

# Korpora modular, verteilt, vernetzt in Text +

## Leinen, Peter

P.Leinen@dnb.de  
Deutsche Nationalbibliothek

## Trippel, Thorsten

thorsten.trippel@uni-tuebingen.de  
Leibniz-Institut für Deutsche Sprache

## Weimer, Lukas

weimer@sub.uni-goettingen.de  
Niedersächsische Staats- und Universitätsbibliothek  
Göttingen

## Witt, Andreas

witt@ids-mannheim.de  
Leibniz-Institut für Deutsche Sprache

## Einführung: Motivation

Text+ ist ein Konsortium der nationalen Forschungsdateninfrastruktur (NFDI). Seine Partner vereinigen die Expertise der bestehenden Infrastrukturverbünde CLARIN-D (Hinrichs und Trippel 2017) und DARIAH-DE (Neuroth u. a. 2016) und integrieren weitere Partner aus den Bereichen Datenzentren, Bibliotheken, Universitäten, Akademien und Rechenzentren. Durch diesen Zusammenschluss entsteht ein einzigartiger Schatz textueller Korpora, die Text+ durch sein Angebot nicht nur unter Beachtung der FAIR- (Wilkinson u. a. 2016) und CARE-Prinzipien (Carroll u. a. 2021) der Community zur Verfügung stellt, sondern Dienste und Beratung anbietet, um auf vielfältige Weise mit ihnen zu arbeiten. Dieser Beitrag wirft einen beispielhaften Blick auf vorhandene Korpora in Text+ (in Text+ vor allem innerhalb der Daten-domäne Sammlungen), veranschaulicht die Ebenen des gemeinsamen Zugriffs und zeigt Möglichkeiten zur weiteren Integration von Ressourcen auf.

Text+ kann bei der Integration von Ressourcen auf vielfältige Vorarbeiten zurückgreifen. Die Vorarbeiten beziehen sich dabei etwa auf verschiedene Infrastrukturen, darunter CLARIN-D als nationalem Zweig der europäischen Infrastruktur CLARIN (Hinrichs und Krauwer 2014) und DARIAH-DE als nationalem Zweig der europäischen Infrastruktur DARIAH (Kálmán u. a. 2019; bzw. überblicks-artig Gray 2021). Allerdings umfasst Text+ auch weitere Partner, die zuvor nicht an diesen geisteswissenschaftlichen Infrastrukturverbünden beteiligt waren, die aber unabhängig davon insbesondere große Inventare an Referenzdaten erstellt haben und weitere Erfahrungen in das Konsortium einbringen.

## Beispielhafte vorhandene Korpora

Auch wenn das NFDI-Konsortium Text+ erst am 1. Oktober 2020 formal begonnen hat, gibt es zahlreiche Vorarbeiten, die durch die Partner bereits zur Verfügung stehen und von vielen Forschenden (nach-)genutzt werden. Einige dieser Referenzdatensätze, die in Text+ zu den Sammlungen gehören, werden hier kurz vorgestellt.

### DeReKo

Das deutsche Referenzkorpus (DeReKo, siehe Kupietz u. a. 2018; 2010; Lungen 2017; Kupietz und Keibel 2009) ist eine Sammlung elektronischer deutschsprachiger Korpora. Mit über 53 Milliarden Wörter ist diese linguistische Sammlung die weltweit größte ihrer Art. DeReKo enthält belletristische und wissenschaftliche Texte, Periodika und viele weitere Textarten der Gegenwart und früheren Vergangenheit. Über die Werkzeuge COSMAS II und KorAP ist es niedrigschwellig zugänglich und für diverse linguistische Fragestellungen nutzbar.

### Zeitungskorpora der DNB

Die Deutsche Nationalbibliothek (DNB) sammelt gemäß ihres gesetzlichen Auftrags Zeitungen und Zeitschriften, die in Deutschland publiziert werden. Der umfangreiche Bestand beläuft sich aktuell auf ca. 311.000 gedruckte Titel und ca. 3,2 Millionen Ausgaben E-Paper. Insgesamt umfasst die digitale Sammlung der DNB aktuell ca. 12 Millionen Objekte.

