CLARIN-D Report R2.3: Third Year Activity Report
Liaison with the CLARIN ERIC and Liaison with DARIAH-DE

May 2014
Deliverable: R3.2a/b
Third Year Activity Report
Liaison with the CLARIN ERIC and Liaison with DARIAH-DE

Responsible: Erhard Hinrichs
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2 Liaison activities within CLARIN</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Liaison activities with CLARIN ERIC</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Liaison activities with foreign CLARIN partners</td>
<td>8</td>
</tr>
<tr>
<td>3 Liaison with DARIAH-DE</td>
<td>10</td>
</tr>
<tr>
<td>3.1 Joint Technical Advisory Board</td>
<td>10</td>
</tr>
<tr>
<td>3.2 International Meetings and Dissemination</td>
<td>10</td>
</tr>
<tr>
<td>4 Liaison with other initiatives</td>
<td>11</td>
</tr>
<tr>
<td>4.1 BMBF-funded eHumanities projects</td>
<td>11</td>
</tr>
<tr>
<td>4.2 EUDAT – European Data Infrastructure</td>
<td>11</td>
</tr>
<tr>
<td>4.3 Liaison activities with other initiatives</td>
<td>12</td>
</tr>
<tr>
<td>5 Conclusion and further perspectives</td>
<td>13</td>
</tr>
</tbody>
</table>
1 Introduction

CLARIN-D report R3.2 "Third Year Activity Report – Liaison with the CLARIN ERIC and Liaison with DARIAH-DE" documents liaison activities during the third year of the CLARIN-D project. It presents an overview of cooperation within the CLARIN-D project as a whole, as well as the cooperative efforts of the individual CLARIN-D centres.

The CLARIN-D project1 (Web and Centre-based Infrastructure for the Social Sciences and Humanities, HSS) is motivated by a vision of providing humanities scholars with single-sign-on access to diverse repositories of language and textual data and the tools to work with them. It is a collaborative effort involving institutions and researchers at nine academic centres:

- Bavarian Archive for Speech Signals (BAS), Ludwig Maximilians University of Munich
- Berlin Brandenburg Academy of Sciences and Humanities (BBAW)
- Institute for the German Language, Mannheim (IDS)
- Max Planck Institute for Psycholinguistics, Nijmegen (MPI)
- University of Tübingen, Department of General and Computational Linguistics
- University of Hamburg, Hamburg Centre for Speech Corpora (HZSK)
- University of Leipzig, Institute of Computer Science
- Saarland University, Department of English and Translation Studies
- University of Stuttgart, Institute for Natural Language Processing

CLARIN-D is building on the achievements of the European CLARIN (Common Language Resource and Technology Infrastructure) initiative as well as the German national project D-SPIN (Deutsche Sprachressourcen-Infrastruktur). CLARIN and D-SPIN developed a set of common standards and goals for humanities research infrastructure as well as plans for the sustainable provision of digital resources and their long-term archiving.

D-SPIN’s funding was provided by the German Federal Ministry of Education and Research (BMBF), while CLARIN was funded directly by the European Commission. Both were part of the European Seventh Framework Programme (FP7)2 as well as the ESFRI Roadmap3 (European Strategy Forum on Research Infrastructures).

FP7 is a strategic plan to encourage research and innovation in the European Union and associated states. It follows the first six framework programmes, which operated from 1984 through 2006, and covers the period from 2007 through 2013, encompassing roughly 50

1 http://www.clarin-d.de/
2 http://cordis.europa.eu/fp7/home_en.html
3 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
billion euros of funding activities. It is due to be replaced by the Eighth Framework Programme (FP8) in 2014.

ESFRI is an on-going European planning framework to coordinate research activities across Europe in order to establish competitive, open-access, continent-wide infrastructure designed to better serve research communities. It promotes the development of cross-border research facilities and services covering broad areas of scientific and academic interest.

