Originally published in: Jozić, Željko/Kirchmeier, Sabine (Eds.):
The role of national language institutions in the digital age. Contributions to the EFNIL conference 2021 in Cavtat.
– Budapest: Nyelvtudományi Kutatóközpont, 2022. Pp. 71-80.

POSTPRINT

Andreas Witt
Paweł Kamocki

The future is now. The digital transformation in the German linguistics community and the key role of the IDS

Abstract

The Leibniz-Institute for the German Language (IDS) was established in Mannheim in 1964. Since then, it has been at the forefront of innovation in German linguistics as a hub for digital language data. This chapter presents various lessons learnt from over five decades of work by the IDS, ranging from the importance of sustainability, through its strong technical base and FAIR principles, to the IDS' role in national and international cooperation projects and its expertise on legal and ethical issues related to language resources and language technology.

1. Introduction

The Leibniz Institute for the German Language (Leibniz-Institut für Deutsche Sprache, hereinafter: IDS) is the central academic institution for the study and documentation of the contemporary usage and recent history of the German language.[1] Since its establishment in 1964, the IDS has been at the forefront of innovation in German linguistics.

This chapter describes the IDS' role as a digital hub for German language data (Section 1) and presents several "lessons learnt" from the (nearly) sixty years of its existence (Sections 2–7). These sections focus on the importance of digital language data for the IDS (Section 2), the importance of sustainability in its many dimensions (Section 3), and the role of the technological base (Section 4). The remaining sections present the efforts made at the IDS towards recognising and addressing the user's specific needs when it comes to language data and tools (Section 5), and making the data and tools findable– a task considerably facilitated by international and national cooperation projects (Section 6), where legal and ethical issues are one of the focal points (Section 7).

2. The IDS as a digital hub

The IDS was established in 1964 in the city of Mannheim. The choice of this city was not accidental, as Mannheim has long had strong links with German linguistics.

[1] Cf. https://www.ids-mannheim.de/?id=1491&L=1 (last accessed 04-05-2022).

This is where the seat of the *Bibliographisches Institut*, publisher of the Duden dictionary, was moved in 1953, together with the *Institut*'s large archive. Although, after the reunification of Germany, the main seat of the publishing house was relocated to Berlin, its Language Technology Division remained in Mannheim. In 2013, the *Bibliographisches Institut*'s archive was donated to the library of the University of Mannheim. Moreover, Mannheim is also the city where the Council of German Orthography[2] (established in 2004) is based.

The IDS was not only established in a very special place but also in a very special era. In the 1960s, Germany's state propaganda was still in living memory, and the IDS was committed from the start to strict empiricism, which was a politically innovative approach at the time. In this spirit, the IDS follows a descriptive rather than a prescriptive approach to language research. The choice of digital methods, with their cold objectivism, is one of the ways to guarantee freedom from any ideological influences.

It is therefore only natural that the IDS – home to the so-called Mannheim School of Corpus Linguistics (Teubert/Belica 2014) – has long been at the forefront of the digital transformation of language research in Germany. Shortly after its establishment, in the very early days of corpus linguistics, the IDS began collecting German texts in digital form, initially using punch cards as data carriers. The first electronic corpus of German, the *Mannheimer Korpus I* (MK I), completed in 1969, was compiled in this way; it numbered 2.2 million words in 293 texts. Another corpus, LIMAS (*Linguistik und Maschinelle Sprachbearbeitung*) was compiled between 1970 and 1971; it consisted of 500 texts divided into 33 subject areas. In 1975, this and other early IDS corpora were printed on continuous form paper; they are stored to this day in this form in the IDS archive (Fürbacher et al. 2017).

This long tradition of text corpora at the IDS led to the creation of DeReKo (*Das Deutsche Referenzkorpus*), the world's largest collection of German texts designed for language research (Kupietz et al. 2018). As of March 2022, DeReKo contains 53 billion words – and is growing at a steady pace. DeReKo is available for online querying by registered users (the registration process is free and simple) via COSMAS II (Corpus Search, Management and Analysis System)[3] and KorAP (Corpus Analysis Platform).[4] DeReKo is also subdivided into smaller sub-corpora according to various criteria, thereby catering to the users' specific needs (cf. Section 5 below). It has been an inspiration for numerous other national language reference corpora in Europe.

Not only text data but also speech data have been collected at the IDS. In 1971, the German Speech Archive (*Deutsche Spracharchiv*, DSAv, compiled since 1932)

---

2 http://www.rechtschreibrat.com (last accessed 01-07-2022).

