Datenübernahmerichtlinien des Instituts für Deutsche Sprache

Version 1 vom Januar 2018

1. Übernahmerichtline des IDS für Mündliche Korpora

Thomas Schmidt, 06.10.2015

Das IDS ist an der Übernahme mündlicher Korpora ins Archiv für Gesprochenes Deutsch (AGD[1]) und das Langzeitarchiv (LZA [2]) für germanistische Primärdaten interessiert. Für die Übernahme mündlicher Korpora gelten folgende Richtlinien:

Welche Daten kommen für eine Übernahme in Frage?

(1) Unter einem mündlichen Korpus verstehen wir systematische Sammlungen von Audio- und/oder Videoaufnahmen sprachlicher Interaktion und zugehörige Dokumentationen, Transkriptionen und Annotationen, sowie ggf. Zusatzmaterialien, die zur Analyse der aufgezeichneten Interaktionen notwendig sein können.

(2) In aller Regel kommen nur solche Daten für eine Übernahme in Frage, zu denen die originalen Audio- und/oder Videoaufnahmen vorliegen. Reine Transkriptkorpora werden nur in begründeten Ausnahmefällen für eine Übernahme in Betracht gezogen.

(3) Wenn ein Korpus (auch) aus anderen Formen der Aufzeichnung von Aspekten sprachlicher Interaktion – wie etwa Eye-Tracking-Daten, Motion-Sensor-Daten etc. – besteht, bedürfen diese einer gesonderten Betrachtung.

(4) Der Fokus liegt auf Daten des gesprochenen Deutsch in all seinen Varietäten inklusive deutschbasierter Kontaktvarietäten. Daten in anderen Sprachen kommen nur dann für eine Übernahme in Frage, wenn sie integraler Bestandteil der aufgezeichneten Sprechsituationen (z.B. in mehrsprachigen Gesprächen) oder des Korpusdesigns (z.B. bei einem vergleichbaren Korpus oder bei einer Studie bilingualer Sprecher) sind. Korpora, die keinen oder nur einen geringen Anteil gesprochenen Deutschs enthalten, kommen für eine Übernahme nicht in Frage.

(5) Der Fokus liegt des Weiteren auf authentischen und spontansprachlichen Daten. Elizitierte Daten kommen für eine Übernahme in Frage, wenn sie neben authentisch-spontansprachlichen Daten integraler Bestandteil des Korpusdesigns sind. Korpora die ausschließlich oder überwiegend aus elizitierten Daten bestehen, kommen für eine Übernahme nur dann in Frage, wenn die Elizitationsmethode auf quasi-spontansprachliche Daten abzielt (z.B. bei Interviews, Nacherzählungen, Bildbeschreibungen, Maptasks), nicht aber, wenn sie auf (schrift-)sprachlichen Vorlagen basiert (z.B. Wortlisten, Lesetexte, Rezitationen, Aufführungen).

(6) Für konzeptionell mündliche, aber medial schriftliche Daten, insbesondere aus internetbasierter Kommunikation (z.B. Chat, Instant Messaging), verweisen wir auf die Übernahmerichtlinie des IDS für Schriftliche Korpora. Gleiches gilt für schriftsprachliche Vorlagen mündlich vorgetragener Texte (z.B. Lesetexte, Rezitationen, Aufführungen, Redemanuskripte, Drehbücher) sowie Protokolle von Interviews, Rundfunksendungen,

Reden, Vorträgen oder Debatten, zu denen keine zugehörigen Audio- oder Videoaufnahmen vorliegen.

(7) Bei mündlichen Daten, die aus einem der o.g. Gründe für eine Übernahme ans AGD/LZA nicht in Frage kommen – etwa mündliche Korpora in anderen Sprachen als Deutsch, phonetische oder sprachtechnologische („Speech“) Korpora, multimodale Messdaten – kann das IDS bei der Auswahl eines geeigneten Daten-Zentrums behilflich sein. Wir verweisen hier insbesondere auf die CLARIN-Zentren am Hamburger Zentrum für Sprachkorpora[3] und am Bayerischen Archiv für Sprachsignale[4].

