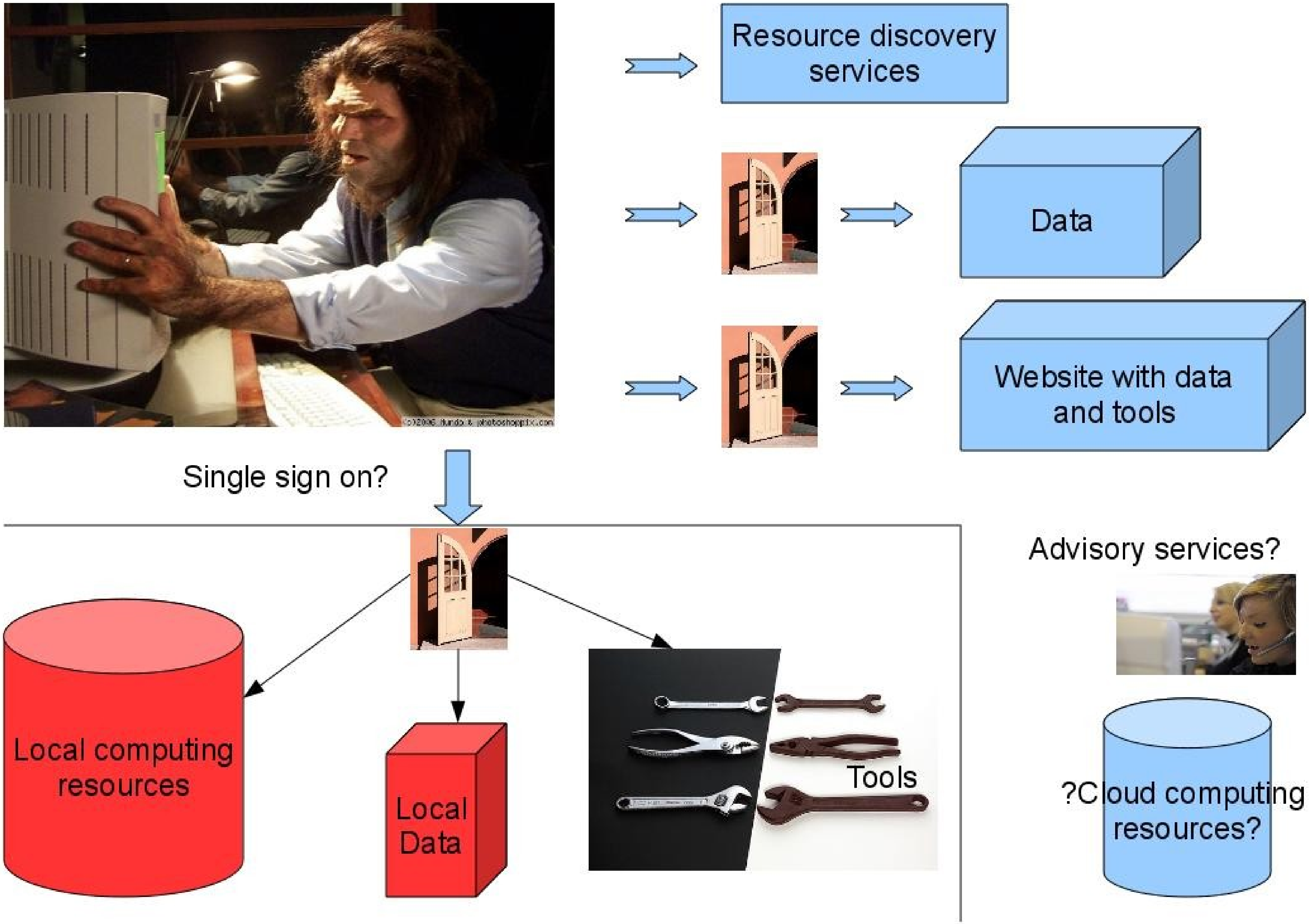
CLARIN: Common Language Resources and Technology Infrastructure

Martin Wynne
Head of the Oxford Text Archive
Oxford e-Research Centre and
Oxford University Computing Services
University of Oxford
martin.wynne@oucs.ox.ac.uk



Resource discovery services

Data

Website with data and tools

Single sign on?

Advisory services?

