neque mihi videtur Socrates in ista Panathenaeorum propinquitate de Minerva veneranda cogitare non potuisse: sed quod simpliciter τὴν ἑορτὴν dicit, numina diversa statuere non sinit" (Schneider). We hear of a temple of Bendis in the Piraeus in 403 B.C. (τὴν ὁδὸν ἡ φέρει πρός τε τὸ ἱερὸν τῆς Μουνυχίας Ἀρτεμίδος καὶ τὸ Βενδίδειον **Xen. Hell. II 4. 11**). See also Introd. § 3 and App. I.

νῦν πρῶτον. Perhaps 410 B.C. Introd. § 3.

Forhold til Språkbanken

Språkbanken:

- *"det naturlige samlingspunktet for lagring og distribusjon av offentlige og private digitale språkressurser"*
- *"å være infrastruktur i språkteknologisk forskning, utvikling og produkt- og tjenestetilpasning for norsk språk"*

CLARIN

- Europeisk satsing som inkluderer alle språk
- Alle slags språkrelaterte forskningsbehov innen humanistiske fag (f.eks. Wittgensteinarkivet; Bosnisk-korpuset)
- Språkteknologi er til dels et instrument heller enn et mål

Forhold til META-NORD / META-SHARE

META-NORD = nordisk-baltisk prosjekt, kortvarig (til 31. jan. 2013)

META-SHARE = katalog for metadata

- Rettet mot språkteknologisk utvikling og anvendelser
- Arbeid med metadata og katalogisering

