CLARIN
AAI Requirements

Daan Broeder & Dieter van Uytvanck

Max-Planck Institute for
Psycholinguistics

TF-EMC2 Meeting Loughborough May 5, 2009

Contents

- CLARIN Project
- AAI scenarios in the CLARIN domain
- Long term objectives & eduGAIN role
- Short term objectives

What is CLARIN

The CLARIN project is a large-scale pan-European collaborative effort to coordinate and make language resources and technology available and readily usable for Language & SSH (Social Sciences & Humanities) researchers.

- Resources: Lexica, text corpora, multi-media/multi-modal recordings, …
- Technology: applications & (web-)services as parsers, tokenizers, speech recognizers & segementators, …
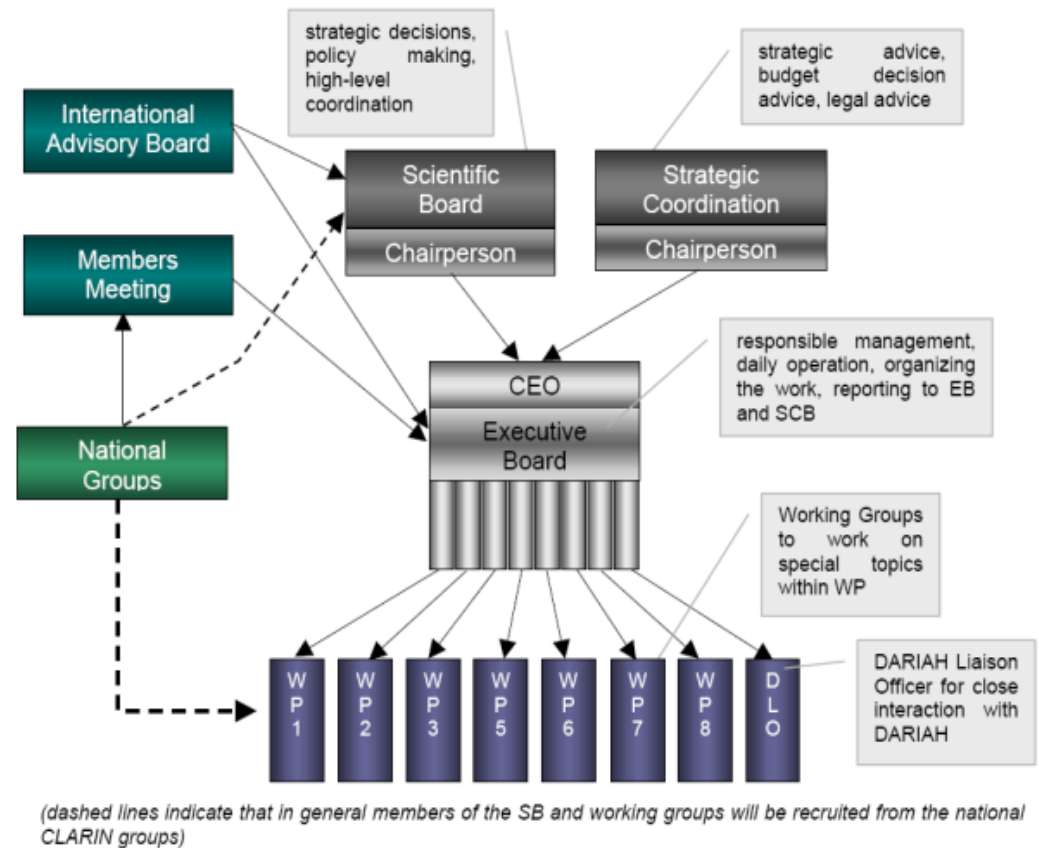
CLARIN overview

- CLARIN is an EU Infrastructure project with 4.2 ME funding for a 3 year preparatory phase
- Additional funding from national governments (at this moment at least 7 ME + 9 ME)
- The CLARIN consortium has now 32 partners from 26 EU countries
- The CLARIN community has 151 member organizations in 32 countries (mostly from NLP organizations)
- CLARIN is based on earlier initiatives with many participants: LangWeb, EARL, TELRI, LIRICS and more recent DAM-LR

CLARIN Organization

WP1	Management & Coordination, Steven Krauwer, OTS U. Utrecht NL
WP2	**Technical Infrastructure, Peter Wittenburg, MPI Nijmegen NL**
WP3	Humanities Overview, Tamas Varadi, Hung Ac. Sc., Hun
DLO	DARIAH Liaison, Martin Wynne, Oxford Univ. UK,
WP5	Language Resources & Technology Erhard Hinrichs, U. Tuebingen, Ger
WP6	Dissemination, Dan Crista, U. Iasi Rom
WP7	**IPR and Business Models, Kimmo Koskiennemie, Univ. Helsinki Fin**
WP8	Organizational Agreements, Bente Maegaard, Univ. Copenhagen, Dk