Local computing resources

Local Data

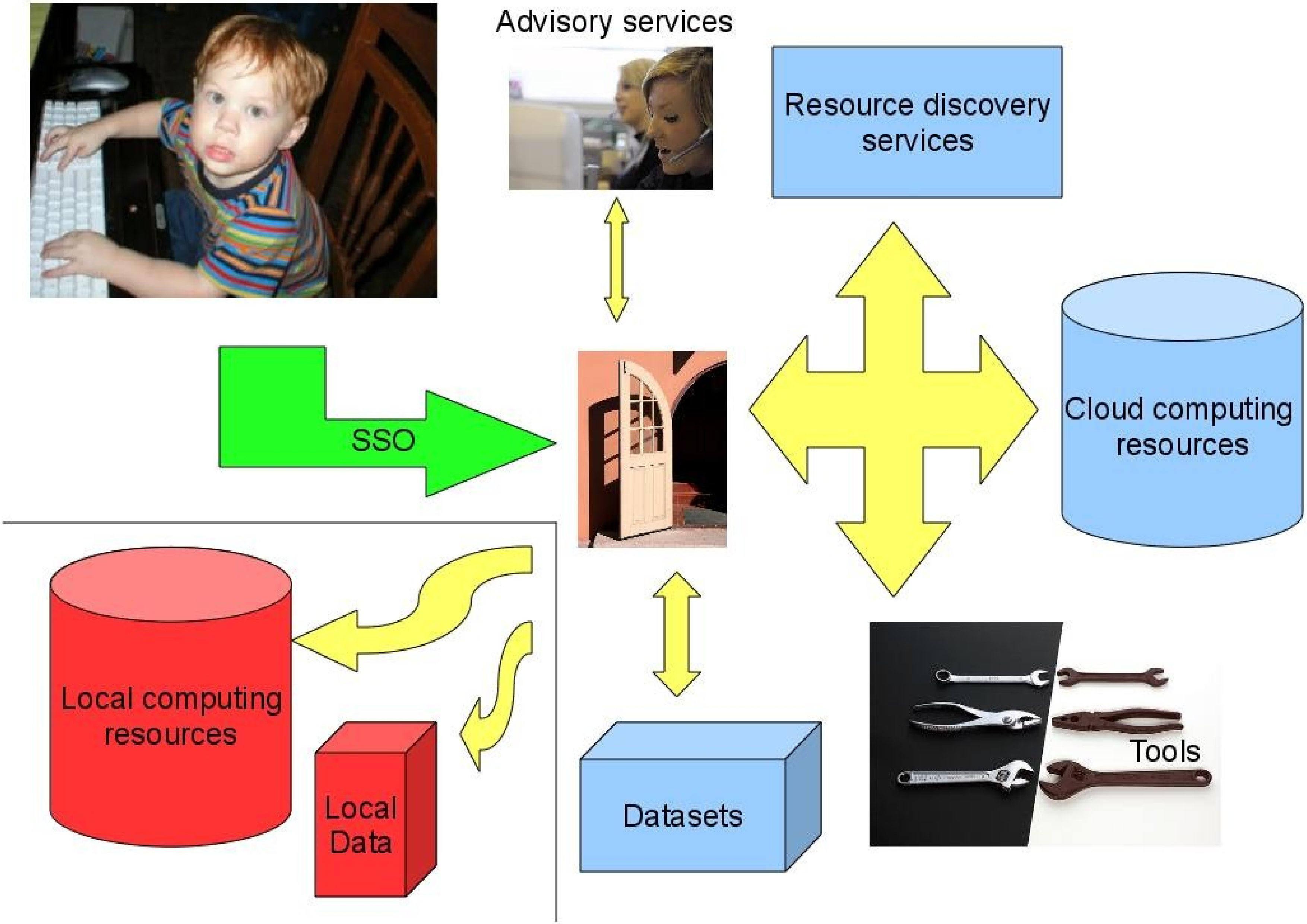
Tools

?Cloud computing resources?

