Inhaltserschließung im Zeitalter von Suchmaschinen und Volltextsuche

b.i.t.online-Gespräch mit Elisabeth Mödden. Die studierte Bauingenieurin leitet bei der Deutschen Nationalbibliothek seit 2014 standortübergreifend das Referat Automatische Erschließungsverfahren, Netzpublikationen.

Nachdem die Deutsche Nationalbibliothek kürzlich ihr Konzept zur Inhaltserschließung vorgestellt hat, sind in der Fachcommunity einige Diskussionen entstanden. Noch sind viele Fragen zum Thema maschinelle Inhaltserschließung offen. Allem voran die Frage: Brauchen wir angesichts von Suchmaschinentechnologie und der Möglichkeiten von Volltextsuchen überhaupt noch Inhaltserschließung?

Elisabeth Mödden Gut, dass Sie das ansprechen. Ja, Inhaltserschließung hat nach wie vor ihre Berechtigung. Suchmaschinen mit Volltextsuche bzw. Stichwortsuche alleine reichen nicht aus, um umfangreiche Datenbestände so zu strukturieren, dass die inhaltlich relevanten Bücher und Artikel bei einer thematischen Suche auch tatsächlich gefunden werden. Ohne inhaltserschließende Metadaten führt eine Recherche nur zu den Publikationen, die den Suchterm wörtlich im Volltext oder im Titel enthalten. Die Nutzerinnen und Nutzer werden also nicht fündig, wenn der Sachverhalt umschrieben ist oder wenn Synonyme verwendet wurden. Ein weiteres Problem für die richtige thematische Zuordnung sind Homonyme, also gleichlautende Begriffe für ganz unterschiedliche Sachverhalte. Sie führen die Informationssuchenden auch zu sachfremden Themen, sofern keine Beziehung zum richtigen semantischen Kontext hergestellt wurde. Genau diese semantische Kontextualisierung leisten wir mit der verbalen Inhaltserschließung, indem wir die Publikationen mit den Schlagwörtern in der GND – also der Gemeinsamen Normdatei – verknüpfen.

Diesen Herausforderungen müssen sich natürlich auch die großen Suchmaschinen wie Google, Bing und Yahoo stellen. Sie haben dafür das Auszeichnungssystem schema.org entwickelt. Es dient der Kennzeichnung und Strukturierung von Inhalten auf Webseiten. Auch die Vergabe von Hashtags, wie sie in den sozialen Medien üblich ist, ist eine Form der Inhaltserschließung. Die Frage ist also nicht, ob wir Inhaltserschließung brauchen, sondern wie wir dabei vorgehen.

Die Deutsche Nationalbibliothek hat bisher gedruckte Bücher und vergleichbare Medien intellektuell erschlossen und nur die Netzpublikationen maschinell. Warum wollen Sie das jetzt ändern?

Elisabeth Mödden Die Deutsche Nationalbibliothek möchte gerne einen möglichst großen Anteil ihres Bestandes inhaltlich verbal erschließen, und zwar möglichst einheitlich unabhängig von der Medienform. Bisher konnte das so nicht geleistet werden. So werden Publikationen der Bibliografiereihe A – also Monografien und Periodika des Verlagsbuchhandels – mit Schlagwörtern aus dem Vokabular der GND erschlossen. Publikationen, die außerhalb des Verlagsbuchhandels erscheinen (Reihe B), und die Hochschulschriften in der Reihe H erhalten derzeit keine verbale Erschließung. Für Netzpublikationen nutzt die DNB die Metadaten der Ablieferer und setzt seit 2012 zunehmend statistische und computerlinguistische Erschließungsverfahren ein. Der Zugang an Netzpublikationen wächst dynamisch an, so hat er sich 2016 im Vergleich zum Vorjahr auf 1,3 Millionen sogar verdoppelt. Eine intellektuelle Erschließung dieser Mengen wäre nicht zu leisten.