ESFRI research infrastructure projects are planned with three sequential phases: a preparatory phase, an implementation phase, and an exploitation phase. D-SPIN was a preparatory phase project, during which plans for digital humanities infrastructure were developed through technical research, prototypes, models for training researchers, and through clarification of the legal framework for the project. CLARIN-D is the implementation phase successor to D-SPIN. Funding during the implementation phase of the ESFRI roadmap comes entirely from participating European countries that directly fund national projects to fulfil the roadmap's goals.

D-SPIN’s European preparatory phase partner (CLARIN) has been replaced by CLARIN ERIC, which is structured as an association of national member projects. This new organization is a European Research Infrastructure Consortium (ERIC)\(^4\), – a new category of legal entity established in European community law as part of ESFRI in order to facilitate the establishment and operation of joint research infrastructures of Europe-wide interest.

CLARIN ERIC currently has nine members, eight of whom are national project organizations joining together researchers and institutions in their respective states, and one intergovernmental body:

- Austria\(^5\)
- Bulgaria
- Czech Republic\(^6\)
- Germany\(^7\)
- Denmark\(^8\)
- Estonia\(^9\)
- The Netherlands\(^10\)

\(^5\) [http://www.oeaw.ac.at/icltt/node/14](http://www.oeaw.ac.at/icltt/node/14)
\(^7\) [http://www.clarin-d.de](http://www.clarin-d.de)
\(^8\) [https://clarin.dk/clarindk/forside.jsp](https://clarin.dk/clarindk/forside.jsp)
• Poland\textsuperscript{11}

• The Nederlandse Taalunie\textsuperscript{12}, a joint Dutch-Flemish intergovernmental organization responsible for promoting and maintaining standards for the Dutch language

Norway has also joined as an observer in January 2013 with the intent to become a full member. Membership is open to European Union countries and EU-associated entities such as intergovernmental organizations as well as individual centres. Several other countries are in the process of joining CLARIN ERIC.

CLARIN-D and the other national CLARIN participants are encouraged to cooperate and coordinate their activities with other projects, at the national and local levels as well as internationally. On the national level, CLARIN-D has aligned some of its activities with DARIAH-DE (Digital Research Infrastructure for the Humanities), another FP7/ESFRI roadmap project with a structure similar to that of CLARIN, focused on digital research infrastructure for arts and humanities research. DARIAH-DE is funded by the German Federal Ministry of Education and Research (BMBF) and is the German national participant in the European DARIAH-EU research infrastructure project.

CLARIN-D members are also involved in a number of BMBF-funded eHumanities projects, often bringing their experience with CLARIN-D and its infrastructure to these initiatives. Furthermore, as academic centres and universities, all CLARIN-D members are involved in multiple national and international projects, and they are encouraged to promote and use the CLARIN-D infrastructure whenever it can serve the scientific goals of researchers.

This report encompasses the liaison and cooperative activities of CLARIN-D centres and CLARIN-D as a whole with CLARIN ERIC, with other CLARIN ERIC members, with DARIAH-DE, with BMBF eHumanities projects, and with other national and international initiatives.

\textsuperscript{10} http://www.clarin.nl

\textsuperscript{11} http://www.clarin-pl.eu/en/

\textsuperscript{12} http://taalunie.org/wat-doet-taalunie/activiteiten/clarin
2 Liaison activities within CLARIN

This section summarizes the liaison activities of CLARIN-D centres with CLARIN ERIC\(^\text{13}\) and foreign CLARIN centres. CLARIN-D centres have a great amount of autonomy to initiate contact with CLARIN ERIC and work together with foreign CLARIN centres within the framework of this project.

2.1 Liaison activities with CLARIN ERIC

CLARIN-D partners participate at all levels of CLARIN ERIC. Erhard Hinrichs of the University of Tübingen is a member of the CLARIN ERIC Board of Directors (BoD) in his capacity as chair of the CLARIN ERIC National Coordinators Forum (NCF). Dieter van Uytvanck of the Max Planck Institute for Psycholinguistics also sits on the board of directors by virtue of being the chair of the Standing Committee for CLARIN Technical Centres (SCCTC). Dr. Heige Kahler (BMBF) is President of the CLARIN ERIC.