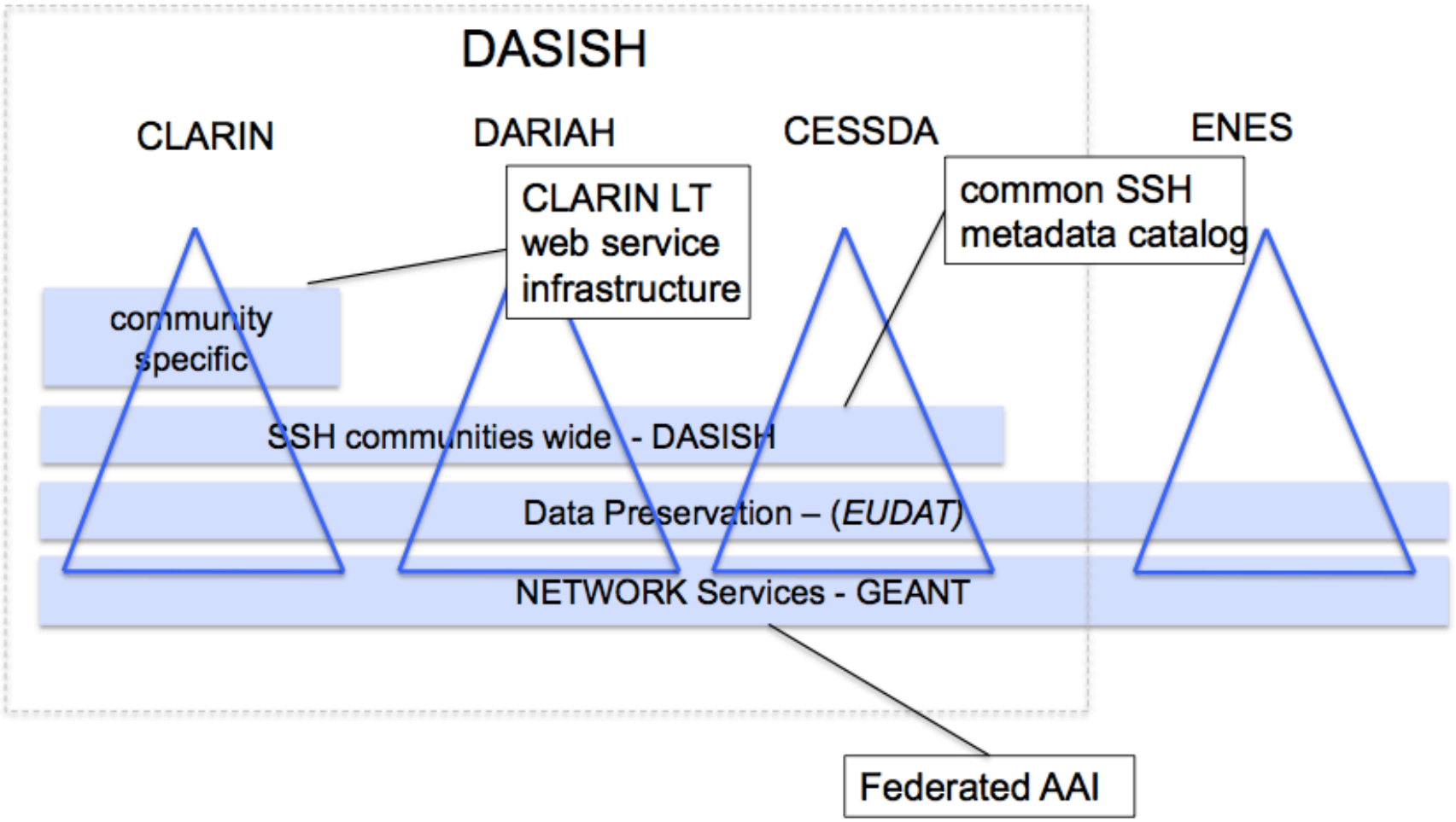
DASISH

Europeisk INFRA 2011 prosjekt, samarbeid mellom:

- CESSDA (Council of European Social Science Data Archives)
- CLARIN (Common Language Resources and Technologies Infrastructure)
- DARIAH (Digital Research Infrastructure for the Arts and Humanities)
- ESS (European Social Survey)
- SHARE (Survey of Health, Ageing and Retirement in Europe)

Felles strategier for forskningsdata i samfunnsvitenskap og humanistiske fag

DASISH / EUDAT / GEANT



Alternativ driftsmodell: ANDS

Forhold til Notur/NorStore

- Lagring og tungregning i CLARINO (også for relatert infrastruktur f.eks. INESS)
- Trenger mer permanent strategi for eInfrastruktur
- Feide og Kalmar2 har vist seg å være nyttig for brukerautentisering i CLARIN og INESS: en effektiv matrise av id-leverandører og tjenesteleverandører

Lokale tiltak

Fakultetene, instituttene og bibliotekene bør ta sitt ansvar ifm. en samordnet strategi for humanistiske forskningsdata

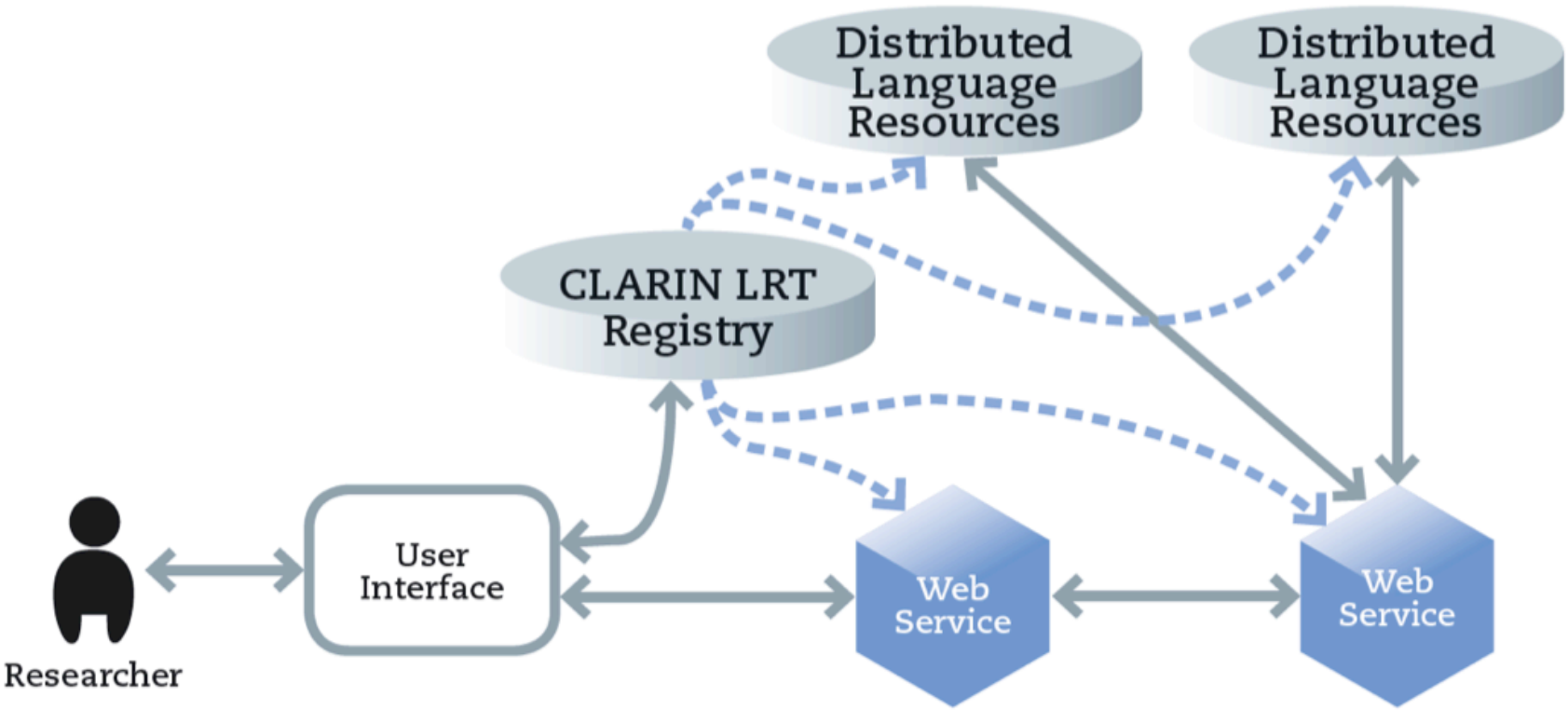
- uttelling for å gjøre forskningsdata tilgjengelig
- offentliggjøring og gjenbruk må være en del av hvert prosjekt (også masteroppgaver)
- rutiner for å lagre *nedlastbare data, metadata og bruksbetingelser* i CLARINO
- rutiner for å håndtere søknader ifm. § 16 i åndsverkloven
- faste stillinger for forskning og utvikling av infrastruktur