### Das Zeitungsportal der DDB.

Über das neu geschaffene Zeitungsportal der Deutschen Digitalen Bibliothek (DDB) ist eine große Sammlung historischer Zeitungen aus den Jahren 1671–1950 zugänglich.

### ELTeC-Korpus

Das Korpus der European Literary Text Collection (ELTeC, Schöch u. a. 2021) im TextGrid Repository (TextGrid Repository 2020) ist eine Textsammlung von 2500 linguistisch annotierten Romanen in mindestens zehn Sprachen. Sie ermöglicht es, Methoden und Verfahren der Textanalyse über mehrere Nationalliteraturen hinweg zu vergleichen.

### Baumbanken

Baumbanken haben in der Sprachwissenschaft eine lange Tradition. Im Gegensatz zu anderen Arten von Korpora und Sammlungen enthalten sie geparste Texte, die

syntaktische Annotationen enthalten. Dadurch werden sie für die Forschung zur Grammatik, Typologie, Sprachtechnologie, etc. eingesetzt (siehe z. B. Kübler, McDonald und Nivre 2009; Hinrichs, Filippova und Wunsch 2005). Ein Beispiel für eine Baumbank ist die Tübinger Baumbank des Deutschen/Zeitungskorpus (TüBa-D/Z, siehe auch Telljohann, Hinrichs und Kübler 2004), die auf Artikeln der Zeitung 'die tageszeitung' (taz) basiert und in der letzten Ausgabe (Version 11) auf 3816 Artikeln und über 100.000 Sätzen mit fast 2 Millionen Token beruht (siehe <https://uni-tuebingen.de/de/134290>). Mit spezialisierten Suchprogrammen wie Tundra (Martens 2013) kann nach Wörtern und syntaktischen Strukturen gesucht werden. Tundra ist dabei nicht auf das Deutsche beschränkt. Die Spannweite reicht dabei von Werken von Thomas von Aquin (Martens und Passarotti 2014) bis zu den vielen Baumbanken der Universal Dependency Treebanks Initiative (<https://universaldependencies.org/>), die für viele Sprachen Baumbanken zur Verfügung stellt.

## Beispiele weiterer Korpora in Text+

Daneben gibt es in Text+ noch zahlreiche weitere wichtige Korpora. Dazu gehört beispielsweise das Deutsche Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften mit Texten ab ca. 1600 bis 1900. Gesprochensprachliche Korpora sind z. B. am Bayerischen Archiv für Sprachsignale an der LMU München, mit der Datenbank Gesprochenes Deutsch (DGD) am IDS Mannheim oder auch in der Digitalen Bibliothek an der SUB Göttingen vorhanden. Diese Aufzählung ist bei weitem nicht vollständig, zeigt aber auf, dass die obige, beispielhafte Auswahl von Ressourcen nur einen kleinen Teil des Portfolios abdeckt und daher Lösungen innerhalb von Text+ zwar mit Hilfe einiger Beispiele implementiert werden, aber die Lösungen immer die Erweiterungsmöglichkeit auch auf weitere Daten und Partner erlauben muss. Alle Ressourcen, Einrichtungen und Angebote können als Module von Text+ gesehen werden, die durch die Infrastruktur vernetzt werden, so dass auf sie gemeinsam zugegriffen werden kann.

## Ebenen des gemeinsamen Zugriffs

Ressourcen in Text+ liegen an unterschiedlichen Institutionen verteilt vor, teilweise sind sie abgeschlossen, teilweise werden sie noch aktiv weiterentwickelt. Da viele bestehende und im Aufbau befindliche Ressourcen Rechte Dritter berühren, z. B. über Verlagsrechte an Texten, bedeutet ein FAIRer Zugang nicht, dass der Zugang offen und frei sein kann. Vielmehr stehen viele Ressourcen unter expliziten Lizenzen, die die Nutzungsart einschränken. So kann beispielsweise auf viele digitale textuelle Daten der DNB nur innerhalb der Bibliothek und über deren Infrastruktur zugegriffen werden, um rechtliche Bestimmungen zu wahren. Für andere Sammlungen gibt es Zugangsbeschränkungen, die über ein Login den Zugang auf autorisierte Nutzende einschränkt, um Lizenzbestimmungen und -verträge einhalten zu können. Eine