Mit welchem Ziel erfolgt eine Übernahme?

(8) Das IDS übernimmt mündliche Korpora mit dem Ziel, sie dauerhaft zu bewahren, nutzbar zu halten und unter Beachtung rechtlicher Vorgaben der wissenschaftlichen Community zur Verfügung zu stellen.

Wie wird über eine Übernahme entschieden?

(9) Die Entscheidung zur Übernahme eines mündlichen Korpus erfolgt auf der Basis einer Kosten-Nutzen-Analyse unter Berücksichtigung der im Programmbereich Mündliche Korpora vorhandenen personellen Kapazitäten.

(10) Die Kosten-Nutzen-Analyse wird von Mitarbeitern des AGD in Zusammenarbeit mit dem potentiellen Datengeber anhand des „Leitfadens zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache“ [9] durchgeführt. Demnach sind eine vollständige Inventarisierung des Korpus, eine ausreichend detaillierte Dokumentation des Korpusdesigns und der Korpusbestandteile, sowie die verbindliche Klärung rechtlicher Bedingungen der Datenweitergabe Mindestanforderungen für die Übernahme eines Korpus. Weitere Orientierungshilfen für die Kriterien zur Beurteilung eines mündlichen Korpus finden sich in den DFG-Handreichungen „Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora“ [10] und „Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora“ [11].

(11) Mit Blick auf die begrenzten personellen Kapazitäten von AGD und LZA sind Beiträge des Datengebers zur Unterstützung der Korpusaufbereitung ausdrücklich erwünscht.

(12) Fällt die Entscheidung zur Übernahme positiv aus, wird zwischen dem Datengeber und der Leitung des IDS eine schriftliche Vereinbarung über die wesentlichen Punkte der Datenübernahme getroffen.

Was geschieht mit übernommenen Daten?

(13) Übernommene Korpora werden nach den Standards des AGD aufbereitet. Je nach Beschaffenheit der Ausgangsdaten beinhaltet dies eine Digitalisierung analoger Daten, die Überführung unstrukturierter Daten in strukturierte Form, eine systematische Dokumentation von Metadaten, ein Text-Ton-Alignment von Transkript und Audio/Video und/oder die Anreicherung von Daten mit zusätzlichen Annotationen.

(14) Nach Abschluss der Aufbereitung werden die Daten der wissenschaftlichen Community über geeignete, vom IDS bereitgestellte Dienste zur Verfügung gestellt. Derzeit beinhaltet dies insbesondere eine Integration der Daten in die Datenbank für Gesprochenes Deutsch (DGD[7]) und/oder in den persönlichen Service des AGD[8] sowie die Einspeisung der Daten ins IDS-Repositorium[5], welches Metadaten zum Korpus für geeignete Kataloge digitaler Infrastrukturen (derzeit insb. CLARIN VLO[6]) zur Verfügung stellt. Künftig mögen weitere, hier nicht genannte, Mechanismen der Datenhaltung und -weitergabe zur Anwendung kommen. Die Weitergabe der Daten erfolgt unter Beachtung rechtlicher Vorgaben.

(15) Die Daten werden im Rahmen der technischen und personellen Möglichkeiten des IDS gesichert und gepflegt, d.h. an geeignete Backupmechanismen angeschlossen bzw. an zukünftig geltende Standards des AGD, des LZAs oder digitaler Infrastrukturen angepasst.

(16) Etwaige analoge Materialien (insb. Ton- und Videobänder) werden im AGD inventarisiert und aufbewahrt, sofern eine Aufbewahrung aus Sicht des AGD notwendig erscheint. Erscheint dies nicht notwendig, werden analoge Materialien nach Abschluss der Aufbereitung entweder an den Datengeber zurückgegeben oder vernichtet. Aufbewahrte analoge Materialien werden in aller Regel nicht an Personen außerhalb des AGD weitergegeben.