Outline

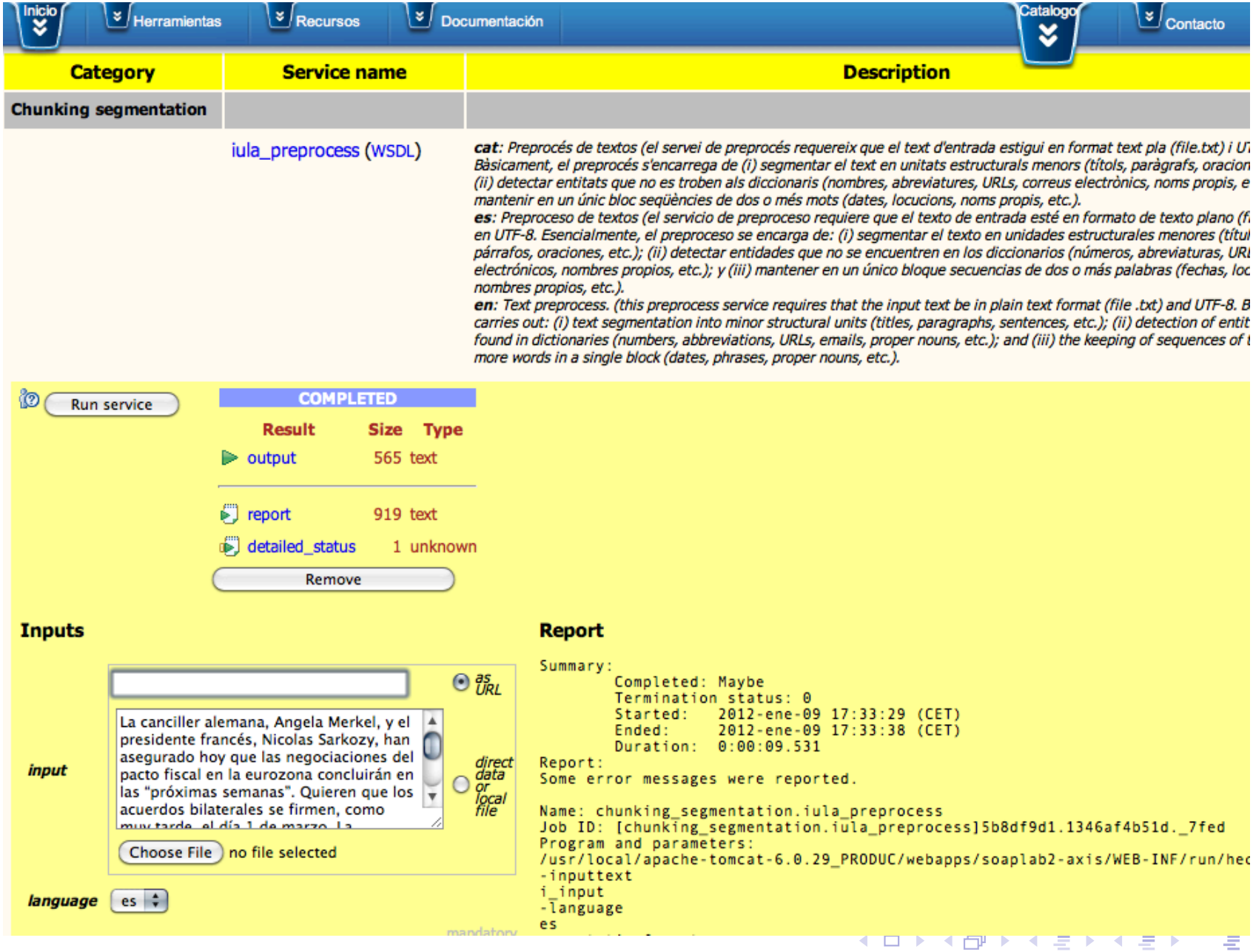
Prosjektplan (12. april 2012 – 28. mars 2019)

