



D-SPIN

**D-SPIN Report R1.2a:
Progress Report of the Project and the
various Working Packages**

June 2009

D-SPIN, BMBF-FKZ: 01UG0801A

Deliverable: R1.2a: Progress Report of the project and the various working packages

Responsible: Erhard Hinrichs

© All rights reserved by the University of Tübingen on behalf of D-SPIN

Editors: Erhard Hinrichs, Lothar Lemnitzer and Kathrin Beck

Summary.....	4
Introductory Remarks.....	5
WP 1: Management and Dissemination.....	6
Cooperation agreement and R&D contracts	6
International Advisory Board	6
Wiki.....	6
Meetings.....	6
Dissemination	6
Liaison with CLARIN.....	7
Project publications.....	7
WP 2: Technical Infrastructure	9
WP 3: Relations of the Project to Research in the Humanities	10
WP 4: Liaison with other National Projects and Initiatives	11
WP 5: Language Resources and Tools.....	12
Integration of language resources	12
Integration of tools and resources via web services.....	12
WP 6: Education and Training.....	13
WP 7: Legal and Ethical Issues	14

Summary

D-SPIN is the German counterpart of the European Research Infrastructure project CLARIN (Common Language Resources and Technology Infrastructure; cf. <http://www.clarin.eu>). The ultimate objective of CLARIN is to create a European federation of digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide available language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience are researchers in the humanities and in the social sciences, and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organizational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, intellectual property rights).

Similar goals are pursued by the D-SPIN project on the national level. Work in D-SPIN is carried out in close collaboration with CLARIN. Within the CLARIN federation, the focus of D-SPIN is on German resources, tools and their integration through web services. Besides these localisation efforts, D-SPIN has a special focus in addressing potential users of the infrastructure by preparing training material and teaching activities.

This report gives an overview of the project activities in the first year, i.e. April 2008 – March 2009 and reports problems that have been encountered so far.

Introductory Remarks

1. D-SPIN is the German counterpart of the CLARIN project. We will therefore devote sufficient space to report on CLARIN activities whenever they are relevant to the D-SPIN project. On the CLARIN website, there is a list of CLARIN documents – in particular, minutes of key meetings and deliverables – which will enable readers to follow the discussions and decisions taken on the CLARIN level.
2. D-SPIN activities are supported by both the BMBF funding as well as by third party funding which has been dedicated explicitly to the CLARIN/D-SPIN activities. In this report, we will report on the D-SPIN activities without mentioning the source of funding for each activity. For details we refer to the intermediate reports (“Zwischenbericht”) of each of the partners where the sources of funding are listed.
3. Activities are listed per work package, if not mentioned otherwise.
4. The purpose of this report is to provide a short overview of the activities carried out in the D-SPIN project during the first year.
5. According to the schedule of deliverables, work packages 2, 3, 5 and 7 have produced more detailed reports about specific aspects of their work. These deliverables are entitled:
 - R2.1 Definition of service centres and selection of centres in Germany
 - R3.1 Requirements analysis
 - R5.1 Guidelines for the standard conformant adaptation and creation of language resources
 - R7.1 Inventory of legal aspects regarding the provision of language resources

WP 1: Management and Dissemination

Work package 1 is devoted to the management of the project as well as to the dissemination of the results.

Cooperation agreement and R&D contracts

In the first three months of the project, a **cooperation agreement** („Kooperationsvertrag“) was drafted, discussed and signed by all partners. The agreement can be produced on request.

Bilateral research and development contracts („FE-Vertrag“) were drafted and signed by the parties. A sample contract provided by the BMBF was adapted to the needs of the project. The model contract can be produced on request.

International Advisory Board

The International Advisory Board of the project was constituted. According to the project work plan, the mission and tasks of the board are to control and to safeguard the progress of the project, to advise the project on the basis of the delivered reports and to enable contacts and cooperations with other German projects where cooperation is advisable.