(dashed lines indicate that in general members of the SB and working groups will be recruited from the national CLARIN groups)

TF-EMC2 Meeting, May 5 2009

Time plan

- ## 2008 - 2010 Preparatory Phase
  - Limited set of federated centers (10+)
  - Showcases, demonstrators
  - WP2: Make cost estimates for infra. maint.
  - WP8: Investigate embedding in national funding schemes for construction phase & maintenance
- ## 2010 - 2020 Construction Phase
  - No important European funding
  - Depend on national project commitments
- ## 2020? - …   Maintenance Phase

CLARIN as a SP Organization

- CLARIN sees itself developing into a SP organization offering:
  - Language Resources in its broadest sense
    - Corpora: (annotated) texts, multi-media/multi-modal recordings, etc.
  - Tools & web-services
    - Multi-media & text annotation tools
    - Parsers, tokenizers, entity recognizers, ASR components
- Infrastructure services:
  - registries to find resources, tools & services
  - Where needed typical infrastructural services like: PID service

CLARIN Executive Board on a "typical" quest

CLARIN "Holy Grail"
Use Case II

- A researcher authenticates at his/her own organization and creates a "virtual" collection of resources from different repositories.

- He does this on the basis of browsing a catalogue, searching through metadata, or searching in resource content.

- He is then able to use a workflow specification tool and process this virtual collection with possibly a mix of home grown and remote service components.

- Resulting data can be added to the origin repositories with proper access rights and the "virtual" collection specification can be stored for future reference.

- For our domain this is very ambitious and challenging, but even a partial realization is worthwhile!

CLARIN AAI Scenarios

- SSO with a web-browser

- SSO with a locally installed application

- Delegation with a web application

- Delegation in a workflow of web-service chains

- Access to GRID resources

SSO with a web client

- User visualizes resources from many different resource providers

- Not of much use for complicated processing of distributed resources

- Although some processing may be implemented in applets

SSO with local application

- User runs a locally installed application that gathers resources from different resource providers

- Considered solutions:
  - Use a X.509 certificate (apache + mod_ssh)
  - Possibly obtained via SLCS
  - (SLCS option being tested within Dutch BiG Grid project)

Delegation with web-application

- User authenticates to web-app that then needs to access resources from different providers on behalf of the user.

- Considered solutions:
    - Accept the web-application's claim
    - Use a certificate
    - possibly obtained via SLCS