Grundlage für die konsortiumsweite und communityintegrierende Arbeit an und mit den Korpora ist daher ein gemeinsamer, niederschwelliger Zugriff. In Text+ wie in verschiedenen Vorarbeiten wird dazu eine gemeinsame AAI-Lösung (Authentication and Authorization Infrastructure) verwendet, über die Nutzende die verschiedenen Ressourcen und Werkzeuge verwenden können.

Da der Zugriff auf die Daten oft eingeschränkt ist, gibt es häufig keinen Vollzugriff auf die (digitalen) Daten. Über ein föderiertes Anmeldeverfahren, das auf Shibboleth und den Diensten des Deutschen Forschungsnetzes (DFN) und deren europäischer Zusammenarbeit im Rahmen von GÉANT aufsetzt, sind viele Dienste institutionsunabhängig zugänglich, z. B. Dienste, die über die AcademicCloud unseres Partners GWDG angebunden sind. Dienste wie die Federated Content Search oder die DARIAH Collection Registry bieten bereits jetzt übergreifende Suchmöglichkeiten nicht nur in Korpora. Daten werden über verschiedene Archive bereitgestellt, darunter das DARIAH-DE Repository an der SUB Göttingen sowie die zahlreichen zertifizierten CLARIN-Zentren.

In vielen Fällen besteht ein freier Zugriff auf Informationen in Katalognachweissystemen. Diese enthalten in der Regel zumindest Informationen über die Existenz der Korpora (bzw. anderer Forschungsdaten) und andere grundlegende, beschreibende Metadaten. Im Bereich der Metadaten besteht eine Herausforderung in den unterschiedlichen verwendeten Metadatenformaten, die nicht zuletzt bezüglich ihrer integrierten semantischen Interoperabilität divergieren. Hier reicht die Spannbreite von Standardformaten (z. B. dem TEI-Header, Dublin Core, Marc21, ISO 24622-X CMDI) bis zu relativ frei definierten Beschreibungen von Ressourcen. Mit mehreren Hunderttausend Datensätzen der Text+ Partner insbesondere im Bereich der Sammlungen stellt die Vernetzung der Daten eine erhebliche Herausforderung dar. In Arbeitsgruppen von Text+, die über die verschiedenen Daten- und Aufgabendomänen hinweg gebildet wurden, werden für den Zugang Lösungen entwickelt. Die Arbeitsgruppe zur Text+ Registry arbeitet dabei an den einzusetzenden Verfahren und Schnittstellen, um ein Inventar der verschiedenartigen Ressourcen anbieten zu können. Dabei zeichnet sich ab, dass für eine solche große Datenmenge eine Listendarstellung nicht ausreichend ist und zumindest grundlegende gemeinsame Informationseinheiten in den verschiedenen Metadaten enthalten sein müssen, um eine sinnvolle Überblicksdarstellung des Inventars zu ermöglichen. Die unterschiedlichen Anforderungen an Metadaten durch Bestandsdaten, Datentypen und neuere Entwicklungen werden dabei gewahrt werden. Eine weitere Arbeitsgruppe beschäftigt sich mit Fragen des Linked Data, wobei hier zwischen Linked Data für beschreibende Metadaten und Objektdaten unterschieden wird. Durch die Verknüpfung von Metadaten mit z. B. Normdaten und dem Angebot im Rahmen von Linked Data-Initiativen wird untersucht, welche Mehrwerte für Forschungsfragen der Text+ Community geschaffen werden können. Die Darstellung von Objektdaten wie Korpora in Linked Data-Formaten sowie die Schaffung rechtskonformer Angebotsmöglichkeiten sind hier noch am Beginn der Entwicklungen.