The members of the Scientific Board are:

- Helge Kahler (BMBF)
- Axel Horstmann (Volkswagen Stiftung)
- Christiane Fellbaum (Princeton University)
- Bernhard Neumair (GWDG, University of Göttingen)
- Neil Freistat (Maryland Institute for Technology in the Humanities)
- Paul Doorenbosch (Koninklijke Bibliotheek NL)
- Bente Maegaard (University of Copenhagen, CLARIN Liaison)

Wiki

A project-internal wiki has been launched at the MPI in Nijmegen. The purpose of this wiki is to organize the internal work of the project and to provide workspace for the work packages.

Meetings

The following project-internal meetings were organised:

- the kick-off meeting in Saarbrücken, 15th - 16th May 2008, plenary meeting and meeting of the executive board
- a plenary meeting in Berlin, 2nd October 2008
- a plenary meeting in Frankfurt, 12th March 2009

Dissemination

The following public relations and dissemination activities have been initiated in this work package:

- a public project website was launched (cf. <http://www.sfs.uni-tuebingen.de/dspin>)
- a flyer was produced for presentation of the project at conferences etc.
- the project was presented at a linguistic workshop in Dortmund, July 2008 and at the D-SPIN workshop “cumulative work with diachronic corpora” which took place in March 2009 at BBAW in Berlin
- D-SPIN was co-organizer of the “Text Mining and Services” Conference, March 2009 in Leipzig
- D-SPIN was introduced to the CLARIN community by an article in CLARIN newsletter #2

Liaison with CLARIN

There is a close and regular cooperation with the EU project CLARIN. Erhard Hinrichs, University of Tübingen and Peter Wittenburg, MPI Nijmegen are members of the CLARIN executive board and have participated in the meetings of the executive board.

- 17 – 20/03 2008 Nijmegen: Kick-off Meeting
- 24 – 25/04 2008 Utrecht
- 27/05 2008 Marrakesh: LREC Consortium Meeting
- 17 – 18/09 2008 Helsinki
- 03 – 05/12 2008 Budapest
- 16 – 17/03 2009 Oxford

D-SPIN is collaborating with CLARIN by holding joint workshops of mutual interest. During the first year, the following workshops were held:

- CLARIN WP2/WP5 workshop in Berlin, October 2008 („technical infrastructure, language resources and Usage Scenarios”)
- CLARIN WP5 Workshop on metadata, January 2009 in Athens
- CLARIN WP2 workshop and tutorials about details of the technical infrastructure, February 2009 in Oxford

Project publications

The following papers were presented by project partners at conferences and/or published:

Fabienne Fritzing, Max Kisselew, Ulrich Heid, Andreas Madsack, Helmut Schmid: *Werkzeuge zur Extraktion von signifikanten Wortpaaren als Web Service*. Paper presented at: *GSCL Symposium Sprachtechnologie und eHumanities*, Duisburg, 26th - 27th February 2009 (<http://cl-pc101.inf.uni-due.de/g scl2009/>)

Andreas Witt, Ulrich Heid, Felix Sasaki, Gilles Serasset: *Multilingual Language Resources and Interoperability*, will be published in: *Language Resources and Evaluation (2009)*

Helmut Schmid: *The technical details of the D-SPIN Architecture* (Internal technical report 2009)

Lothar Lemnitzer, Erhard Hinrichs, and Andreas Witt: *Language Resources, Taxonomies and Metadata*. Paper presented at: *Text Mining and Services*, Leipzig, March 2009

Volker Boehlke: *A prototype infrastructure for D-Spin-Services based on a flexible multilayer architecture*. Presentation and publication: *Text Mining and Services*, Leipzig, March 2009

Alexander Geyken and Wolfgang Klein: *Deutsche Sprachressourcen-Infrastruktur*. In: Berlin-Brandenburgische Akademie der Wissenschaften (Hg.): *Jahrbuch 2008*. Berlin: Akademie Verlag 2009.

Christina Bankhardt: *D-SPIN – Eine Infrastruktur für Deutsche Sprachressourcen*. In *Sprachreport* Heft 1/2009 (25. Jg.)