Weil wir die Vorteile einer einheitlichen Erschließung für die Recherche in unseren Beständen sehen, betrachten wir die Ausweitung der maschinellen Prozesse als einzig gangbaren Weg, um möglichst vielen Publikationen möglichst gute Inhaltserschließungsdaten mitzugeben. Die DNB führt neben der verbalen auch eine klassifikatorische Erschließung mit der Dewey Dezimalklassifikation (DDC) durch. So werden alle Medienwerke in Sachgruppen eingeordnet, die meist auf den obersten 100 Klassen der DDC basieren.[1]

Die Vereinheitlichung und Vervollständigung der inhaltlichen Erschließung haben wir übrigens in unserem *Strategischen Kompass Deutsche Nationalbibliothek 2025* und unseren *Strategischen Prioritäten 2017 – 2020* als wichtiges Handlungsfeld benannt. Wir wol-

1 http://www.dnb.de/DE/Erwerbung/Inhaltserschliessung/inhaltserschliessung.html

Elisabeth Mödden studierte Bauingenieurwesen an der Technischen Universität Braunschweig und absolvierte an der Universitäts- und Landesbibliothek Darmstadt das Bibliotheksreferendariat. Seit 2007 arbeitet sie an der Deutschen Nationalbibliothek, zunächst als Fachreferentin für Informatik und Technik, seit 2014 leitet sie das standortübergreifende Referat Automatische Erschließungsverfahren, Netzpublikationen.
Kontakt: e.moedden@dnb.de

len auch weitere Vorteile maschineller Prozesse konsequent nutzen. Dazu gehört zum Beispiel die bisher nicht vorhandene Möglichkeit, auch Zeitschriftenartikel mit inhaltsbeschreibenden Metadaten anzureichern, wenn diese in digitaler Form vorliegen.

Stichwort „gute Inhaltserschließungsdaten". Was bedeutet das für die Deutsche Nationalbibliothek?

Elisabeth Mödden „Gut" bedeutet zielführend bei Recherche und Retrieval. Die inhaltserschließenden Metadaten sollen ja den Zweck erfüllen, die relevanten Publikationen im Bestand zu finden. Das ist die Grundlage für unseren Qualitätsmaßstab. Wir haben ein Qualitätsmanagement eingerichtet, mit dem die Fehlerquote der maschinellen Verfahren und ihre Auswirkungen auf den Datenbestand kritisch beobachtet wird und bei Bedarf nachgesteuert werden kann.

Wie überprüfen Sie die Qualität der maschinell erschlossenen Daten?

Elisabeth Mödden Wir kontrollieren täglich die technischen Abläufe und führen fortlaufend Stichproben zur fachlichen Überprüfung durch. Dabei werden die maschinell erschlossenen Publikationen aus den Stichproben durch die Fachreferentinnen und Fachreferenten noch einmal intellektuell klassifiziert und beschlagwortet. Alle Erschließungsdaten werden im bibliografischen Datensatz mit Angaben zur Herkunft, also einer Kennzeichnung, ob sie intellektuell oder maschinell erstellt wurden, versehen.
Die Qualität der maschinellen Sachgruppenvergabe kann so durch einen statistischen Abgleich der maschinellen und intellektuellen Einordnungen kontinuierlich überwacht werden. In den letzten fünf Jahren haben wir etwa 18 % der maschinell vergebenen Sachgruppen der Reihe O – also der Netzpublikationen – betrachtet. Im Durchschnitt stimmten in 76 % der Vergleichsfälle die maschinell und die intellektuell vergebenen Sachgruppen überein. Der Wert variiert in den verschiedenen Fächern: In Fächern mit großem Publikationsaufkommen wie dem Recht liegt er bei 92 %, in der Medizin bei 87 %. Bei Klassen mit geringem Publikationsaufkommen funktioniert die maschinelle Einordnung allerdings noch nicht so gut, wie wir uns das wünschen, weil die Trainingsbeispiele für die Lernprozesse nicht ausreichen.
Bei der Überprüfung der Schlagwörter gehen wir anders vor, weil kein direkter Vergleich möglich ist. Dort führen wir eine differenzierte Einzelbetrachtung durch, ob ein maschinell vergebenes Schlagwort für das Retrieval nützlich ist, oder ob der Sucheinstieg zu einem falschen Ergebnis führen würde. Auch diese Bewertungsergebnisse werden regelmäßig statistisch ausgewertet. Für den Jahrgang 2016 der Reihe O haben die Qualitätskontrollen zu dem Ergebnis geführt, dass etwa 27 % der Schlagwörter in die Bewertungskategorie „sehr nützlich" eingeordnet wurden, 18 % als „nützlich" und 33 % als „wenig nützlich" bewertet wurden. Etwa 22 % der maschinell vergebenen Schlagwörter sind nicht zutreffend.
Unbefriedigende Ergebnisse werden übrigens insbesondere immer dann erzielt, wenn die aktuelle Terminologie eines Fachgebiets noch nicht ausreichend in die GND eingearbeitet ist. Deshalb betrachten wir die systematische Pflege und die Aktualität der GND als wichtige Ansatzpunkte für die Verbesserung der Ergebnisse bei der Beschlagwortung. Wir wollen zusätzliche maschinelle Verfahren entwickeln, um die noch nicht in der GND enthaltenen, aber inhaltlich relevanten Schlagwörter zu erkennen und sie in ein Redaktionssystem einfließen zu lassen. Die GND-Redakteure sollen auf diese Weise fundierte Vorschläge für die weitere Pflege der GND erhalten.