The Vision

A researcher in Amsterdam from his desktop computer can:

- do a single sign-on with local authentication, and then:
- search for, find and obtain authorization to use corpora in Utrecht, Prague and Berlin
- select the precise dataset to work on, and save that selection
- run semantic analysis tools from Budapest and statistical tools from Nancy over the dataset
- use computational power from the local or national computing centre where necessary
- save the workflow and results of the analysis, and share those results with collaborators in Oxford, Vienna and Zagreb
- discuss and iteratively adopt and re-run the analyses with collaborators



Advisory services

Resource discovery services

SSO

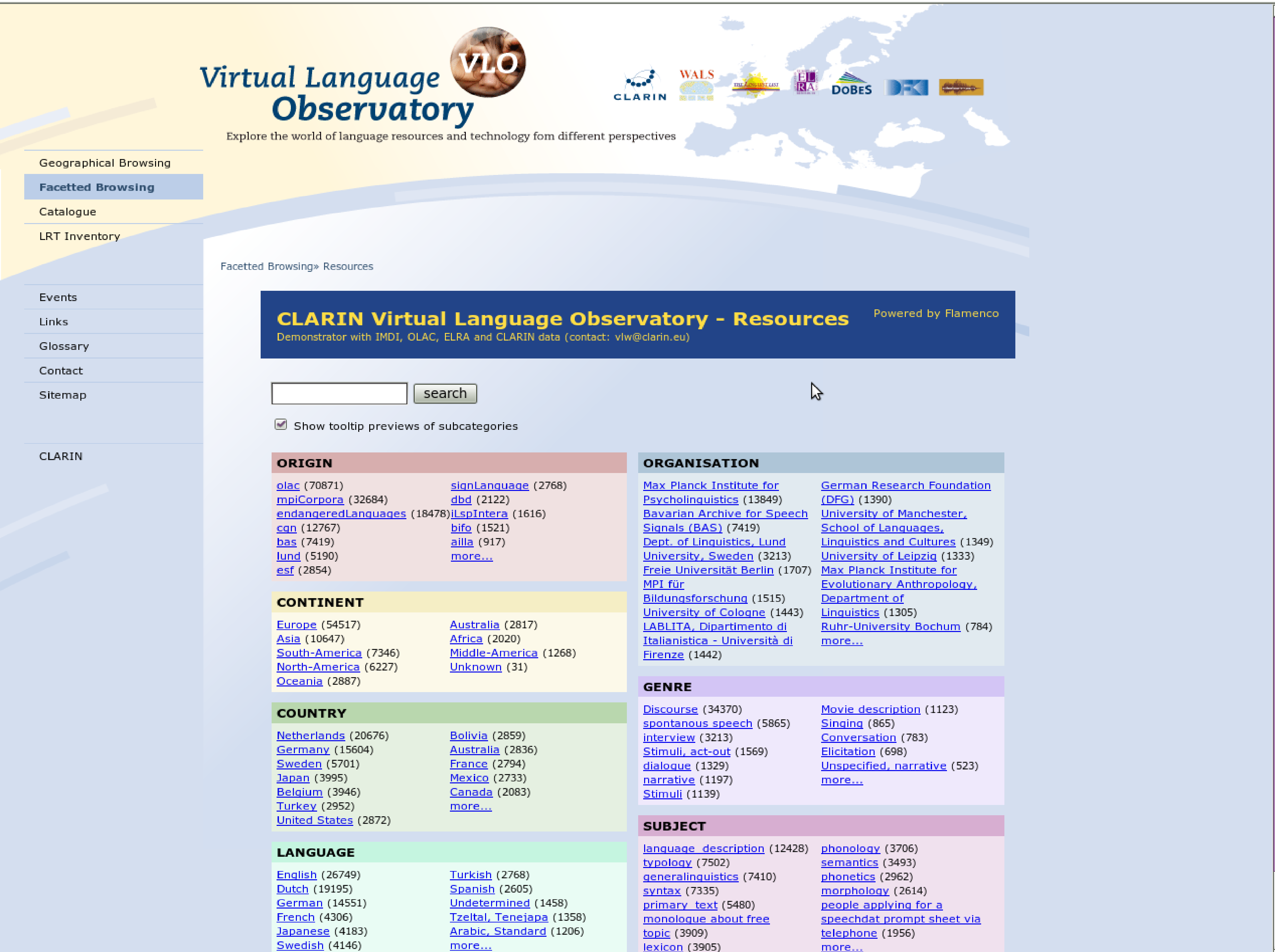
Cloud computing resources

Local computing resources

Local Data

Datasets

Tools



Virtual Language Observatory

Explore the world of language resources and technology fom different perspectives

- Geographical Browsing
- **Facetted Browsing**
- Catalogue
- LRT Inventory

- Events
- Links
- Glossary
- Contact
- Sitemap

- CLARIN

Facetted Browsing» Resources

CLARIN Virtual Language Observatory - Resources

Demonstrator with IMDI, OLAC, ELRA and CLARIN data (contact: vlw@clarin.eu)

Powered by Flamenco

search

☑ Show tooltip previews of subcategories

ORIGIN

olac (70871)
mpiCorpora (32684)
endangeredLanguages (18478)
cgn (12767)
bas (7419)
lund (5190)
esf (2854)
signLanguage (2768)
dbd (2122)
iLspIntera (1616)
bifo (1521)
ailla (917)
more...