Other CLARIN-D personnel serve in various roles within CLARIN ERIC, including:

- Axel Herold of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) is a member of the CLARIN ERIC Standards Committee as well as National Content Manager for CLARIN-D.

- Erik Ketzan of the Institut für Deutsche Sprache (IDS) is chair of the CLARIN ERIC Legal Issues Committee.

- Heike Stadler (IDS) serves as Deputy National Content Manager.

- Thorsten Trippel of the University of Tübingen is a member of the CLARIN Office, providing support services for CLARIN ERIC.

- Andreas Witt (IDS) is the Chairman of the CLARIN ERIC Standards Committee.

CLARIN-D signed the renewal of the CLARIN Agreement (coordinated by Erhard Hinrichs and Kathrin Beck, University of Tübingen), which specifies the national contributions of CLARIN-D members to CLARIN ERIC. Germany’s CLARIN Agreement renewal was signed in November 2012 and is effective until April 2014.

CLARIN-D centres and personnel participated in the CLARIN ERIC conference held in Prague, October 21st-23rd 2013:

- Showcases
  - Textual emigration analysis, Stuttgart, presented by Andre Blessing: Historians, literary scientists, and others are interested in the semantic interpretation of text. With automatic pre-processing of texts, e.g. named entity recognition, coreference resolution, and dependency parsing, relevant semantic relations can be extracted. The Stuttgart tools extract relations of migration, i.e. name and job of the migrant and date, reason and destination of migration. A graphical

\(^{13}\) http://www.clarin.eu/
interface showing timeline and geographic migrations will be implemented.

- WebMAUS, MPI and BAS, presented by Dieter van Uytvanck and Thomas Kisler: The web application WebMAUS allows the user to automatically align speech recordings to their corresponding text form. Two input files need to be uploaded by the user: a media file containing a recorded speech signal and a file containing some textual encoding of the words spoken in the recording.

- Achievements in the “points of pride” section
  - TLA Archive & Virtual Language Observatory, Dieter Van Uytvanck, The Language Archive/Max Planck Institute
  - TLA applications, Sebastian Drude, The Language Archive/Max Planck Institute
  - IMS Web Services: Toolchains and Applications, Jens Stegmann, IMS, University of Stuttgart
  - CLARIN-D: Working closely with the community, Erhard Hinrichs, University of Tübingen
  - CLARIN-D user relations, Volker Boehlke, University of Leipzig

- Presentations on the CLARIN infrastructure: State of the technical infrastructure at the Annual Meeting 2013, presented by Dieter van Uytvanck

Several CLARIN-D centres participated in the CLARIN ERIC Federated Content Search (FCS) workshop in Copenhagen on 24 April 2013 demonstrating and discussing the solutions for Federated Content Search and how this could be used on a European scale.

2.2 Liaison activities with foreign CLARIN partners

CLARIN ERIC appointed a director for user involvement who will function as a dedicated liaison officer for CLARIN as a whole. CLARIN-D also appointed a liaison coordinator to reach out to the interested scientific community outside of the CLARIN-D Discipline-specific Workinggroups ("Facharbeitsgemeinschaften", F-AGs). These two appointments will further increase the visibility of CLARIN among potential users and related projects.

CLARIN-D participates in the CLARIN ERIC booth at the European village at LREC 2014, demonstrating the potential and infrastructure of CLARIN-D to a wider community. Together with the following accepted publications, LREC is a major venue for CLARIN-D:

- Andre Blessing; Jonas Kuhn: Textual Emigration Analysis (TEA)
- Anders Björkelund; Kerstin Eckart; Arndt Riester; Nadja Schaufler; Katrin Schweitzer: The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution
- Chris Culy; Marco Passarotti; Ulla König-Cardanobile: A Compact Interactive Visualization of Dependency Treebank Query Results
- Corina Dima; Verena Henrich; Erhard Hinrichs; Christina Hoppermann: How to Tell a Schneemann from a Milchmann: An Annotation Scheme for Compound-Internal Relations
• Christoph Draxler: Online Experiments with the percy Software Framework - Experiences and some Early Results