Wie gehen Sie vor, um die Fehlerquote zu senken und die Qualität zu verbessern?

Elisabeth Mödden Zur Verringerung der Fehlerquote setzen wir auf ein ganzes Bündel verschiedener Maßnahmen, die sich ergänzen sollen, zum Beispiel die Weiterentwicklung der Softwareverfahren und die Optimierung der Erschließungsprozesse, sowie eben auch die Neuausrichtung der GND-Pflege und die Fortentwicklung des Qualitätsmanagements. Auch die engere Verzahnung der intellektuellen und maschinellen Erschließung ist ein Thema.
Die große Vielfalt der Medienwerke, die wir als Deutsche Nationalbibliothek sammeln, stellt hohe Anfor-

derungen an die Erschließungsalgorithmen und Prozessabläufe. Die Erschließung auf Artikelebene erfordert beispielsweise eine andere Herangehensweise als die Erschließung von E-Books oder von gedruckten Publikationen. Auch durch die sprachliche Heterogenität der Sammlung ergeben sich große Herausforderungen. Sprachliche Muster und Begriffe unterscheiden sich nicht nur von Fachgebiet zu Fachgebiet, sondern oft auch innerhalb einer Fachsprache. Hinzu kommen die Unterschiede zwischen Wissenschaftssprache und Allgemeinsprache. Mit diesen Phänomenen müssen die Erschließungsalgorithmen zurechtkommen. Wir arbeiten deshalb weiter daran, die Fähigkeiten der Erschließungssoftware zu verbessern.
Last but not least ist ein Monitoring durch die Mitarbeiterinnen und Mitarbeiter der Deutschen Nationalbibliothek auf der Basis von technisch gut unterstützten Erschließungs-und Qualitätsmanagementprozessen unerlässlich. Unser Ziel ist eine hohe Verlässlichkeit der Erschließungsdaten, unabhängig davon, ob sie intellektuell oder maschinell erzeugt wurden. Aufgrund der Analyseergebnisse und unserer strategischen Leitlinien entscheiden wir letztlich, welche Publikationsgruppen maschinell erschlossen werden können und welche weiterhin intellektuell bearbeitet werden müssen.

Was genau hat die Deutsche Nationalbibliothek aktuell bei der Erschließung geändert?