Searching through annotations

Parsers "normalize" the structural format
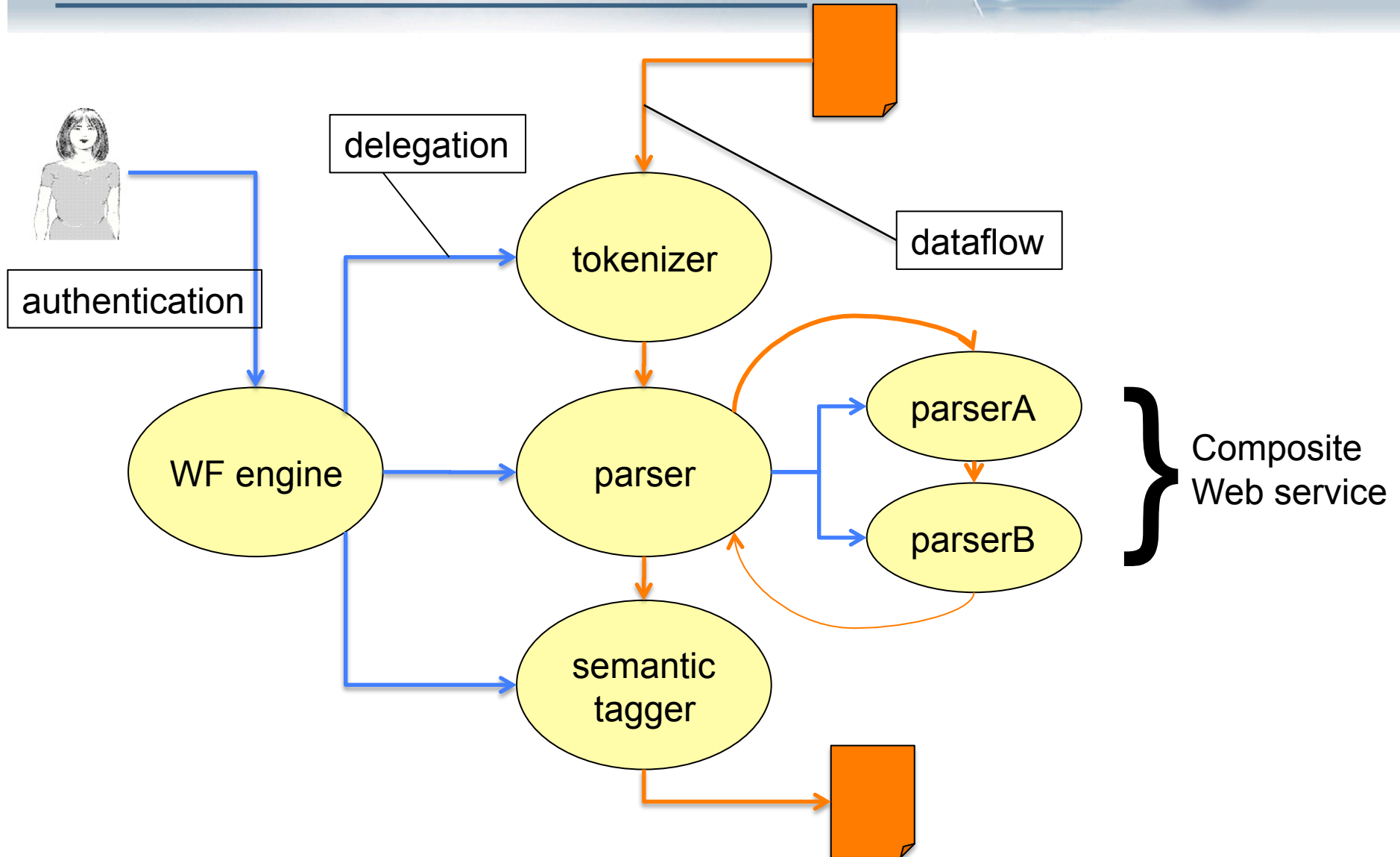
Archive A

Archive B

TF-EMC2 Meeting, May 5, 2009
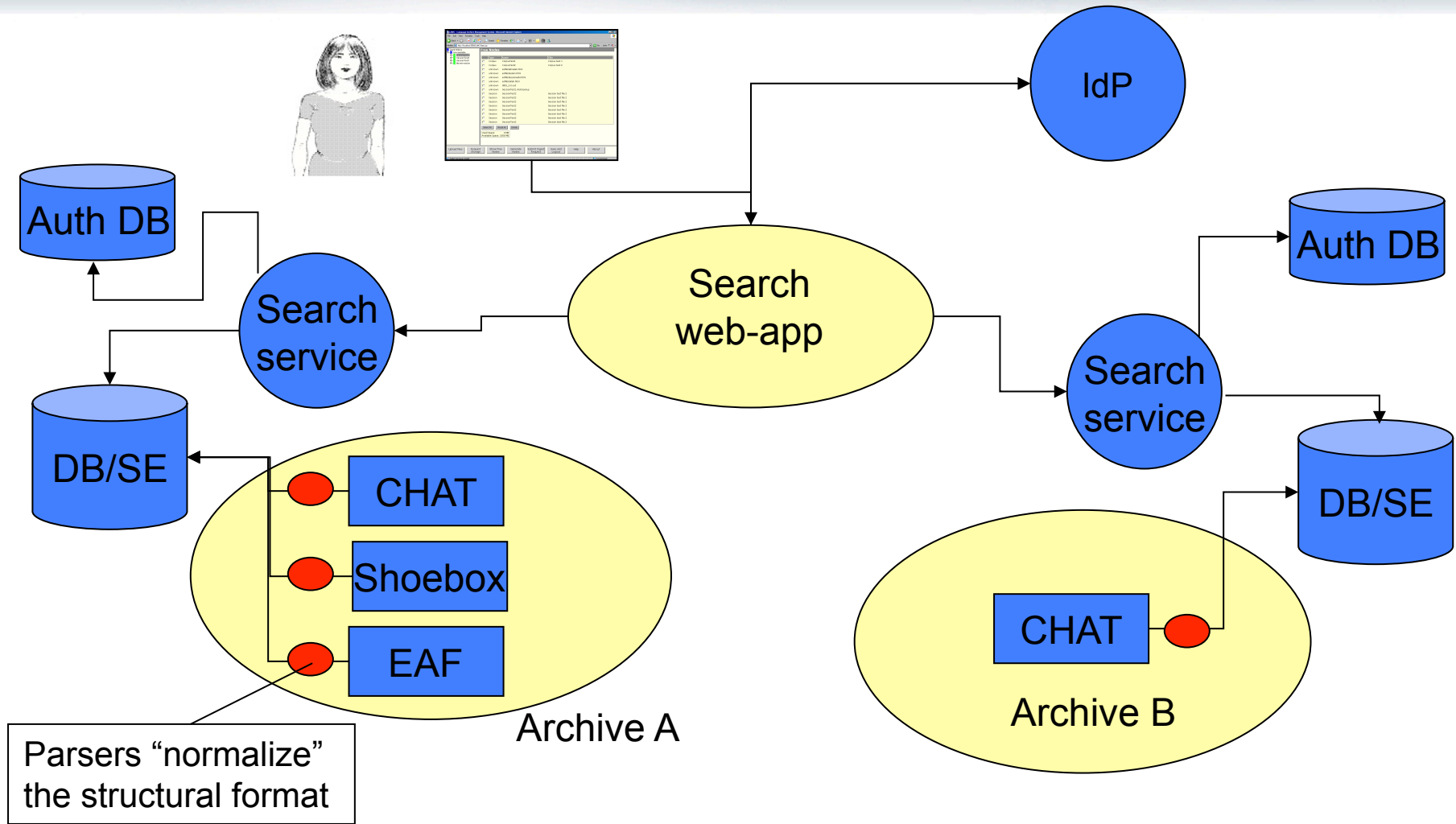
The search web-app acts on behalf of the user when it accesses the search services.

