CLARIN
Common Language Resources and Technology Infrastructure

CLARIN and the Humanities

Daan Broeder

The Language Archive – MPI for Psycholinguistics

CLARIN EU/NL

Workshop on Federated Identity Management
CERN, June 9-10 2011

CLARIN

Common Language Resources and Technology Infrastructure

CLARIN is an ESFRI roadmap Research Infrastructure project

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**.

Its target audience is mainly academic researchers, not only linguists but all from the wider SSH

- Text mining technology on historical texts for historians
- Opinion mining from newspaper corpora for social scientists

Language Resources

Any resource used to study language

- Text Corpora
  - Newspapers,…, email, sms messages
- Multi-media corpora
  - Audio recordings to study phonetics, train speech recognizers
  - Video recordings for Sign-Language studies
  - Language Documentation (language use in cultural context)
- Multi-Media Lexica
  - Lexical entries linked with pictures, sound

Our data collections are not particularly large. ~100 TB for the MPI-PL archive. But the possible relations between language resources and their constituent parts can be complex

CLARIN “Holy Grail” User Scenario

- A researcher authenticates at his 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.
- To be granted access to this distributed dataset he signs the appropriate licenses
- He is then able to use a workflow specification tool and process this virtual collection using LT tools in the form of reliable distributed web services which he is authorized to use.
- Results are stored in a user specific workspace
- After evaluation, the resulting data (including metadata) can be added to a repository and the “virtual” collection specification can be stored for future reference using PIDs with proper access rights.

CLARIN AAI

- Purpose is to create one single domain of CLARIN resources and services for our users
  - Where users have only one identity (and since we hope to have very many users) preferably maintained at their home institute
  - and can use SSO between services at different centers
  - Users have to sign a license only once
- Our users are linguists and SSH academics spread out over Europe, CLARIN can not hope to influence the way their user accounts are set-up.
  - But CLARIN can profit from existing AAI infrastructure in the research & education domain.
- CLARIN centers are part of the CLARIN organization and they can be asked to conform to CLARIN needs.