• Christoph Draxler; Thomas Eckart; Susanne Haaf; Daniel Jettka; Thomas Kisler; Dieter van Uytvanck: Online Speech and Language Resources

• Thomas Eckart, Erla Hallsteinsdóttir, Sigrún Helgadóttir, Uwe Quasthoff, Dirk Goldhahn: A 500 Million Word POS-Tagged Icelandic Corpus

• Dirk Goldhahn, Uwe Quasthoff: Vocabulary-Based Language Similarity using Web Corpora

• Susanne Haaf; Matthias Schulz: Historical Newspapers & Journals for the DTA

• Fritz Kliche; Andre Blessing; Jonathan Sonntag; Uli Heid: The e-Identity Exploration Workbench

• Scott Martens; Marco Passarotti: Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus Treebank into CLARIN-D

• Uwe Quasthoff, Dirk Goldhahn, Thomas Eckart, Erla Hallsteinsdóttir, Sabine Fiedler: High Quality Word Lists as a Resource for Multiple Purposes

• Florian Schiel; Thomas Kisler: German Alcohol Language Corpus - the Question of Dialect

• Thorsten Trippel; Daan Broeder; Matej Durco; Oddrun Ohren: Towards automatic quality assessment of component metadata

Supporting the establishment of CLARINO (CLARIN in Norway) CLARIN-D participated at the kick-off meeting on 12 and 13 September 2013 in Bergen, Norway.
3 Liaison with DARIAH-DE

This section outlines recent liaison activities between CLARIN-D and DARIAH-DE, including establishing a joint advisory board on technical issues, forming cooperative infrastructure-related working groups, and holding a joint workshop.

3.1 Joint Technical Advisory Board

The joint Technical Advisory Board for CLARIN-D and DARIAH-DE has a face-to-face meeting in Lausanne at the Digital Humanities 2014 in July 2014. A preparatory meeting of DARIAH-D and CLARIN-D in Göttingen was conducted in May 2014. The two projects discussed areas of common interest concerning technical infrastructure, particularly persistent identifiers for resources, identity verification services, and cooperative repository services.

The board consists of the following international experts in digital humanities and computer science:

- Jonas Beskov (University of Stockholm)
- Carol Goble (University of Manchester)
- Jan Hajic (Head of the Prague CLARIN Centre)
- Ed Hovy (University of Southern California)
- Michael Lautenschlager (German Research Centre for Geosciences, Potsdam)
- Gerhard Schneider (University of Freiburg)
- Toma Tasovac (Digital Humanities Centre, Belgrade)
- Melissa Terras (University College London)
- Claire Warwick (University College London)

3.2 International Meetings and Dissemination

CLARIN is involved in discussions with DARIAH to develop a curriculum in digital humanities. As part of that effort, the workshop “Curricula in DH-Studiengängen”, organized by DARIAH, was conducted on 4. September 2013 at Cologne.
4 Liaison with other initiatives

CLARIN-D and the individual CLARIN centres participate in a variety of other research initiatives. CLARIN-D centres are involved in many national eHumanities projects, generally funded by the German Federal Ministry of Research and Education (BMBF). CLARIN-D also plays a major role in the ESFRI/FP7 project EUDAT (European Data Infrastructure), focused on building a European collaborative data infrastructure. Furthermore, as important academic centres in their various specialties, CLARIN-D members are naturally involved in a wide range of other initiatives, to which they bring their CLARIN-D experiences.

4.1 BMBF-funded eHumanities projects

With the establishment of the eHumanities project of the German Federal Ministry of Research and Education (BMBF), CLARIN-D is in close contact with various eHumanities projects. The following eHumanities projects include CLARIN-D centres as partners: AUVIS\(^{14}\), e-Identity\(^{15}\), eXchange/ECCTA\(^{16}\), ePoetics\(^{17}\), GeoBib\(^{18}\), KobRA\(^{19}\) and Postdemokratie\(^{20}\). Additional cooperations exist with BMBF funded eHumanities projects ARGUMENTUM\(^{21}\), SeNeReKo\(^{22}\), VisArgue\(^{23}\).