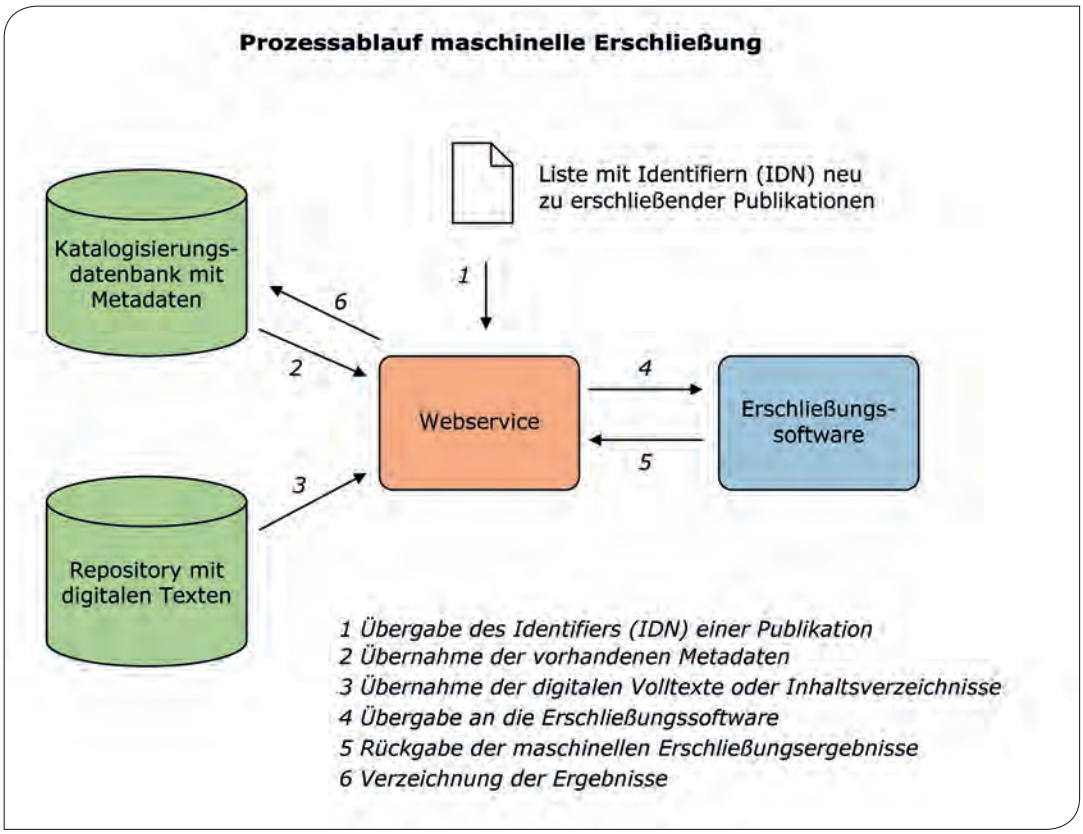
Elisabeth Mödden Dazu muss ich weiter ausholen: Die Deutsche Nationalbibliografie ist thematisch nach 104 Sachgruppen gegliedert, die auf der DDC basieren. Seit etwa 10 Jahren werden die Reihen A, B und H zudem mit vollständigen DDC-Notationen erschlossen. Nur Publikationen der Reihe A, also Verlagspublikationen, werden intellektuell auch mit Schlagwörtern erschlossen.
Seit September 2017 setzen wir die Analysesoftware auch für gedruckte Hochschulschriften (Reihe H) ein, sowie für Publikationen, die außerhalb des Verlagsbuchhandels erscheinen (Reihe B). Die maschinelle Erschließung wurde damit erstmalig auf physische Medien ausgeweitet. Für die Reihen B und H verzichten wir fortan auf die Tiefenerschließung mit der DDC und arbeiten an einem Klassifikationsschema mit verkürzten Notationen für alle Fächer, um diese Form der Erschließung künftig auch maschinell durchführen zu können. Für die medizinischen Dissertationen wurde bereits Ende 2005 ein Schema mit 140 Kurznotationen eingeführt. Entsprechende Systematiken für die Informatik, Chemie und Sozialwissenschaften werden zurzeit erprobt. Und, ganz wichtig, weil es hier offenbar zu Missverständnissen gekommen ist: Die Publikationen des Verlagsbuchhandels (Reihe A) werden mit Ausnahme der Belletristik und der Kinder- und Jugendliteratur weiterhin intellektuell bearbeitet.

Wie ist das machbar? Auf welcher Grundlage können Sie gedruckte Publikationen maschinell erschließen?

Elisabeth Mödden Nun, bei physischen Publikationen sollen künftig alle digital vorhandenen Informationen für die maschinelle Analyse herangezogen werden, also zum Beispiel parallele Online-Ausgaben, digitalisierte Inhaltsverzeichnisse oder Abstracts, Klappen- und Umschlagstexte. Zurzeit erfolgt die maschinelle Erschließung auf der Basis der Inhaltsverzeichnisse und bibliografischen Angaben zum Titel. Das bedeutet, dass geringere Textmengen für die Analysen zur Verfügung stehen. Leider ist auch der Informationsgehalt der Inhaltsverzeichnisse fallweise nicht ausreichend, um mit statistischen und computerlinguistischen Methoden zu einem guten Ergebnis zu gelangen. Das wird zu Recht eingewendet. Wir haben durchaus einen kritischen Blick auf die Prozesse. Die maschinell vergebenen Sachgruppen werden zurzeit durchgängig intellektuell überprüft.