Ein zusätzlicher, gemeinsamer Zugang zu den Daten über die Metadaten hinaus wird in Text+ mit sogenannten abgeleiteten Textformaten (Schöch u. a. 2020) weiter-

entwickelt. Abgeleitete Textformate sind dabei Informationen über die Texte, die nicht die Lizenzbedingungen verletzen, aber trotzdem jene Informationen enthalten, die für konkrete Forschungsfragen erforderlich sind. Ein Ziel besteht dabei darin, z. B. die bei Partnern von Text+ befindlichen vollständigen Sammlungen rechtskonform nutzbar zu machen, etwa die Zeitungen, die an der DNB gesammelt werden oder Korpora des IDS, die nicht im Volltext außerhalb der Lizenzbestimmungen bereitgestellt werden dürfen. Eine zusätzliche Zugangsmöglichkeit zu den zugrundeliegenden Texten unter Berücksichtigung der Rechte Dritter entwickelt Text+ auf Grundlage einer föderierten Suche über den Inhalt von Ressourcen, die Federated Content Search (FCS, Vorarbeiten siehe Stehouwer u. a. 2012). Eine Herausforderung bei der Inhaltssuche sind die unterschiedlichen Annotationssebenen, welche die teils aufwändig annotierten Korpora enthalten, seien es linguistische aber auch andere Annotationen, die z. B. in anderen TEI-Repräsentationen von Texten enthalten sind. Hier muss – ebenso wie für andere Suchen – eine gemeinsame Anfragesprache definiert werden. Da diese gemeinsame Anfragesprache zwangsläufig nur eine Teilmenge der Möglichkeiten spezialisierter Werkzeuge bieten kann, können Anfragen, die z. B. durch hochspezialisierte Forschungsfragen motiviert sind, häufig besser mit den spezialisierten Werkzeugen bearbeitet werden. Ein konkretes Beispiel sind Anfragen zu bestimmten syntaktischen Strukturen, die in Baumbanken enthalten sind, die aber auf nicht syntaktisch annotierten Daten nicht zu einem Ergebnis führen können und daher in einer föderierten Inhaltssuche nur eingeschränkt zur Verfügung stehen.

## Integration weiterer Ressourcen

Ein Anspruch, den eine nationale Forschungsdateninfrastruktur adressieren muss, ist der, dass eine Vollständigkeit kaum gewährleistet werden kann und daher eine Offenheit gegenüber neuen Daten und Datengegebenen bestehen muss. Insbesondere vor dem Hintergrund zunehmender Anforderungen an das Forschungsdatenmanagement im Rahmen des Forschungsprozesses – etwa der Anforderung, Daten, die als Grundlage von Forschungsergebnissen dienen, für mindestens 10 Jahre aufzubewahren (siehe DFG-Leitlinien zur Sicherung guter wissenschaftlicher Praxis<sup>1</sup>) – können Forschende darauf angewiesen sein, verlässliche Partner zu finden, die sie beim Datenmanagement unterstützen. Text+ adressiert diesen Bedarf auf unterschiedliche Weise: (1) durch Kooperationsprojekte zur Förderung neuer Angebote zur Integration in die Text+ Infrastruktur; (2) durch offene Schnittstellen, durch die Datenzentren ihre Angebote auch über die Infrastruktur von Text+ bereitstellen können und (3) durch Angebote von Partnern von Text+, Daten zu hosten.

Kooperationsprojekte sind Projekte, die im Rahmen von Text+ das Portfolio an Daten und Diensten für die wissenschaftliche Gemeinschaft erweitern. Jedes Jahr erfolgt in Text+ eine Ausschreibung, durch die zusätzliche Arbeiten gefördert werden können, die eine Integration im Rahmen der NFDI-Angebote von Text+ ermöglichen.