3 https://cosmas2.ids-mannheim.de/cosmas2-web/ (last accessed 01-07-2022).

4 https://korap.ids-mannheim.de (last accessed 01-07-2022).

was transferred to the IDS. Later, it became the Archive of Spoken German (*Archiv für Gesprochenes Deutsch*),[5] which is still being added to today (Fürbacher et al. 2017).

The directors of the IDS and their affinity for digital matters played a crucial role in the institution's becoming a hub for digital language data. Before he was appointed director of the IDS (a position he occupied between 1976 and 2002), Prof. Dr. Gerhard Stickel worked as a researcher at the German Computing Centre (*Deutsche Rechenzentrum*, DRZ) in Darmstadt; he was also involved in early-stage AI research. Prof. Dr. Ludwig M. Eichinger (director of the IDS between 2002 and 2018) already used the IDS' digital data as a PhD Student. Prof. Dr. Henning Lobin (Director of the IDS since 2018) obtained his habilitation in Computational Linguistics at the University of Bielefeld in 1996, and then served as Professor of Applied and Computational Linguistics at the Justus Liebig University in Giessen for nearly two decades. This proves that since the institution's early days, the IDS' directors understood the importance of digital technology and realised its potential for language research. Some distinguished members of the IDS' Scientific Advisory Board, like Hans Uszkoreit and John Nerbonne, were also pushing the institution up the digital path.

In 2019, a dedicated Department for Digital Linguistics (*Digitale Sprachwissenschaft*) was created at the IDS, headed by one of the co-authors of this chapter. As of mid-2022, there are eighteen researchers in the department, working on the collection and curation of language data, the long-term archiving of language data, and national and international infrastructure projects as well as legal and ethical issues related to the above-mentioned domains (cf. Section 7). The establishment of the department was crucial for infrastructure projects at the IDS. The department's associates are (or were) involved in such projects as D-SPIN[6] (the predecessor of CLARIN-D,[7] the German national branch of CLARIN ERIC (see below), and later CLARIAH-DE[8]), TextGRID (a virtual research environment for the humanities optimised to work with TEI-coded resources),[9] *Verwertug Geist*[10] (exploring the potential of knowledge transfer in the humanities and related domains) and Text Transfer[11] (on the application of corpus-based methods to predict the impact of scientific texts).

---

5 https://agd.ids-mannheim.de/index.shtml (last accessed 01-07-2022).

6 https://weblicht.sfs.uni-tuebingen.de/publikationen.shtml (last accessed 06-07-2022).

7 https://www.clarin-d.net/en/ (last accessed 06-07-2022).

8 https://www.clariah.de/en/ (last accessed 06-07-2022).

9 https://textgrid.de/en/ (last accessed 06-07-2022).

10 https://www.ids-mannheim.de/fi/abgeschlosseneprojekte/verwertung-geist/ (last accessed 06-07-2022).

11 https://www.ids-mannheim.de/fi/projekte/texttransfer/ (last accessed 06-07-2022).

3. The role of sustainability

The many resources, tools and activities mentioned in the previous section could not have been developed at the IDS if the institution had not provided sufficient guarantees of sustainability.

Organizational sustainability is a pre-condition of trust. It is indeed hard to trust an organization that cannot guarantee its survival over a long period of time. This is clearly visible in the world of education, where older establishments (such as Oxford and Cambridge universities, among the first universities in the world) have an obvious reputational advantage over newly created ones, no matter how generously funded and how enthusiastically advertised they are. The fact that an establishment has been issuing internationally recognized diplomas for decades if not centuries is perceived as a guarantee that a diploma from this establishment will retain its value in the foreseeable future. The same reasoning also applies to research organisations.

According to the Practical Guide for Sustainable Research Data recently published by Science Europe (2021),[12] the sustainability of a Research Performing Organisation (RPO) is to be evaluated in the following areas: Organisational Engagement and Commitment; Policy Environment; Financial Aspects; Training; Technical Preparedness; and Communication and Awareness Raising. As a research institution with over 50 years of tradition and stable sources of funding (the German Federation and the Federal State of Baden-Wuerttemberg), the IDS has a high score in all of the above-mentioned domains.

In a narrower sense, sustainability refers specifically to the perennial archiving of research data (technical sustainability), i.e., providing guarantees that the data will be re-usable and available (preferably at their original location, even if the data themselves are marked as outdated) over long periods of time. This is achieved via the standardization of data formats, and especially via continuous conservative (or, rather, preservative) development. Both of these necessitate a strong technological base.