(17) Zu den Daten gehörende Software-Umgebungen (z.B. Web-Portale, Suchmaschinen) können in aller Regel bei der Datenübernahme nicht berücksichtigt werden, d.h. sie können im AGD oder im Rahmen der LZA nicht dauerhaft gepflegt werden.

Wo gibt es weitere Informationen?

Wenn Sie ein mündliches Korpus haben, das für eine Übernahme in Frage kommt, kontaktieren Sie uns unter (0621) 1581-313 oder agd@ids-mannheim.de. Den Leitfaden [9] sowie die DFG-Handreichungen [10,11] finden Sie unter den unten angegebenen URLs.

[1] http://agd.ids-mannheim.de
[2] https://repos.ids-mannheim.de/
[3] https://corpora.uni-hamburg.de/
[4] http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html
[5] https://repos.ids-mannheim.de/
[6] https://vlo.clarin.eu/
[7] http://dgd.ids-mannheim.de
[8] http://agd.ids-mannheim.de/korpora.shtml
[9] http://ids-pub.bsz-bw.de/frontdoor/index/index/docId/1331
[10] http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf
[11] http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf

2. Übernahmerichtlinie des IDS für schriftliche Korpora

Marc Kupietz, Harald Lüngen, 02.10.2015

Das IDS ist an der Übernahme schriftlicher Korpora in das Deutsche Referenzkorpus (DeReKo) und das Langzeitarchiv (LZA) für germanistische Primärdaten interessiert. Für die Übernahme schriftlicher Korpora gelten folgende Richtlinien:

Welche Daten kommen für eine Übernahme in Frage?

(1) Der Fokus liegt auf Korpora mit Texten der deutschen Gegenwartssprache (ab ca. 1956), die im Original von Muttersprachlern auf Deutsch verfasst wurden, also nicht übersetzt wurden. Von Interesse sind darüber hinaus auch Archive älterer Texte, die eine besondere Relevanz für die Gegenwartssprache haben (etwa Korpora aus den Werkausgaben bedeutender Autoren). Textdaten in anderen Sprachen kommen nur dann für eine Übernahme in Frage, wenn sie integraler Bestandteil des Korpusdesigns sind, z.B. bei einem alignierten oder vergleichbaren Korpus, das z.B. für kontrastive Untersuchungen angelegt wurde. Korpora, die keinen oder nur einen geringen Anteil an Schriftdeutsch enthalten, kommen für eine Übernahme nicht in Frage.

(2) Außer Lyrik wird kein Genre und keine Textsorte generell ausgeschlossen, d.h. Interesse besteht z.B. an Korpora mit belletristischen Texten, Pressetexten, Lehrbüchern, Gebrauchstexten, wissenschaftlichen Texten, amtlichen Texten, Werbetexten, Broschüren oder Bedienungsanleitungen, handschriftliche Texte (Briefe u.a.) sowie Webtexte (aus dem Web gewonnene Korpora). Unter Schriftkorpora verstehen wir auch Korpora internetbasierter Kommunikation (IBK, z.B. Chat, Wiki, Foren, Instant Messaging), schriftsprachliche Vorlagen mündlich vorgetragener Texte (z.B. Lesetexte, Rezitationen, Aufführungen, Redemanuskripte, Drehbücher) sowie Protokolle von Interviews, Rundfunksendungen, Reden, Vorträgen oder Debatten.

(3) Zu übernehmende Korpora sollten vollständige Texte enthalten, d.h. nicht nur aus Textausschnitten bestehen und sollten digital, in einer textbasierten Form (d.h. nicht nur als gescannte Bilder) vorliegen. Eine basale Korpusstruktur (Textgrenzen) und basale Metadaten der Korpustexte (Titel, Autor, Publikationsdatum oder Entstehungszeit) sollten vorhanden sein oder sich zumindest automatisch rekonstruieren lassen. Das Encoding der Texte sollte bekannt und einheitlich sein.