Workflow using chains of web services

- User specifies processing chains of web-services
- Every web service must be able to act on behalf of the user
  - Access resources
  - Store results
  - Call other web-services (and again delegate)
- Solutions under investigation:
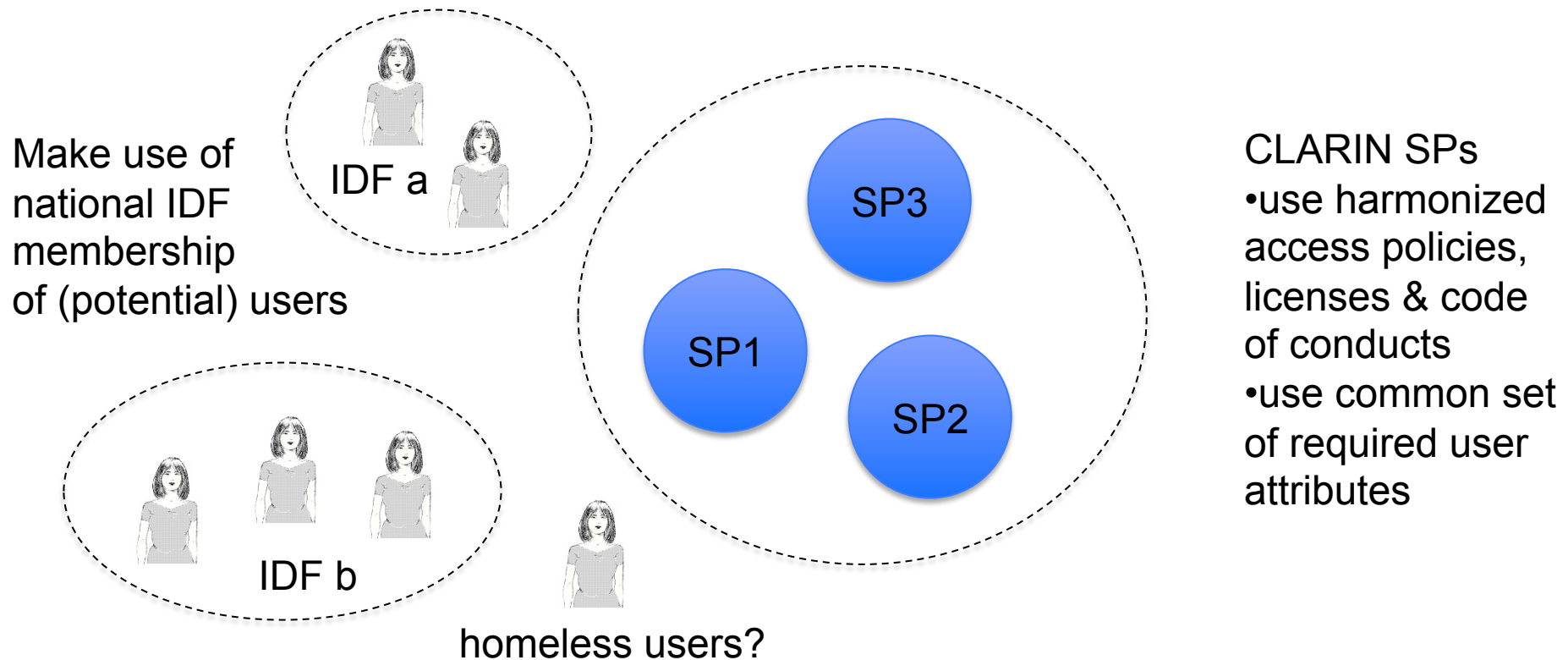 - WS Security (security token, X.509, ...)
  - ...