4. The role of a strong technological base

Another lesson learnt from the IDS' experience as a digital hub for language data concerns the importance of a strong technological base.

The implementation of the current technological base at the IDS has been described by Witt/Schonefeld (2011). The authors identify the following aspects of the technological base:

[12] https://scienceeurope.org/media/b3odxx3s/se-practical-guide-sustainable-research-data.pdf (last accessed 01-07-2022).

- *Services* provided to users are the most important part of the infrastructure; they include internet access (e.g., via Eduroam), e-mail, cloud storage, virtual workspace (e.g., an online text editor), an online library catalogue, etc. During the COVID-19 pandemic and generalized home office, it has become particularly important for services to be available not only on site, but also remotely, via virtual private networks (VPN);
- *Identity Management*: access to most services requires user authentification; it is simplified considerably if the user's personal data are managed centrally (e.g., by the HR department), and each service synchronises its access data with a central identity database. This minimizes the risk of errors due to typos, facilitates the recommended periodic changes of passwords and changes of usernames (e.g., following marriage), and enables the accounts of former employees to be deleted quickly;
- *Operating and Maintenance*: all components of the technological base (servers, workstations, internet connection, printers, etc.) should be classified according to their importance; critical components (such as the internet connection) should be backed up by redundant systems, and the whole infrastructure should be constantly monitored so that immediate action can be taken in case of failure;
- *Security*: research data, especially those held by a language research institution, may be thought of as presenting little to no interest for hackers; this, however, is not true. Many attacks are quantity-oriented and their perpetrators simply want to affect as many computers as possible, regardless of their "quality". Moreover, IT security is increasingly a legal requirement for storing and processing personal data (cf. Articles 5.1(f) and 32 of the General Data Protection Regulation) and corpora based on the Text and Data Mining exception (Article 3.2 of the 2019 Directive on Copyright in the Digital Single Market). Protection against unauthorized access is, therefore, an essential feature of a strong technological base in a research institution.

5. Transfer depends on technology

Language institutes, just like any other establishments or organisations, should never lose sight of the needs of their "customers" (no matter whether they are called "clients", "users" or "target groups", the idea remains the same).

In the case of language institutes, this is particularly difficult as the "customers" are indeed particularly difficult to define. It might be tempting to say that a language institute's work is carried out "for science", "for the greater good" and "for future generations" – all of these are true – but there are also actual people, here and now, who can benefit from the results of a language institute's work.

The first and largest group of the IDS' customers is undoubtedly other research institutions, and especially universities: the place where future teachers of national languages are educated. Public administrations also count among the IDS' "cli-

ents". On occasions, the IDS also works with the private sector, such as the publishers of dictionaries and encyclopaedias, interested in keeping their publications up to date. This pool of "clients" is expected to grow steadily, as more and more actors realize that the ability to process and analyse digital text data is an important component of 'digital literacy', a fundamental skill in the contemporary world and not just limited to the job market.[13]

Each of the above-mentioned groups has its specific needs which the IDS is trying to cater for. In particular, representatives of each of these groups expect to receive empirical data pre-processed in a specific way. Responding to this expectation requires skilful usage of the possibilities offered by digital technology.

6. The role of national and international cooperation

There is a great variety of language data and language corpora. They can be divided according to their modality (text, speech, audio-visual data), their context (e.g. parliamentary debates, poetry, everyday speech, simplified language, L2 and learner's speech), their time periods and their media (e.g., computer-mediated communication). This variety makes it complicated for users to find the exact type of resource that corresponds to their needs. In order to facilitate this task, it is crucial for the resources to be marked with appropriate metadata.

However, even a very complete metadata description of a language resource does not guarantee that it will be found and re-used. The metadata should also be 'advertised' through appropriate channels, such as catalogues provided as part of national and international cooperation projects.

National and international cooperation in the field of language resources is, to a large extent, motivated by the idea of FAIR data, i.e. making research data Findable, Accessible, Interoperable and Re-Usable (Wilkinson et al. 2016). Cooperation between institutions, within and across borders, is necessary to achieve this ideal, as it allows them to mutualize and coordinate efforts towards addressing some of the common problems, such as legal and ethical issues (cf. Section 7).

The IDS has been involved in CLARIN (Common Language Resources and Technology Infrastructure, formally established in 2012[14]) since its conception phase. CLARIN's mission is to create an online environment in which digital language resources and tools from all over Europe are accessible through a single sign-on for researchers in the humanities and social sciences (Fišer/Witt 2022).