(4) Bei schriftlichen Daten, die aus einem der o.g. Gründe für eine Übernahme in DeReKo nicht in Frage kommen – etwa schriftliche Korpora in anderen Sprachen als Deutsch oder historische Korpora – kann das IDS bei der Auswahl eines geeigneten Daten-Zentrums behilflich sein. Wir verweisen hier insbesondere auf das CLARIN-Zentrum an der Berlin-Brandenburgischen Akademie der Wissenschaften [3].

Mit welchem Ziel erfolgt eine Übernahme?

(5) Das IDS übernimmt schriftliche Korpora mit dem Ziel, sie dauerhaft zu bewahren, nutzbar zu halten und unter Beachtung rechtlicher Vorgaben der wissenschaftlichen Community zur Verfügung zu stellen.

Wie wird über eine Übernahme entschieden?

(6) Die Entscheidung zur Übernahme eines schriftlichen Korpus erfolgt auf der Basis einer Kosten-Nutzen-Analyse unter Berücksichtigung der in den zuständigen Programmbereichen vorhandenen personellen Kapazitäten.

(7) Die Kosten-Nutzen-Analyse wird von Mitarbeitern der Projekte Korpusausbau und LZA in Zusammenarbeit mit dem potenziellen Datengeber durchgeführt. Mindestanforderungen für die Übernahme eines Korpus sind eine Inventarisierung des Korpus, eine Dokumentation des Korpusdesigns und der Korpusbestandteile, sowie die verbindliche Klärung rechtlicher Bedingungen der Datenweitergabe. Weitere Orientierungshilfen für die Kriterien zur Beurteilung eines schriftlichen Korpus finden sich in den DFG-Handreichungen „Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora“ [1] und „Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora“ [2].

(8) Mit Blick auf die begrenzten personellen Kapazitäten von Korpusausbau und LZA sind Beiträge des Datengebers zur Unterstützung der Korpusaufbereitung ausdrücklich erwünscht.

(9) Fällt die Entscheidung zur Übernahme positiv aus, kann zwischen dem Datengeber und der Leitung des IDS eine schriftliche Vereinbarung über die wesentlichen Punkte der Datenübernahme getroffen werden.

Was geschieht mit übernommenen Daten?

(10) Übernommene Korpora werden nach den Standards des Deutschen Referenzkorpus (DeReKo) aufbereitet. Je nach Beschaffenheit der Daten beinhaltet dies die Überführung unstrukturierter Daten in eine strukturierte Form, eine systematische Dokumentation von Metadaten und/oder die Anreicherung von Daten mit zusätzlichen Annotationen.

(11) Nach Abschluss der Aufbereitung werden die Daten der wissenschaftlichen Community über geeignete, vom IDS bereitgestellte Mechanismen zur Verfügung gestellt. Derzeit beinhaltet dies insbesondere die Einspeisung der Daten ins IDS-Repositorium, welches Metadaten zum Korpus für geeignete Kataloge digitaler Infrastrukturen (derzeit insb. CLARIN VLO) zur Verfügung stellt, sowie ggf. eine Integration der Daten in das Deutsche Referenzkorpus (DeReKo) und das zugehörige Korpusrecherchesystem (derzeit COSMAS II oder KorAP). Künftig mögen weitere, hier nicht genannte, Mechanismen der Datenhaltung und -weitergabe zur Anwendung kommen. Die Weitergabe der Daten erfolgt unter Beachtung rechtlicher Vorgaben.

(12) Die Daten werden im Rahmen der technischen und personellen Möglichkeiten des IDS gesichert und gepflegt, d.h. an geeignete Backupmechanismen angeschlossen bzw. an zukünftig geltende Standards von DeReKo, des LZAs oder digitaler Infrastrukturen angepasst.