The national IDFs & eduGAIN

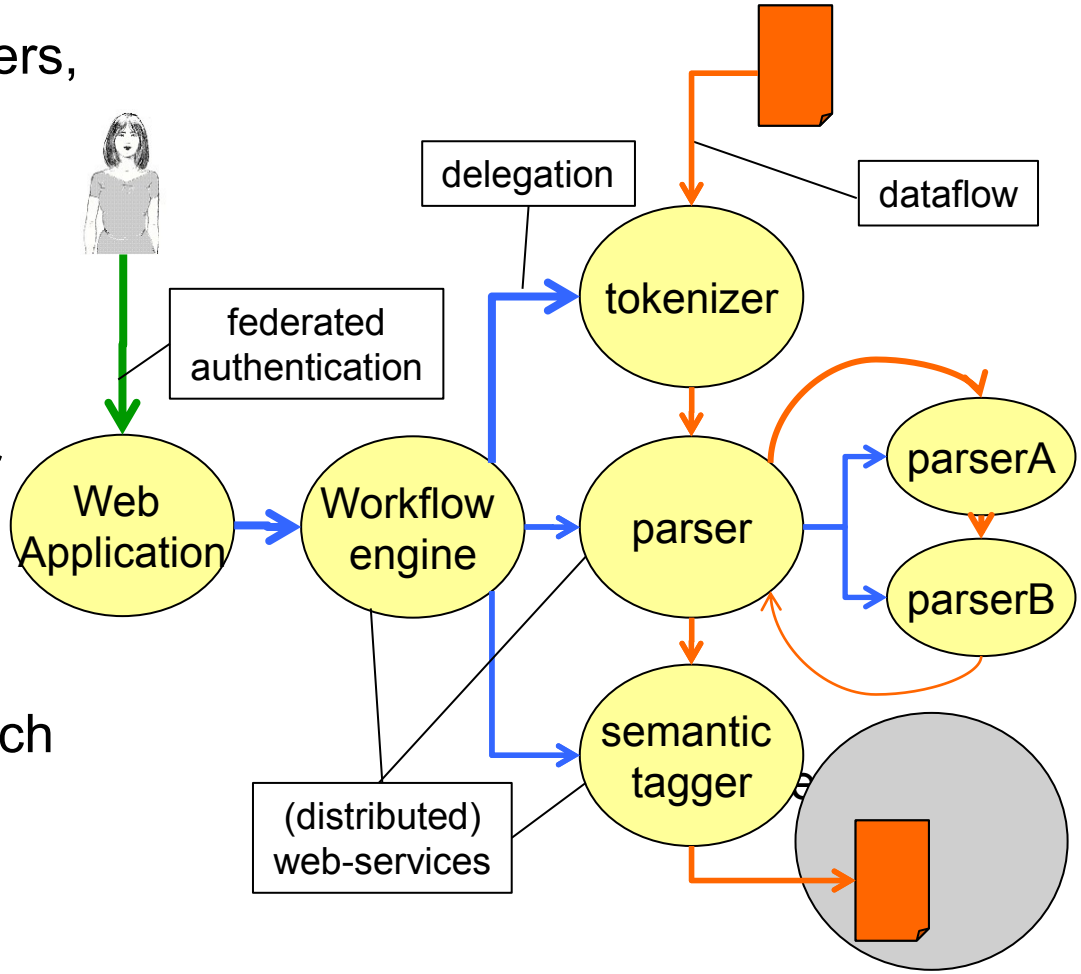
- Seemed obvious the use the national IDFs
- … and in particular the “eduGAIN” interfederation at that moment a pilot project. Hoped for:
  - transparent participation for SPs and IdPs
  - attribute harmonization
- CLARIN authz on basis of identity, signed licenses
- Only use ePPN although (email & organization would be nice).
- If specific attributes required then probably set-up CLARIN VO-Platform
- Delay in availability eduGAIN led to creating the **CLARIN SP Federation**
  - 3 IDFs: HAKA, DFN-AAI, SURFfed
  - 9 CLARIN SPs (4 on-line), one with power of attorney as coordinating party.
  - Asymmetric relations with FR, TSJ, A,
- Created a home for the homeless

Obstacles for federated identity use & acceptance

- Unfamiliarity of users with the technology
  - WAYF: where do I find my organization, what is my IDF (two step)
  - ARC: prompting user consent for attribute release (uApprove)
  - Need careful guiding of inexperienced users
- Scaling problems
  - Does eduGain have an opt-in policy? Every IdP has to allow its users access the inter-federation or worse individual SPs
  - Individual IDF can also have an opt-in policy. Every IdP has to agree to have its users access CLARIN SPs
  - Hopefully they can treat the CLARIN SPs as a single entity
- WAYF SPOF, deploying several will break the SSO

Web service security/delegation in workflows

- CLARIN is also about language technology: parsers, tokenizers, etc.
- In CLARIN SOA these are offered as (REST) web services and operated by workflow engines
- Problem of delegating user control from the controlling web application to the participating WSs
- In cooperation with the Dutch NGI investigating solutions using ‘security tokens’ as OAuth2



DARIAH
RI for the Arts and Humanities

DARIAH
Digital Research Infrastructure for the Arts and Humanities

- Goal
  - Shibboleth-based federation across Europe, ideally eduGAIN
  - shared approach with other SSH infrastructures, e.g. CLARIN and CESSDA in DASISH
  - explore integration with user-centered approaches (e.g. OpenID)

- Experiences and existing systems
  - VRE-Integration of homeless users [TextGrid/D-Grid]
  - Job-Submission (e.g. Globus, gLite) through Shibboleth, based on Robot Certificates and Short-Lived-Credentials [GAP-SLC/D-Grid]
  - Design of attributes and attribute integration [with DFN/AAI]

Humanities & Social Sciences

- 5 ESFRI projects:
  - CLARIN - Language Resources
  - DARIAH - Wider Humanities
  - CESSDA - Social Sciences
  - SHARE, ESS - Survey Oriented