CONTINENT

Europe (54517)
Asia (10647)
South-America (7346)
North-America (6227)
Oceania (2887)
Australia (2817)
Africa (2020)
Middle-America (1268)
Unknown (31)

COUNTRY

Netherlands (20676)
Germany (15604)
Sweden (5701)
Japan (3995)
Belgium (3946)
Turkey (2952)
United States (2872)
Bolivia (2859)
Australia (2836)
France (2794)
Mexico (2733)
Canada (2083)
more...

LANGUAGE

English (26749)
Dutch (19195)
German (14551)
French (4306)
Japanese (4183)
Swedish (4146)
Turkish (2768)
Spanish (2605)
Undetermined (1458)
Tzeltal, Tenejapa (1358)
Arabic, Standard (1206)
more...

ORGANISATION

Max Planck Institute for Psycholinguistics (13849)
Bavarian Archive for Speech Signals (BAS) (7419)
Dept. of Linguistics, Lund University, Sweden (3213)
Freie Universität Berlin (1707)
MPI für Bildungsforschung (1515)
University of Cologne (1443)
LABLITA, Dipartimento di Italianistica - Università di Firenze (1442)
German Research Foundation (DFG) (1390)
University of Manchester, School of Languages, Linguistics and Cultures (1349)
University of Leipzig (1333)
Max Planck Institute for Evolutionary Anthropology, Department of Linguistics (1305)
Ruhr-University Bochum (784)
more...

GENRE

Discourse (34370)
spontanous speech (5865)
interview (3213)
Stimuli, act-out (1569)
dialogue (1329)
narrative (1197)
Stimuli (1139)
Movie description (1123)
Singing (865)
Conversation (783)
Elicitation (698)
Unspecified, narrative (523)
more...

SUBJECT

language_description (12428)
typology (7502)
generalinguistics (7410)
syntax (7335)
primary_text (5480)
monologue about free topic (3909)
lexicon (3905)
phonology (3706)
semantics (3493)
phonetics (2962)
morphology (2614)
people applying for a speechdat prompt sheet via telephone (1956)
more...

CLARIN technical work

Promoting collaboration and interoperability between European language resource repositories, particularly in relation to:

- Persistent identifiers
- Component metadata
- A trust domain and a service provider federation
- Service centres
- Virtual collections
- Standards and best practices
- ...and more!

See the CLARIN Short Guides at http://www.clarin.eu/

A cautionary note...

CLARIN is building a research infrastructure for the use of all researchers in the Humanities and Social Sciences, and beyond, but **NOT** for computer scientists. We can have few or no expectations about technical expertise and support.

Are GRID services sufficiently mature and user-friendly for us?

Requirements

Service providers

Persistent identifiers

Single sign-on

Trust federation

Web services

Resource discovery

Resource preservation

Monitoring of service availability

Brokering and accounting compute resources

Requirements

Service providers	*existing and emerging*
Persistent identifiers	*EPIC handle service*
Single sign-on	*national identity federations*
Domain of trust	*CLARIN initial federation*
Web services	*RESTful web services*
Resource discovery	*our own metadata sets, OAI-PMH*
Resource preservation	*separate centre policies*

Monitoring of service availability ?

Brokering and accounting compute resources ?

CLARIN Centres

Perhaps our biggest concern is the sustainability of the centres that provide the CLARIN services.

What can we do to embed our centres more firmly in national and European infrastructure to make them more sustainable?

Virtual research communities

CLARIN does not serve one *community.* Some thoughts on how the Humanities and Social Sciences might be different...

- Groups of potential users are not easily classified as *virtual research communities*
- A “project” might be a life's work
- Legal and ethical restrictions apply to many datasets
- Tools are not necessarily associated with particular resources, subject domains or research activities
- 'Multilingual research' can mean discrete domains of monolingual research
- *humani nil a me alienum puto:* the objects of study are largely outside of our repositories and are potentially boundless

Thank you for your attention

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