Daan Broeder, Malte Dreyer, Marc Kemps-Snijders, Andreas Witt, Marc Kupietz, and Peter Wittenburg: *Persistent and Unique Identifiers*. CLARIN-Deliverable D2.2

WP 2: Technical Infrastructure

The work of WP2 is closely related to the corresponding activities in CLARIN WP 2. This work is documented in detail in the following deliverables:

- D2R-1a Centres Network Formation
- D2R-1b Centres Network Formation - Centre Types
- D2R-2a Centres Network Formation - Centre Types
- D2R-2b Federation Foundation – Persistent and Unique Identifiers
- D2R-4 Registry Requirements – Metadata Infrastructure for LRT
- D2R-6a Web Services and Workflow Requirements

The deliverables are available on the CLARIN website and on http://www-sk.let.uu.nl/cgi/clarin_deliverables_list.ksh?due.

The following D-SPIN-related activities have been undertaken in this context:

- D-SPIN partners filled in a centre self assessment form which has been provided by MPI. The data have been analysed and after bilateral consultations between MPI and individual partners, the following (preliminary) decisions were taken. Centres of type A will be in Germany: DFKI, IDS, MPI; Centres of type B will be in Germany: BBAW, University of Leipzig, University of Tübingen. Bilateral talks have been initiated by MPI with these partners to make sure that the requirements for the types are met (see deliverable D2R-1b for further details about the centre types, their tasks and responsibilities).
- In collaboration with Max-Planck Gesellschaft and DFN Verein, a prototype LRT federation based on SAML 2.0 has been established. In a further step a prototype will be implemented, which will cover Dutch, Finnish and German partners. From the D-SPIN consortium, the BBAW, IDS and MPI will take part. Contracts with Identity Federations will be concluded so that researchers from the participating institutions can be identified federation-wide and can be assigned access rights to resources.
- Partners have registered their language resources and tools in the ad-hoc repository on the CLARIN website. This repository serves as a starting point for the further description and for the integration of (selected) language resources and tools.
- MPI provides a service for the partners who want to acquire Persistent Identifiers for their language resources (for further details cf. Deliverable D2R-2b).
- On request, MPI will help partners to install components of the technical infrastructure that is needed for the LRT federation.
- D-SPIN partners are actively involved in establishing a CLARIN Metadata infrastructure.
- D-SPIN partners are actively involved in setting a CLARIN framework for LRT web services.

WP 3: Relations of the Project to Research in the Humanities

In year one, the following tasks were undertaken:

1. Launching a questionnaire to explore the needs and requirements of humanities and of social science researchers
2. Launching case studies a) language resources for Computer Aided Language Learning, b) language resources for the study of endangered languages and c) cumulative work with diachronic corpora.

Ad 1. Due to problems in finding adequate personnel, there have been delays in the drafting of a questionnaire to be addressed to humanities and to social science researchers. The questionnaire is now ready to be distributed. The first target has been a large mailing list of the cultural and social science community („HsozKult“).

Ad 2. The case study on the use of language resources in computer-aided language learning (CALL) is currently being worked on. Three technical reports have been drafted so far, in which the state of the art is presented.

WP 4: Liaison with other National Projects and Initiatives

On the level of the CLARIN project, a deliverable has been published which summarizes the liaison activities of this project with other projects and activities. The results of these activities are relevant for the D-SPIN project (see Deliverable D5C-1).

Liaison activities that are specific to the D-SPIN project:

- Interactions with the TextGrid project (<http://www.textgrid.de>). Two partners of D-SPIN (MPI and IDS) are also partners in the TextGrid consortium.
- Interactions with the DFN-Verein, which should provide AAI services to D-SPIN centres.
- Some partners are active participants of ISO TC37/SC4 (Management of Language Resources) and of the German mirror group DIN NAT AA-G. The D-SPIN work with respect to LRT standards has been presented at committee meetings.
- A German coordination initiative on eHumanities has been launched under the auspices of the Volkswagen Stiftung. Members are Dr. Peter Wittenburg (DOBES), Prof. Fotis Jannidis (TextGrid), Prof. Erhard Hinrichs (CLARIN/D-SPIN), Prof. Gerhard Heyer (eAuqa), Dr. Heike Neuroth (Dariah), Prof. Gerhard Lauer, coordinator: Dr. Axel Horstmann (Volkswagen Stiftung). The purpose of this initiative is to coordinate the work of several ongoing projects in the area of research infrastructures and to coordinate future funding actions at the national level. A task force has been formed whose members represent the various ongoing projects.