CLARIN
Common Language Resources and Technology Infrastructure

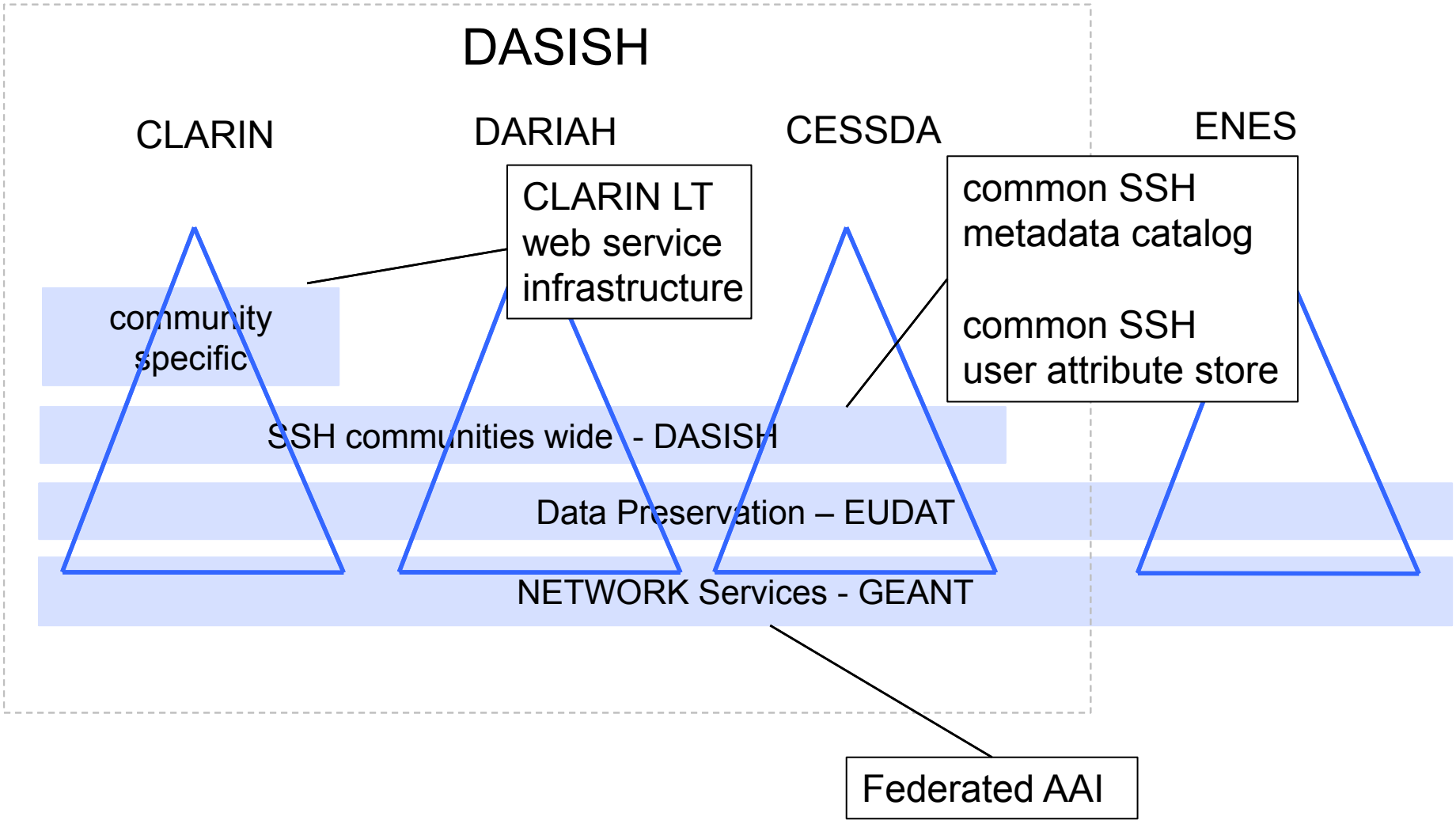
European Social Survey

metadata accessing data data harmonisation ethics cessda social science data data catalogue confidentiality data preservation