Welche Methoden wenden Sie für diese Art der maschinellen Erschließung an?

Elisabeth Mödden Für die Vergabe von DDC-Sachgruppen und DDC-Kurznotationen verwenden wir ein lernendes Klassifikationsverfahren. Ausgangsbasis sind Trainingskorpora mit möglichst zahlreichen, intellektuell bearbeiteten Erschließungsbeispielen für jede Klasse. In der Lernphase erstellt die Software anhand der sprachlichen Merkmale der Referenzbeispiele ein Modell für jede Klasse. Im Produktionsbetrieb werden neu eintreffende Publikationen in die Klassen eingeordnet, indem ihre sprachlichen Merkmale mit den erlernten Mustern verglichen werden. Die Klassen mit dem höchsten Maß der Übereinstimmung werden als Sachgruppen und Kurznotationen zugeordnet. Damit erreichen wir eine thematische Strukturierung unseres Bestandes.
Die maschinelle Schlagwortvergabe hingegen basiert auf einem computerlinguistischen Verfahren mit einem Wörterbuch als Kernbestandteil. Das Wörterbuch für die Erschließung deutschsprachiger Texte erhalten wir dadurch, dass wir die Sachbegriffe, Personen, Geografika, Körperschaften, Kongresse und Werke als semantische Konzepte aus der GND in die Erschließungssoftware übernehmen. Bei der Erschließung wird eine komplexe linguistische Analyse durchgeführt, um die inhaltstragenden Begriffe der Publikation zu ermitteln und sie mit dem Wörterbuch in Be-



ziehung zu setzen. Dabei müssen auch die eingangs schon erwähnten mehrdeutigen Begriffe in den richtigen Bedeutungszusammenhang eingeordnet werden. Das heißt, bei gleichlautenden Begriffen mit verschiedenen Bedeutungen müssen die richtigen Anknüpfungspunkte im Wörterbuch ausgewählt werden. Das ist ein sehr komplexer Vorgang, der viele Stufen durchläuft. Als Analyseergebnis werden schließlich bis zu sieben Schlagwörter ausgewählt. Konkret bedeutet das eine Verknüpfung der Publikation mit den entsprechenden Normdatensätzen, also eine Vernetzung mit den semantischen Konzepten in der GND. Somit können dann auch die kooperativ gepflegten und vielfältig vernetzten Normdaten als Einstieg in eine thematische Suche verwendet werden. Das betrachten wir als einen bedeutenden Mehrwert.

Wie kann man sich den Erschließungsprozess ganz konkret vorstellen?

Elisabeth Mödden Die maschinelle Erschließung startet täglich automatisch zu einer festgelegten Zeit, selektiert neue Publikationen und übergibt die Liste an einen Webservice. Dieser holt die Volltexte oder Inhaltsverzeichnisse aus dem Repository und die Metadaten aus der bibliografischen Datenbank, wandelt die Daten in einfache Textdateien um, bestimmt die vorwiegende Sprache und übergibt alle Informationen an die Erschließungssoftware. Die zurückgelieferten Analyseergebnisse werden im bibliografischen Datensatz des Titels verzeichnet, und Auffälligkeiten im Verarbeitungsprozess werden in Systemdateien protokolliert.

Zur Verarbeitung der verschiedenen Publikationsgruppen existieren speziell angepasste Konfigurationen. Dabei handelt es sich um Softwareeinstellungen, die zuvor in systematischen Testreihen optimiert wurden. So wird bei der Sachgruppenvergabe beispielsweise das zu verwendende Klassifikationsmodell definiert. Auf diese Weise können über die Konfigurationen spezifische Erschließungsalgorithmen angesteuert werden, um monografische Netzpublikationen anders zu prozessieren als Zeitschriftenartikel, deutschsprachige Texte anders als englischsprachige, Volltexte anders als Inhaltsverzeichnisse.

Diese Verfahren können also auch bei fremdsprachigen Publikationen angewendet werden?