WS Workflow chains

delegation

dataflow

authentication

tokenizer

WF engine

parser

parserA

parserB

} Composite
Web service

semantic
tagger

TF-EMC2 Meeting, May 5, 2009

Long term AAI objectives I

- Establish a CLARIN SP organization as a legal entity able to sign contracts where needed.

- Rely on user's home organization membership of national IDFs for establishing trust relations with the SPs

Make use of national IDF membership of (potential) users

IDF a

IDF b

homeless users?

SP3

SP1

SP2

CLARIN SPs
•use harmonized access policies, licenses & code of conducts
•use common set of required user attributes

Long term AAI objectives II

- Agree on accepted set of user attributes that should be provided by all IdPs involved.
- If possible harmonize licenses & code of conducts so a central registry may be established for these untypical user attributes
    - Implement as External AA for use in authorization process
- Provide for all identified CLARIN AAI scenarios

SAML

SP

IdP

WS API

SAML

Ext. AA
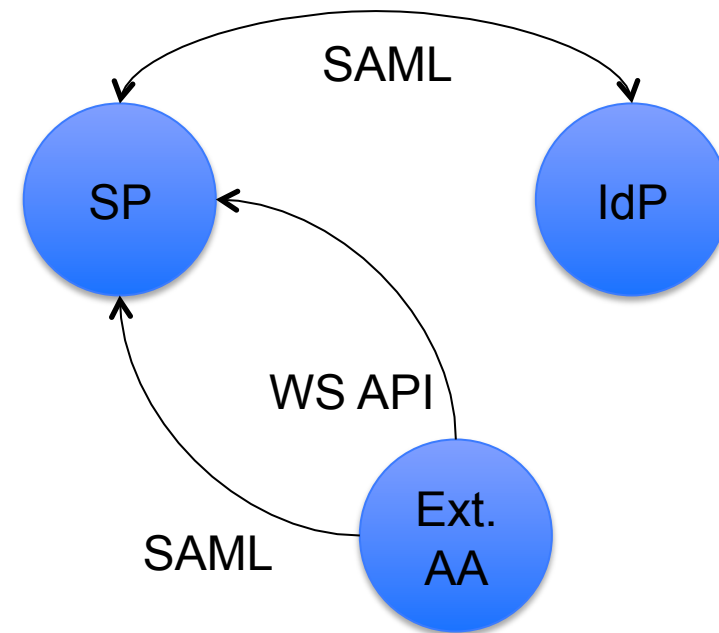
eduGAIN

- Recent meeting WP2/WP7 discussing possible eduGAIN role
- Important to maximize potential user group: all those part of national IdFs
- Minimize bureaucratic overhead (contracts) and configuration for the SPs
- Guide or solve the user attribute harmonization question
- Advice & help in realizing the (more complicated) AA scenarios

Short term AAI objectives I

- To safeguard further financing for the national CLARIN initiatives
- Demo of federated identity and SSO access by end 2009
    - Users with different home organizations in EU
    - Representative set of CLARIN SPs in EU
- Show simple case of SSO with a web browser.
- SPs in this context are repository systems that are able to deliver resources on request

Short term AAI objectives II

- Decided for demo setup:
    - 3 IDFs: HAKA, DFN, SurfFed
    - 6 SPs: BBAW, IDS, INL, Meertens Inst, MPI Psyl, CSC
    - All SPs will use SAML 2.0 (Shibboleth, SimplSAMLPhP)
- Legal entity formation for demo setup is being investigated. Can be quick & dirty.
- Technical realization
    - All will use SAML 2.0
    - SPs will need to harvest IDF metadata and become member of all IdFs
    - ePPN will be the only required attribute

Thank you for your attention