SHARE
Survey of Health, Ageing and Retirement in Europe
50+ in Europe

- DASISH – Digital Services Infrastructure for the SS and Humanities
- A EU cluster project of the SSH ESFRI projects: CLARIN, CESSDA, DARIAH, ESS, SHARE
- Exploiting the commonalities of those projects and building on their achievements

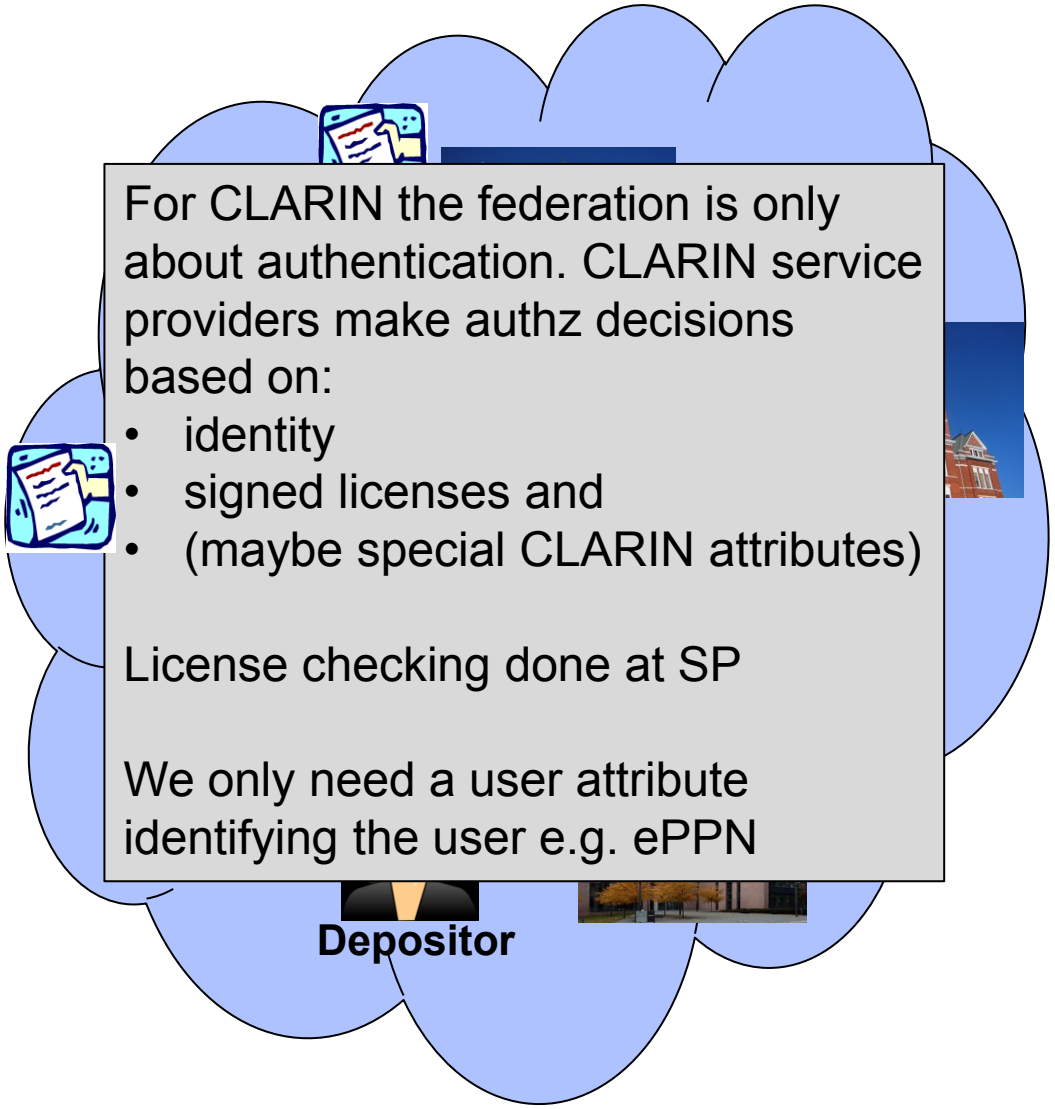
CLARIN in context



Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230

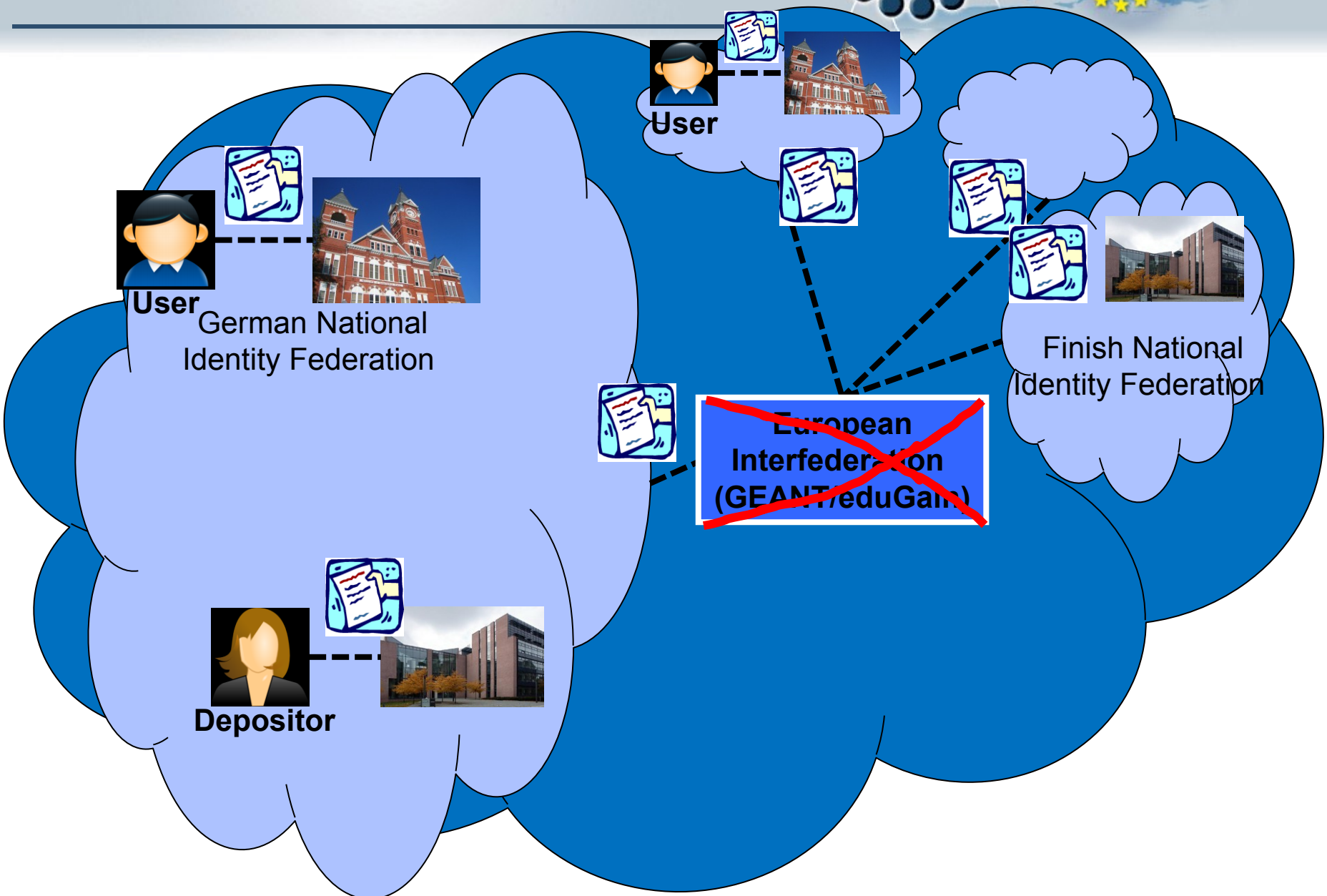
National Trust Domain



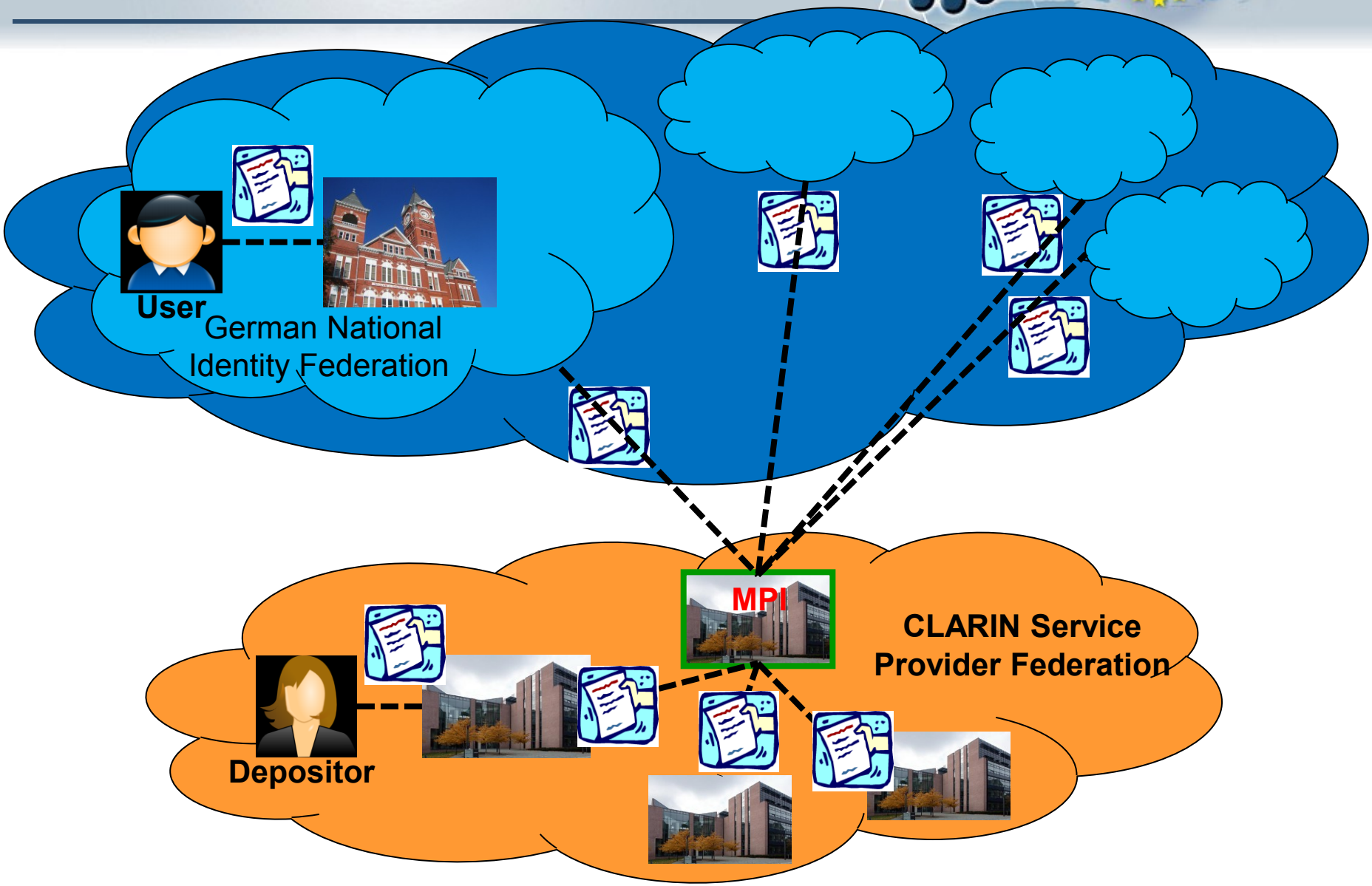
- Depositor:
  - Mary Lamb may see my data
  - If she signs the code of conduct “only for academic use”
- User organization
  - This is Mary Lamb

Seems very scalable provided users are easily connected to new service providers without much overhead for them