(13) Zu den Daten gehörende Software-Umgebungen (z.B. Word-Prozessoren, Dokumenten-Management-Systeme, Blog-/Chat-Plattformen Wiki-Systeme) können in aller

Regel bei der Datenübernahme nicht berücksichtigt werden, d.h. sie können vom Projekt Korpusausbau oder im Rahmen der LZA nicht dauerhaft gepflegt werden.

Wo gibt es weitere Informationen?

Wenn Sie ein schriftsprachliches Korpus haben, das für eine Übernahme in Frage kommt, kontaktieren Sie uns unter (0621) 1581-437 oder dereko@ids-mannheim.de.
Diie DFG-Handreichungen finden Sie unter folgenden URLs:

[1]
http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

[2]
http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf

[3] https://clarin.bbaw.de/

3. Übernahmerichtlinie des IDS für andere wissenschaftlich aufbereitete Ressourcen deutscher Sprache

Peter Fankhauser, 06.09.2017

Welche Daten kommen für eine Übernahme in Frage?

(1) Unter einer wissenschaftlich aufbereiteten Ressource deutscher Sprache verstehen wir eine systematische Sammlung von digitalisierten Textdaten zusammen mit manuellen oder automatischer Annotationen. Dazu gehören Textsammlungen für einen bestimmten Zeitraum und Themenbereich (z.B. Korpus Diskurs in der Weimarer Republik [1]) und Zusatzmaterialien zu germanistischen Veröffentlichungen (z.B. Zusatzmaterialien zu einer Dissertation [2]).

(2) Ressourcen, die von den Übernahmerichtlinien des IDS für mündliche oder schriftliche Korpora abgedeckt sind, sind hiervon ausdrücklich ausgenommen.

Mit welchem Ziel erfolgt eine Übernahme?

(3) Das IDS übernimmt deutsche Sprachressourcen mit dem Ziel, sie dauerhaft zu bewahren und unter Beachtung rechtlicher Vorgaben der wissenschaftlichen Community zur Verfügung zu stellen.

Wie wird über eine Übernahme entschieden?

(4) Die Entscheidung zur Übernahme einer Sprachressource erfolgt auf Basis einer Kosten-Nutzen-Analyse unter Berücksichtigung der in den zuständigen Programmbereichen vorhandenen personellen Kapazitäten.

(5) Mit Blick auf die begrenzten personellen Kapazitäten des LZA sind Beiträge des Datengebers zur Unterstützung der Ressourcenaufbereitung ausdrücklich erwünscht.

(6) Fällt die Entscheidung zur Übernahme positiv aus, wird zwischen dem Datengeber und der Leitung des IDS eine schriftliche Vereinbarung über die wesentlichen Punkte der Datenübernahme getroffen.

Was geschieht mit den übernommenen Daten?

(7) Übernommene Ressourcen werden nach den Standards des LZA mit Metadaten versehen. Im Minimum werden dazu aussagekräftige Dublin-Core-Metadaten erhoben und im CMDI-Format enkodiert, sowie persistente Identifikatoren angelegt.

(8) Auf Basis der Metadaten werden die Ressourcen im IDS-Repositorium[3] abgelegt und so auch in geeigneten Katalogen, wie dem CLARIN VLO[4], sichtbar gemacht.

(9) Soweit vom Datengeber zugelassen, werden auch die Ressourcen selbst allgemein zugänglich gemacht, gegebenenfalls mit entsprechenden Nutzungsbestimmungen. Andere

Ressourcen werden auf Basis von Shibboleth einem eingeschränkten Nutzerkreis zugänglich gemacht.

Referenzen

[1] Korpus Diskurs in der Weimarer Republik. http://hdl.handle.net/10932/00-01B9-43B3-1E1D-7B01-6

[2] Zusatzmaterialien der Dissertation "Automatische Erkennung von Redewiedergabe". http://hdl.handle.net/10932/00-027B-9E8A-9300-0B01-E

[3] https://repos.ids-mannheim.de/

[4] https://vlo.clarin.eu/