---

[13] Cf. the podcast by Andreas Witt and Thorsten Meyer as part of the Max Planck Society's series 'Digital Qualifiziert', recorded in 2021, available at: https://soundcloud.com/max-planckgesellschaft/digital-qualifiziert-andreas-witt-thorsten-meyer (last accessed 06-07-2022).

[14] By the Commission Decision 2012/136/EU of 29 February 2012 setting up the Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium (CLARIN ERIC).

The IDS is a certified CLARIN B-Centre providing long-term storage of Germanic language resources. Apart from the storage facilities, the IDS' contribution to CLARIN involves the institution's expertise in language archives, linguistic tools, long-term preservation, multimedia and multimodal data as well as legal and ethical issues.

The IDS also plays an important role in the Text+ Consortium, whose goal it is to preserve text- and language-based research data in the long term and enable their broad use in science.[15] Formally established in 2021, Text+ has been approved as a consortium for the nationwide initiative to create a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI),[16] based on an application submitted by the applicant institution, the IDS, and the four co-applicant institutions, the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Library, Göttingen State and University Library, and the North Rhine-Westphalian Academy of Sciences, Humanities and the Arts. Apart from the five applicants, more than 25 additional participating institutions contribute their specialist expertise to the initiative, a number which is expected to grow. Erhard Hinrichs – who, alongside his full professorship for General and Computational Linguistics at the University of Tübingen, is also affiliated to the IDS – serves as the spokesperson for the Text+ consortium.

The IDS is also part of many smaller international infrastructure projects, such as DeutUng (Deutsch-ungarischer Sprachvergleich) (with the University of Szeged, Hungary)[17] and DRuKoLA (Deutsch-Rumänische korpuslinguistische Analyse) (with the University of Bucharest and the research institutes of the Romanian Academy in Bucharest and Iaşi)[18] (Kupietz et al. 2019a; Cosma et al. 2016). Both these projects are integrated in a larger EuReCo (The European Reference Corpus) (launched in 2012), which aims to virtually join various national reference corpora by using the same analysis platform, KorAP (cf. above) (Kupietz et al. 2019b; Trawiński/Kupietz 2021).

7. The importance of legal and ethical issues

Since its establishment, the IDS has handled third-party language data, especially provided by such entities as the press and book publishers (cf. Section 2). Re-use of such data for research purposes requires a careful assessment of its legal status. Thanks to the experience acquired over the decades, the IDS has become a national (and, to a certain extent, a European) centre of expertise on the many legal and ethical issues affecting language resources.

---

15 https://www.text-plus.org/en/home/ (last accessed 07-07-2022).

16 https://www.nfdi.de/?lang=en (last accessed 07-07-2022).

17 https://www.ids-mannheim.de/gra/projekte/deutung/ (last accessed 06-07-2022).

18 https://www.ids-mannheim.de/digspra/kl/projekte/drukola (last accessed 06-07-2022).

As a general rule, language data are protected by copyright, as language expressions are in fact the result of their authors' own intellectual creations.[19] Copyright protection expires 70 years after the death of the author, so in principle all born-digital language data are still in copyright. The re-use (reproduction of and communication to the public) of such data requires permission from the copyright holder (i.e., typically, the author, their descendents or the publisher, if copyright was transferred by the author), unless it is expressly allowed by a statutory exception. Such exceptions exist and they are currently expanding (e.g., new exceptions for Text and Data Mining purposes were introduced by the 2019 Directive on Copyright in the Digital Single Market) but they are accompanied by complex requirements which, according to the principle *exceptio est strictissimae interpretatonis*, always need to be interpreted narrowly. This means that before an exception can be relied on, a thorough analysis of each specific case is necessary.

When copyright exceptions are insufficient for the intended use, it is necessary to negotiate a license (Latin: permission) with the copyright holders, which also needs to be carefully drafted and interpreted. On the other hand, researchers who want to make data and content generated by them (e.g., research articles, software tools) available for re-use by granting up-front permission to every member of the public, in the spirit of the Open Access/Open Data/Open Science movements, can achieve this via proper licensing, using the so-called public licenses, such as Creative Commons (CC) or the General Public License (GPL). The use of such licenses for research results is increasingly required by research funding bodies.

Moreover, language data often contain personal data, i.e., as per the legal definition (Article 4, (1) of the GDPR), "*any information relating to an identified or identifiable natural person".* The processing of such data, even for research purposes, must abide by the strict framework of the GDPR, which affects especially speech and multimodal resources.