- **WP1** Centres Setup
- **WP2** National Registry and Long-Term Archiving
- **WP3** Trusted AAI
- **WP4** Electronic Editions Platform with IDP
- **WP5** Glossa Integration
- **WP6** Corpuscle Integration
- **WP7** Terminology Integration
- **WP8** LAP
- **WP9** Tool Adaptation
- **WP10** Data and Metadata Adaptation
- **WP11** Management and Dissemination
- **WP12** Use Cases, Evaluation and Delivery

Webtjenester



CLARIN-ES


Inicio
Herramientas
Recursos
Documentación
Catalogo
Contacto
Category
Service name
Description
Chunking segmentation
iula_preprocess (WSDL)
cat: Preprocés de textos (el servei de preprocés requereix que el text d'entrada estigui en format text pla (file.txt) i UT
Bàsicament, el preprocés s'encarrega de (i) segmentar el text en unitats estructurals menors (títols, paràgrafs, oracion
(ii) detectar entitats que no es troben als diccionaris (nombres, abreviatures, URLs, correus electrònics, noms propis, e
mantenir en un únic bloc seqüències de dos o més mots (dates, locucions, noms propis, etc.).
es: Preproceso de textos (el servicio de preproceso requiere que el texto de entrada esté en formato de texto plano (f
en UTF-8. Esencialmente, el preproceso se encarga de: (i) segmentar el texto en unidades estructurales menores (títul
párrafos, oraciones, etc.); (ii) detectar entidades que no se encuentren en los diccionarios (números, abreviaturas, URL
electrónicos, nombres propios, etc.); y (iii) mantener en un único bloque secuencias de dos o más palabras (fechas, loc
nombres propios, etc.).
en: Text preprocess. (this preprocess service requires that the input text be in plain text format (file .txt) and UTF-8. B
carries out: (i) text segmentation into minor structural units (titles, paragraphs, sentences, etc.); (ii) detection of entit
found in dictionaries (numbers, abbreviations, URLs, emails, proper nouns, etc.); and (iii) the keeping of sequences of t
more words in a single block (dates, phrases, proper nouns, etc.).
Run service
COMPLETED
Result Size Type
output 565 text
report 919 text
detailed_status 1 unknown
Remove
Inputs
input
as URL
La canciller alemana, Angela Merkel, y el presidente francés, Nicolas Sarkozy, han asegurado hoy que las negociaciones del pacto fiscal en la eurozona concluirán en las "próximas semanas". Quieren que los acuerdos bilaterales se firmen, como
direct data or local file
Choose File
no file selected
language
es
Report
Summary:
Completed: Maybe
Termination status: 0
Started: 2012-ene-09 17:33:29 (CET)
Ended: 2012-ene-09 17:33:38 (CET)
Duration: 0:00:09.531
Report:
Some error messages were reported.
Name: chunking_segmentation.iula_preprocess
Job ID: [chunking_segmentation.iula_preprocess]5b8df9d1.1346af4b51d._7fed
Program and parameters:
/usr/local/apache-tomcat-6.0.29_PRODUC/webapps/soaplab2-axis/WEB-INF/run/hec
-inputtext
i_input
-language
es

Opplæring

Kurs i regi av DASISH, EUDAT, META-NET osv.