Ausgestattet mit substantiellen Mitteln können dadurch Bestandsdaten und -dienste die Infrastruktur erweitern.<sup>2</sup>

Neben den Kooperationsprojekten basiert die Offenheit von Text+ auch auf den Schnittstellen. Als ortsverteilte Infrastruktur mit einer verteilten Datenhaltung, unterschiedlichen Schwerpunkten der Beteiligten, diversen Dateiformaten und zugrundeliegenden Technologien der Archivsysteme bei den verschiedenen Partnern sind Schnittstellen zur Zusammenarbeit der Partner in gemeinsamen Angeboten unverzichtbar. Ein Beispiel dafür sind die bereits beschriebenen unterschiedlichen Korpora, die durch die föderierte Inhaltssuche über eine gemeinsame Schnittstelle zugänglich sind. Diese Schnittstellen werden in Text+ offen spezifiziert und sind damit auch für Datenzentren außerhalb des Konsortiums nutzbar. Dadurch können auch Datenzentren außerhalb von Text+ ihre Angebote über die Infrastruktur verfügbar machen und können, so sie ihre Angebote nachhaltig bereitstellen möchten, verlässlich integriert werden. Über diesen Mechanismus können sich z. B. auch Anbietende von bisher nicht einbezogenen Communities, etwas aus dem Bereich der sog. „kleinen Fächer“, über die Infrastruktur vernetzen.

Daneben übernehmen die unterschiedlichen Daten- und Kompetenzzentren von Text+ aber auch selbst Daten Dritter. Wenn Daten und Dienste zur Spezialisierung und zu den Möglichkeiten eines Partners passen, können sie dort gehostet werden. Die Prozesse – anfangen von Dateiformaten bis zu den dafür notwendigen Lizenzen und Übereinkommen – basieren auf den jeweiligen lokalen Kontexten der Partner. Als Fallback steht aber auch eine Bitstream-Archivierung von einem Partner in Text+ zur Verfügung, hier werden Daten archiviert, aber nicht notwendigerweise in weitere Infrastrukturdienste integriert, z. B. weil sie nicht die formalen Voraussetzungen erfüllen oder Daten nicht zu den bestehenden Daten- und Kompetenzzentren passen. Auch dies erlaubt es, die Community zu erweitern.

Die Erweiterungsmöglichkeiten von Text+, gerade im Bereich der Sammlungen, sind dabei sehr vielfältig und erlauben individuelle Absprachen mit Forschenden. Dabei ist es entscheidend, dass die Forschenden durch die Infrastruktur ihre Interessen wahren und die Verantwortung für und Rechte an Ressourcen nicht verlieren.

## Ausblick auf weitere Entwicklungen in Text+

Das Bestreben der Vernetzung der an Text+ beteiligten Institutionen, ihrer Daten und Dienste wird auch in weiterer Hinsicht in diversen Arbeitsgruppen von Text+ vorangetrieben. Ein Bestandteil für diese Vernetzung und Integration ist eine übergreifende Registry, über die ein Zugang zu Ressourcen aus allen Text+-Datendomänen trotz deren Unterschieden, etwa ihrer disziplinären Genese, Institution, Sprache, etc. möglich werden soll. Diese Registry, aber auch soweit möglich die Daten an sich, sollen über eine Erweiterung der Federated Content Search durchsuchbar gemacht werden. Schließlich gibt es zur Verbesserung der Vernetzung der Daten eine Arbeitsgruppe zu Linked Data (LOD) in Text+. Um den Zugriff auf all diese Dienste zu gewährleisten und

zu verbessern, wird die Text+-Homepage (<https://www.text-plus.org>) sukzessive zu einem Text+-Portal als universalen Einstiegspunkt in die Text+-Infrastruktur transformiert.

## Fußnoten

1. Siehe Leitlinie 17 in DFG 2019: Leitlinien zur Sicherung guter wissenschaftlicher Praxis, [https://www.dfg.de/download/pdf/foerderung/rechtliche\\_rahmenbedingungen/gute\\_wissenschaftliche\\_praxis/kodex\\_gwp.pdf](https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf) (korrigierte Version 1.1 vom April 2022, zuletzt aufgerufen am 2022-12-13)
2. Zum Zeitpunkt des Vortrags können zu den Kooperationsprojekten aus dem Bereich der Sammlungen konkrete Beispiele benannt werden, die derzeit noch in der Bewilligung und Kooperationsvertragsabwicklung sind.

## Bibliographie

**Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell und Shelley Stall.** 2021. „Operationalizing the CARE and FAIR Principles for Indigenous data futures“. *Scientific Data* 8 (1): 108. <https://doi.org/10.1038/s41597-021-00892-0>.