All these issues need to be properly addressed already in the conception phase of a language research project. For this reason, it is important to not only provide researchers with guidance and advice but also to educate them so that they are able to identify potential friction points at an early stage. Therefore, the IDS' legal experts not only provide Legal Helpdesk services for CLARIN but also created the Legal Information Platform.[20] Moreover, they have been involved in the creation of two LegalTech tools destined specifically for researchers in the data-intensive humanities and social sciences: the Public License Selector[21] (Kamocki et al. 2016) and the Consent Form Wizard.[22] Recently, the IDS has also published a set of hand-

[19] Cf. CJEU's judgement of 16 July 2009 in the case C-5/08 (Infopaq).

[20] https://www.clarin.eu/content/legal-information-platform (last accessed 01-07-2022).

[21] https://github.com/ufal/public-license-selector (last accessed 01-07-2022).

[22] https://consent.dariah.eu/ (last accessed 01-07-2022).

outs on GDPR compliance, specifically addressing issues related to language research and the archiving of language resources.[23]

Finally, ethical issues are also of growing importance for language resources and language technology. Despite a growing number of ethics-related concerns, the exact content of ethical principles governing language resources and language technology remains unclear. In order to mitigate this, the IDS researchers and authors of this chapter have proposed a tentative taxonomy of ethical issues in the sector, based on five principles: Privacy, Property, Equality, Transparency and Freedom (Kamocki/Witt 2022).

8. Conclusion

As demonstrated above, the IDS plays a key role in digital transformation in the German language community. With its stable and sustainable funding, over half a century of experience with collecting, curating and archiving language data, a rich and diversified portfolio of projects and activities, strong participation in international initiatives and, last but not least, a dedicated department of Digital Linguistics, the IDS is well equipped to assume this role for decades to come.

References

Cosma, R.,/Cristea, D./Kupietz, M./Tufiş, D./Witt, A. (2016): DRuKoLA – Towards contrastive German-Romanian research based on comparable corpora. In: *4th Workshop on Challenges in the Management of Large Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 28–32.

Fišer, D./Witt, A. (eds.) (2022): *CLARIN: The infrastructure for language resources*. (= Digital Linguistics 1). Berlin: de Gruyter.

Fürbacher, M./Varadi, T./Witt, A. (2017): Digitale Forschungsinfrastrukturen: ihre Nutzung durch die Mitglieder der Europäischen Föderation Nationaler Sprachinstitutionen. In: Dąbrowska-Burkhardt, J./Eichinger, L. M./Itakura, U. (eds.): *Deutsch: lokal – regional – global*. (= Studien zur Deutschen Sprache 77). Tübingen: Narr, 103–113.

Kamocki, P./Witt, A. (2022): Ethical issues in language resources and language technology – tentative taxonomy. In: *Proceedings of the 13th Conference on Language Resources and Evaluation* (LREC 2022), Marseille, France. Paris: European Language Resources Association (ELRA), 559–563.

Kamocki, P./Stranak, P./Sedlak, M. (2016): The public license selector: making open licensing easier. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 2533–2538.

[23] https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/10695 (last accessed 01-07-2022).

Kupietz, M./Cosma, R./Witt, A. (2019a): The DRuKoLA project. In: Cosma, R./Kupietz, M. (eds.): *On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*. Bucharest: Editura Academiei Române.

Kupietz, M./Margaretha, E./Diewald, N./Lüngen, H./Frankhauser, P. (2019b): What's new in EuReCo? Interoperability, comparable corpora, licensing. In: *Proceedings of the Workshop on Challenges in the Management of Large Corpora* (CMLC-7, 2019), Cardiff, UK. Mannheim: Institut für Deutsche Sprache, 33–39.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: new developments – new opportunities. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan. Paris: European Language Resources Association (ELRA), 4354–4360.

Teubert, W./Belica, C. (2014): Von der linguistischen Datenverarbeitung am IDS zur "Mannheimer Schule der Korpuslinguistik". In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache.

Trawiński, B./Kupietz, M. (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, H./Witt, A./Wöllstein, A. (eds.): *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. Jahrbuch des Instituts für Deutsche Sprache 2020*. Berlin/Boston, de Gruyter.

Wilkinson, M./Dumontier, M./Aalbersberg, I. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientifix Data* 3, No. 160018. https://doi.org/10.1038/sdata.2016.18.

Witt, A./Schonefeld, O. (2011): Informationsinfrastrukturen am Institut für Deutsche Sprache. In: Stickel, G./Varadi, T. (eds.): *Language, languages and new technologies: ICT in the service of languages. Contributions to the Annual Conference 2010 of EFNIL in Thessaloniki*. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 87). Frankfurt a. M. et al.: Lang, 197–211.