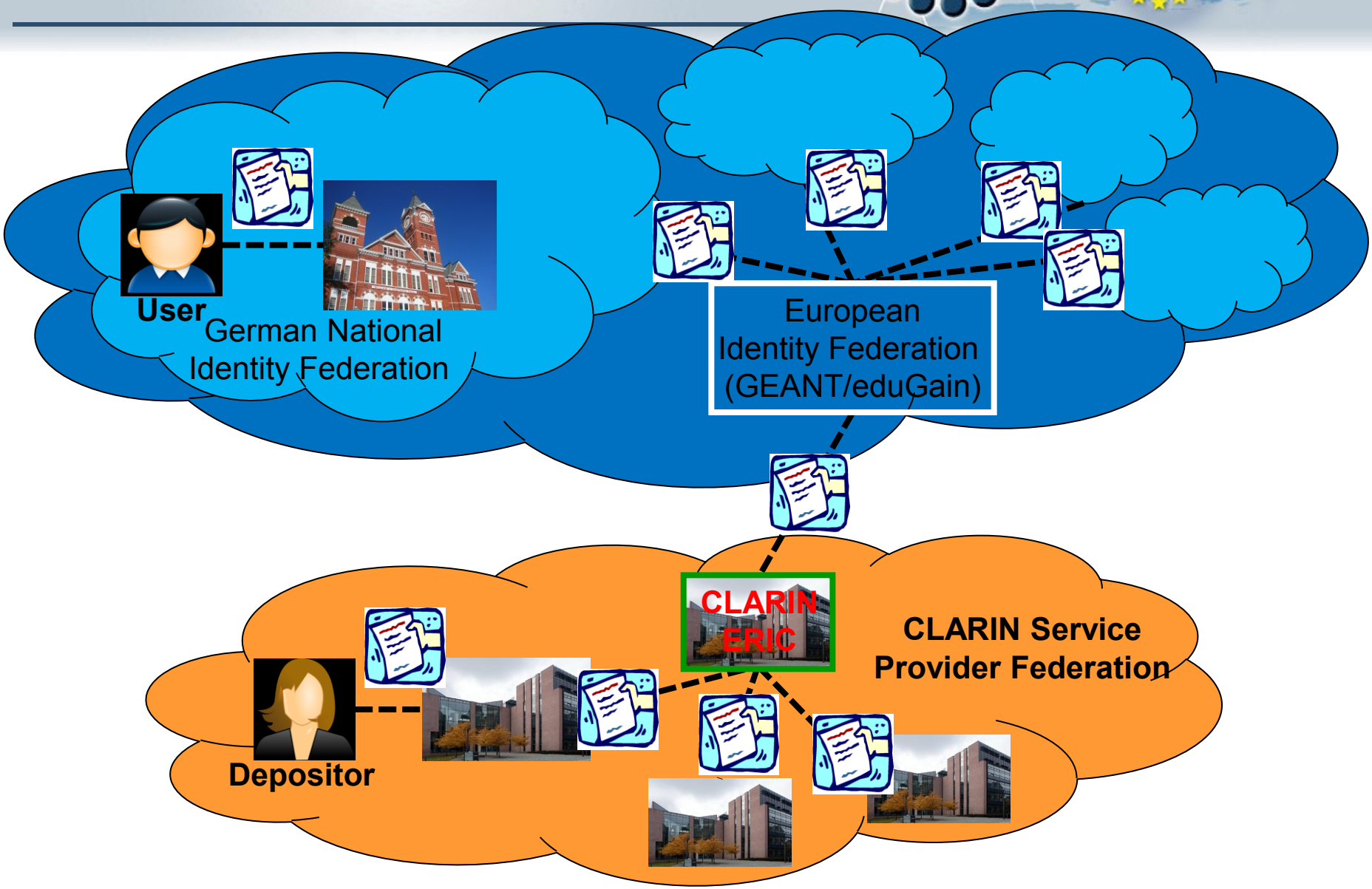
European Trust Domain



CLARIN SPF


User
German National
Identity Federation
MPI
CLARIN Service
Provider Federation
Depositor

CLARIN SPF


User
German National
Identity Federation
European
Identity Federation
(GEANT/eduGain)
CLARIN
ERIC
CLARIN Service
Provider Federation
Depositor

Current State CLARIN SPF