**Gray, Edward J.** 2021. „DARIAH ERIC: Empowering Arts and Humanities Research on a National and International Level“. Zenodo. <https://doi.org/10.5281/zenodo.5596905>.

**Hinrichs, Erhard, Katja Filippova, und Holger Wunsch.** 2005. „What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German“. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005)*, 77–88.

**Hinrichs, Erhard und Steven Krauer.** 2014. „The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars“. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis, 1525–31. Reykjavik, Island: ELRA.

**Hinrichs, Erhard und Thorsten Trippel.** 2017. „CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften“. *Bibliothek - Forschung und Praxis* 1 (41).

**Kálmán, Tibor, Matej Ďurčo, Frank Fischer, Nicolas Larrousse, Claudio Leone, Karlheinz Mörth und Carsten Thiel.** 2019. „A landscape of data – working with digital resources within and beyond DARIAH“. *International Journal of Digital Humanities* 1 (1): 113–31. <https://doi.org/10.1007/s42803-019-00008-6>.

**Kübler, Sandra, Ryan McDonald und Joakim Nivre.** 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Springer. <https://doi.org/10.1007/978-3-031-02131-2>.

**Kupietz, Marc, Cyril Belica, Holger Keibel und Andreas Witt.** 2010. „The German Reference Corpus DeReKo: A primordial sample for linguistic research“. In

*Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner und Daniel Tapias, 1848–54. European Language Resources Association (ELRA) 2010.

**Kupietz, Marc und Holger Keibel.** 2009. „The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research“. In *Workings Papers in Corpus-based Linguistics and Language Education*, herausgegeben von Makoto/Kawaguchi Minegishi, 3:53–59. Tokyo: Tokyo University of Foreign Studies 2009.

**Kupietz, Marc, Harald Lungen, Paweł Kamocki und Andreas Witt.** 2018. „The German Reference Corpus DeReKo: New Developments – New Opportunities“. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, u. a., 4353–60. Miyazaki: European Language Resources Association (ELRA).

**Lungen, Harald.** 2017. „DeReKo - Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim“. *Zeitschrift für germanistische Linguistik* 45 (1): 161–70.

**Martens, Scott.** 2013. „TüNDRA: A Web Application for Treebank Search and Visualization“. In *Proceedings of the 12th International Workshop on Treebanks and Linguistic Theories (TLT 2013)*. Sofia, Bulgaria.

**Martens, Scott und Marco Passarotti.** 2014. „Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus Treebank into CLARIN-D“. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis. Reykjavik, Island.

**Neuroth, Heike, Stefan Schmunk, Mirjam Blümm, Andrea Rapp, Fotis Jannidis, Dirk Wintergrün, Ulrich Schwardmann und Peter Gietz.** 2016. *DARIAH-DE – Digitalität in den Geistes- und Kulturwissenschaften am Beispiel der digitalen Forschungsinfrastruktur DARIAH-DE*. Bd. 40. Bibliothek Forschung und Praxis 2. De Gruyter.

**Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke.** 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“. *Zeitschrift für digitale Geisteswissenschaften*. [https://doi.org/10.17175/2020\\_006](https://doi.org/10.17175/2020_006).

**Schöch, Christof, Roxana Patras, Tomaz Erjavec, und Diana Santos.** 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives“. *Modern Languages Open* 1 (25): 1–19. <https://doi.org/10.3828/mlo.v0i0.364>.

**Stehouwer, Herman, Matej Ďurčo, Eric Auer und Daan Broeder.** 2012. „Federated Search: Towards a Common Search Infrastructure“. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, herausgegeben von Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk und Stelios Piperidis. Istanbul, Türkei.

**Telljohann, Heike, Erhard Hinrichs und Sandra Kübler.** 2004. „The Tüba-D/Z Treebank: Annotating German

with a Context-Free Backbone". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2004/pdf/135.pdf>.

**Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a.** 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship“. *Scientific Data* 3 (März): 160018. <https://doi.org/10.1038/sdata.2016.18>.