Elisabeth Mödden Zumindest für den hohen Anteil englischsprachiger Publikationen in der Sammlung der DNB arbeiten wir gerade daran, das Verfahren der Schlagwortvergabe zu erweitern und auch die Library of Congress Subject Headings (LCSH) als Terminologie einzubinden. Darüber hinaus wollen wir eine Vernetzung mit anderen Datenressourcen wie DBpedia oder YAGO testen. Crosskonkordanzen, zum Beispiel zwischen den LCSH und der GND, bieten zudem die Möglichkeit, bei Bedarf auch mehrsprachige Sucheinstiege zu generieren.

Ist denn im Katalog zu erkennen, ob ein Mensch oder die Maschine das Werk erschlossen hat?

Elisabeth Mödden Ja, natürlich. Wir haben unsere Datenstrukturen entsprechend angepasst, um die Herkunft und die Vertrauenswürdigkeit der bibliografischen Daten zu dokumentieren. Das Erschließungsergebnis maschineller Prozesse wird im Datensatz jeweils zusammen mit dem Tagesdatum, einer Kennung

zur Identifizierung der verwendeten Konfiguration sowie einem Schätzwert zur Informationsgüte, dem sogenannten Konfidenzwert, verzeichnet. Und bei der Anzeige der DDC-Kurznotationen und Schlagwörter im DNB-Portal wird diese Information mit angezeigt. Durch Anpassung des Datenaustauschformats MARC 21 haben wir zudem erreicht, dass die Informationen zur Datenherkunft auch an die Datenbezieher ausgeliefert werden können.

Wie sorgen Sie dafür, dass die Verfahren stets aktuell sind?

Elisabeth Mödden Die Entwicklungen sind tatsächlich sehr dynamisch. So führt die Pflege und Weiterentwicklung der Softwarealgorithmen, der Trainingskorpora und der Wörterbücher dazu, dass wir regelmäßig neue Konfigurationen erstellen und nach einer Testphase in den Produktivbetrieb überführen. Ich möchte ein Beispiel nennen: Die intellektuell bearbeiteten Publikationen eines Jahres nutzen wir, um die Lernprozesse der Klassifikation mit neuen Trainingsbeispielen anzureichern. Ebenso wollen wir erreichen, dass die laufenden Veränderungen in der GND zeitnah in die Wörterbücher der Erschließungssoftware einfließen.

Die Deutsche Nationalbibliothek hat an anderer Stelle ein Umdenken gefordert. Erschließung soll demnach nicht mehr abgeschlossen, sondern eher als zyklischer Prozess begriffen werden. Können Sie das erläutern?

Elisabeth Mödden Bei maßgeblichen Fortschritten stellt sich jeweils die Frage, ob die Erschließungsprozesse retrospektiv noch einmal wiederholt werden sollten, um die Qualität der Metadaten sukzessive zu verbessern. Solche Wiederholungsläufe wurden für die Beschlagwortung bisher schon jährlich durchgeführt, indem die bereits erschlossenen Bibliografie-Jahrgänge erneut prozessiert und zudem auch Jahrgänge mit einbezogen wurden, die zuvor noch gar nicht berücksichtigt werden konnten. Die so aktualisierten Daten stellen wir selbstverständlich über unsere Datendienste zur Verfügung, so dass alle Datenbezieher sie auch nutzen können.

Vorhin haben Sie von einer Verzahnung der Prozesse gesprochen. Was meinen Sie damit?

Elisabeth Mödden Unser Ziel ist es, eine integrierte Erschließungsumgebung für die intellektuelle und maschinelle Erschließung zu entwickeln, um die Mitarbeiterinnen und Mitarbeiter auf effiziente Art und Weise bei der Erschließungsarbeit, bei der Qualitätssicherung und bei der Pflege der Normdaten zu unterstützen. Diese Aufgaben sind eng miteinander verbunden und erfordern eine adäquate technische Infrastruktur. Dazu wollen wir alle Komponenten und Dienste betrachten, die mit der Erschließung und Metadatenverwaltung in Zusammenhang stehen. So denken wir beispielsweise auch an Assistenzfunktionen für die intellektuelle Erschließungsarbeit, die Erschließungsdaten aus anderen Datenquellen oder maschinell erzeugte Erschließungsergebnisse als Vorschläge bereitstellen. Insbesondere aber gehören das Qualitätsmanagement und leistungsfähige Instrumente zur Pflege und Verwaltung der GND in den Katalog der Themen, mit denen wir uns intensiv beschäftigen.

Frau Mödden, vielen Dank für das Gespräch.