4.2 EUDAT – European Data Infrastructure

The EUDAT\(^{24}\) project is building a collaborative data infrastructure for researchers to share data with colleagues from all disciplines, helping them to collaborate effectively. The CLARIN community acts as the liaison to the humanities and social science user group in EUDAT.

MPI Nijmegen and the University of Tübingen are partners in both CLARIN-D and EUDAT, and both centres contribute to various EUDAT work packages. Their current work focuses on:

- B2SHARE: in EUDAT, this is a simple storage mechanism to exchange research data,

\(^{14}\)http://tla.mpi.nl/projects_info/auvis/

\(^{15}\)http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eldentity.html

\(^{16}\)http://www.informatik.uni-leipzig.de/bsv/homepage/de/content/exchange-exploring-concept-change-and-transfer-antiquity

\(^{17}\)http://www.uni-stuttgart.de/ilwndl/news/forschung/news_0017.html

\(^{18}\)http://geobib.info/

\(^{19}\)http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Hauptseite

\(^{20}\)http://www.epol-projekt.de/

\(^{21}\)http://argumentum.eear.eu/

\(^{22}\)http://senereko.ceres.rub.de/

\(^{23}\)http://www.visargue.uni-konstanz.de/index.php?lang=en
providing an upload and a simple metadata input facility. CLARIN-D was consulted for the specification of user needs and to help in adding several essential metadata categories for the linguistic domain.

- **Workflow development:** Working on a Generic Execution Framework (GEF), EUDAT develops mechanisms for running analysis tools close to the data, in order to avoid submission of large data sets in webservice chains. This process should answer imminent scalability issues with big data sets in webservice based architectures. With the expertise of WEBLICHT development, CLARIN-D contributed data, model workflows and requirements for the HSS community. CLARIN-D workflows will also be used as use-cases for the EUDAT community based workflow repository.

- **B2SAFE** is the EUDAT service for safe replication of repositories. CLARIN-D collaborates closely with EUDAT in testing and using the replication of repositories to central computing centres for backup and disaster recovery. The replicas potentially be used for accessing the data (depending on the access rights) through the use of PIDs.

### 4.3 Liaison activities with other initiatives

CLARIN-D members belong to the larger German, European and international academic communities. They are routinely present and active in academic conferences, workshops, and projects on many different scales. This subsection offers only a selection of liaison activities that CLARIN-D personnel have participated in as representatives of CLARIN-D over the last year.

The CLARIN-D team at BBAW is involved in the Deutsches Textarchiv (German Text Archive); with the associated group at the University of Gießen additional contacts to the DTA group exists. The Gießen group also investigates uses of the CLARIN-D showcases for the Digital Cultural Heritage community.

Through the centres in Stuttgart and Tübingen, CLARIN-D is affiliated with the special research centres **SFB 833 Emergence of meaning – the dynamics and adaptivity of linguistic structures** and **SFB 732 Incremental Specification in Context**, providing support in metadata development; and **SFB 991 Conceptual shifts: Statistical evidence**.

The Institut für Deutsche Sprache (IDS) participates in the BMBF-funded **TextGrid** project, developing sustainable, interoperable, homogenous, open source and standards-based digital research infrastructure for the humanities – a project with goals and interests that overlap with several aspects of CLARIN-D.

CLARIN-D is involved in standardization activities for language resources, especially ISO TC 37/SC4 (via IDS and Tübingen), Text Encoding Initiative (IDS and BBAW), and the Research Data Alliance (RDA, via IDS and MPI).
5 Conclusion and further perspectives

During the past year, CLARIN-D was heavily involved in extending the project’s duration and working on liaisons with national and international partners in building and establishing the infrastructure for the Humanities and Social Sciences. The liaison covered individual projects, national initiatives and international cooperations, especially in the context of CLARIN ERIC.

In the future, using the liaison coordinator, CLARIN-D will focus its liaison activities more on eHumanities projects to reach out to the user community currently not affiliated with the CLARIN-D centres. A cooperation with work package 4 will allow the reuse of material and technology when approaching the user communities.