EUDAT 1st training days, Juni 2012

Outline

WP1 Centres setup

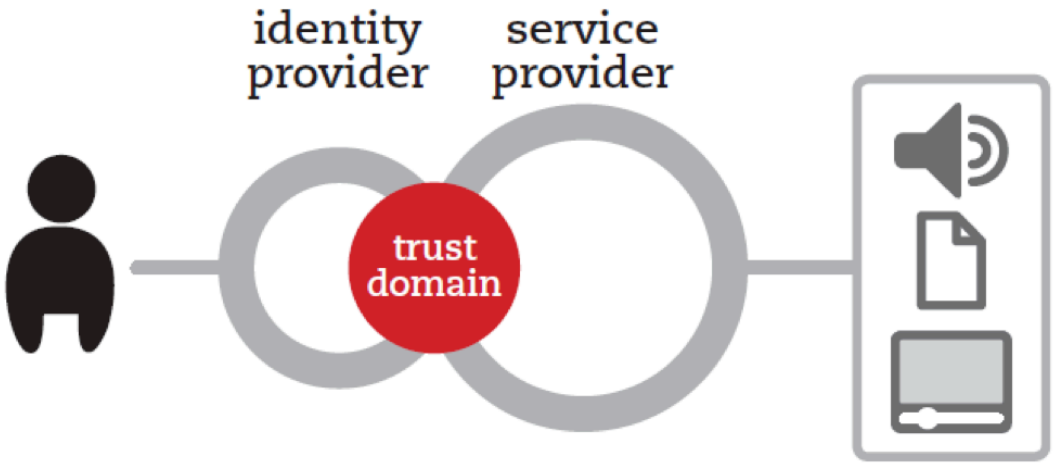
- gjør skjema og koding eksplisitt, bruk ISOcat `http://www.isocat.org`
- bruk versjonering og PID (EPIC, DataCite)
- deponer data i repository
- sett opp høsting av metadata

WP1: Milepæler

- Centres setup for metadata harvesting: 1. juli 2013
- Centres setup for AAI: 1. juli 2013

Outline

Trusted Authentication and Authorization Infrastructure (AAI)



AAI

- Shibboleth, Moonshot, Eduroam (Radius), OpenID, XACML, Oauth2
- Feide, cross-federated authentication: EduGain, Kalmar2

Milepæl 3.1 AAI conneeted to CLARIN AAI: 1. juli 2013

Outline

Management

- Prosjektkoordinator Koenraad De Smedt støttet av prosjektmanager
- Arbeidspakkekoordinatorer
- Styringsgruppe
- Råd
- Årlige konsortiemøter
- Konsortieavtale

Kommunikasjon og formidling

- Språk: norsk internt, engelsk eksternt
- Foreløpig nettsted: `http://clarin.b.uib.no`
- Trenger internt nettsted (Redmine, BaseCamp?)
- Publikasjoner og presentasjoner
- Deltagelse i CLARIN-aktiviteter
- Formidlingsevenement i Norge

ØKONOMI

- Totalt budget (år 1–5): 41,348 M, Støtte fra NFR: 25 M (60%)
- Personalskostnader (år 1–5): 37,8 M (91,5%)
- Utstyr (år 1–5): 0,5 M
- Andre kostnader (reiser mm, år 1–5): 1,5 M
- ERIC-medlemskap (dekket av NFR i år 1–5)
- Egeninnsats personal (år 1–10): 30,8 M (123% av NFRs støtte)
- Egeninnsats fra NB og Uninett

Regnskap og rapportering

- Etterskuddsvis betaling; send fakturaer til Maria Makolska, UiB
- Oppgaver må være utført og tidsplanen må overholdes (men kan skje raskere)
- Klarér reiser og anskaffelser med koordinator
- Send reiseregninger til Maria Makolska, UiB
- Utvidet rapportering til NFR: inkluder egeninnsats; rapporter om bruk av infrastrukturen
- Trenger administrativ kontaktperson ved hver enhet
- Årlig rapportering og tilleggsrapportering 1. halvår

Outline

Use cases

- Hva er eksempler på forskningsprosjekter som kan bruke infrastrukturen?
- Hva slags spørsmål vil humanistiske forskere stille?
- Hvordan vil og kan forskere lettest mulig bruke infrastrukturen?

Milepæl 12.1 Use cases: 1. oktober 2012

Evaluation and delivery

- Evaluering gjennom autentisk bruk blant målgruppen
- Presentasjon gjennom evenement og media

Milepæl 12.2 Overall evaluation and delivery: 1. april 2017