WP 5: Language Resources and Tools

The tasks of WP5 are to identify language resources of the partners and of further interested sites in Germany and to prepare the integration of these resources and tools into the CLARIN/D-SPIN LRT federation.

The following activities have been started, in close collaboration with the CLARIN project:

Integration of language resources

- Partners have registered their resources with the resource repository on the CLARIN website.
- Some partners (e.g. BBAW, University of Frankfurt, University of Tübingen) have started to convert resources to facilitate the integration.
- Some partners (e.g. University of Tübingen, BBAW) contribute to the discussions on the emerging metadata infrastructure, which is first of all a CLARIN task. D-SPIN partners will test the metadata components and profiles that will be provided by CLARIN.

Integration of tools and resources via web services

- Four partners (BBAW, University of Leipzig, University of Stuttgart, University of Tübingen) have implemented and deployed the provided web services for a) the access to lexical resources, e.g. “GermaNet”, “Deutscher Wortschatz” and “Wörterbuch der deutschen Gegenwartssprache” and b) for the processing and linguistic annotation of text (tokenisation, part-of-speech-tagging, parsing, named entity recognition, baseform reduction, cooccurrences); parts of the resources and tools have been combined into complex workflows. Import/Export formats for the interchange of textual data between these tools have been specified. A test GUI for the combination of web services has been implemented. The goal of this exercise is to test the feasibility of such service-oriented workflows and to detect technical problems.

WP 6: Education and Training

The tasks of this work package are a) to provide teaching materials for several kinds of courses (classroom teaching as well as on-line courses), b) to offer courses to young researchers based on these teaching materials including a summer school, and c) to develop a plan for a D-SPIN help desk. Work has started earlier than planned due to the availability of highly qualified personnel at the involved partner institutions.

- In a first meeting of the three partners involved (University of Frankfurt, University of Gießen and University of Tübingen), in August 2008, the topics of the learning materials and a schedule for the development of the learning modules have been fixed.
- In order to be as flexible as possible, learning modules are being developed as SCORM packages; their integration into two learning platforms, i.e. ILIAS and MOODLE, will be tested.
- Various tools for the development of SCORM learning modules have been tested; no definite decision has been taken yet, but the partners work with some of the tools to test their suitability.
- Work on the contents of learning objects has started.
- In a second meeting of the three partners involved (University of Frankfurt, University of Gießen and University of Tübingen) the schedule for the development of the learning modules was revised and adapted.
- The learning modules will be used, tested and further developed during university courses in the summer term 2009 (University of Gießen) and winter term 2009/3020 (University of Frankfurt).

WP 7: Legal and Ethical Issues

The goal of this work package is to analyse the legal and ethical situation with respect to the compilation, the distribution and the use of language resources, to collect and to analyse samples of usage licences, and based on this analysis to give best practice recommendations to language resource providers from the legal and ethical point of view. The following activities have been initiated in the first year of the project:

- Drafting of a model licence which describes the relation between copyright holders on the one hand and language resource providers on the other hand
- Drafting of an End User Licence Agreement (EULA) for NLP tools
- Liaison in these issues with CLARIN WP7 and meetings with the WP7 managers from the University of Helsinki
- Drafting of a model contract between resource providers and end users with respect to the use of collections of citations
- Analysis of contracts for the use of commercial annotation tools and of the legal impact of distributed annotation with such tools
- Collection of typical EULAs, of licence contracts and of memoranda of understanding
- Participation in an expert workshop on “copyright and digital literature”, held at the “Zentrum für Europäische Wirtschaftsforschung” (ZEW)