Medienlinguistische Methodik

Offene Forschung

Arne Rubehn
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau

20.01.2026

Überblick

In dieser Sitzung wollen wir uns mit einigen Grundlagen der offenen Forschung befassen. Dabei wollen wir auch auf konkrete Lösungsvorschläge eingehen, die eigene Forschung selbst offen zu gestalten.

Welche Aspekte sind für die offene Forschung relevant?

Wie können Forschungsdaten und Code langlebig gemacht werden?

Welche Plattformen sind zur Langzeitarchivierung geeignet?

Grundlagen

Mit **offener Forschung** (*Open Science*) bezeichnet man eine Haltung in der Forschung, die nach **maximaler Offenheit** aller Prozesse, die zu neuen Erkenntnissen führt, strebt.

Die offene Forschung wird von vielen Menschen auf der Welt inzwischen mehr oder weniger radikal propagiert, wobei zu beachten ist, dass diese Forderung in einigen Bereichen mit **wirtschaftlichen** oder **politischen Interessen kollidiert**.

In **kleinen, harmlosen Bereichen** (wie eben der multilingualen CL) lässt sich das Ideal der offenen Forschung allerdings meistens auch in seiner Reinform aufrecht erhalten.

Grundlagen

Im Zusammenhang mit offener Forschung werden verschiedene Aspekte diskutiert.

Einer dieser Aspekte ist die **freie Verfügbarkeit von Materialien** (*Open Access*), also etwa Artikeln oder Lehrbüchern. Zwar sind inzwischen deutlich mehr Forschungsmaterialien frei zugänglich verfügbar – allerdings mit der Konsequenz, dass Verläge teils horrende Gebühren von den Institutionen der Autor:innen verlangen.

Autor:innen haben jedoch immer die Möglichkeit, ein **persönliches Manuskript** im Internet frei zugänglich zu machen (etwa über öffentliche Archive wie arXiv).

Grundlagen

Andere wichtige Aspekte der offenen Forschung sind die Offenlegung von **Forschungsdaten** (*Open Data*) und **Software** (*Open Code*).

Das Grundziel der offenen Forschung ist, dass die folgenden drei Eigenschaften auf alle Aspekte zutreffen:

- **wiederholbar**
- **transparent**
- **integrierbar**

Wiederholbarkeit

Wiederholbarkeit bezieht sich darauf, dass eine Forschung, die man selbst durchführt, auch von anderen durchgeführt können werden sollte. Das Wiederholen ist dabei in zweierlei Hinsicht wichtig.

Zum Einen gibt es anderen Menschen die Möglichkeit, von einem bestimmten **Ansatz** zu **lernen** und diesen zu **erweitern**.

Zum Anderen kann man durch das Wiederholen sicherstellen, dass die Messungen, die vorgenommen wurden, auch tatsächlich **unabhängig** von der forschenden Person **Bestand haben**.

Transparenz

Um Wiederholbarkeit zu gewährleisten, muss inbesondere auch **Transparenz** vorausgesetzt werden. Transparenz bezeichnet, dass die Forschung so beschrieben wurde, dass **alle durchgeführten Schritte** und die zugrundeliegenden Daten **klar benannt** werden.

Transparenz wird leider häufig unabsichtlich vernachlässigt: So werden z.B. Vorverarbeitungsschritte der Daten oder die benutzten Softwareversionen nicht ausreichend beschrieben. Mangelnde Transparenz **schadet der Wiederholbarkeit**; auch bei Ansätzen, die per se wunderbar wiederholbar wären.

Integrierbarkeit

Unter **Integrierbarkeit** wird verstanden, dass Forschung, die zu einem Thema durchgeführt wurde, sich auch mit **anderen Forschungsansätzen** integrieren lässt.

Das bedeutet insbesondere, dass man seine eigene Forschung so aufbauen sollte, dass **andere Forscher:innen** diese in ihre eigene Forschung einbauen können.

Grundlage dafür sind **Transparenz** und **Wiederholbarkeit**: Nutzung von etablierten Standards, modularem Code, frei zugänglichen Tools, ...

Einfach gesagt: Forscherin B sollte einzelne Teile von Forscher A mühelos in ihre Forschung integrieren können.

Grundlagen

Illustrieren wir diese Grundlagen mal am Beispiel eines Kuchenrezeptes.

Wiederholbarkeit. Ich sollte selbst das Rezept so gut beherrschen, dass auch jedes Mal der gleiche Kuchen herauskommt.

Transparenz. Mein Rezept sollte alle Zutaten und alle Arbeitsschritte genau aufschreiben, sodass andere Leute selbständig den gleichen Kuchen backen können.

Integrierbarkeit. Andere Leute können Teile meines Rezeptes in ihre eigenen Rezepte integrieren, oder mein Rezept modifizieren (z.B. eine vegane Variante ableiten).

Offene Daten

Der nächste wichtige Aspekt für offene Forschung, ist, dass **Forschungsdaten offen zugänglich** gemacht werden. Wilkinson et al. prägen hierfür das Akronym **FAIR**, das die vier wichtigsten Kriterien zusammenfasst:

- **F**indability (*Auffindbarkeit*)
- **A**ccessibility (*Zugänglichkeit*)
- **I**nteroperability (*Integrierbarkeit*)
- **R**eusability (*Wiederverwertbarkeit*)

Findability

Um Daten **auffindbar** zu machen, muss man sie so gestalten, dass sie so gut beschrieben werden, dass man sie beispielsweise über Suchmaschinen findet.

Dabei ist es auch wichtig, dass Daten in **Archiven** angeboten werden, die sich **leicht durchsuchen** lassen.

Auffindbarkeit ist wohl das schwächste Glied in der Kette, da sie auch von der **Qualität der verfügbaren Server/Archive** abhängt, die Forschende nur bedingt kontrollieren können.

Accessibility

Dass man auf Daten **zugreifen** kann, ist eine Grundvoraussetzung für offene Wissenschaft. Natürlich muss man hierbei darauf achten, dass **sensible Daten** (etwa personenbezogene oder sicherheitsrelevante) angemessen vorverarbeitet werden (z.B. durch Pseudonymisierung).

Die meisten Daten in der **Linguistik** (und generell in den Geisteswissenschaften) sind dahingehend allerdings **unproblematisch** – hier gibt es keine Ausreden!

Mangelnde Zugänglichkeit von nicht sensiblen Daten liegt entweder an **finanziellen Interessen** (z.B. Paywalls) oder **technischen Problemen** (z.B. abgeschaltete Server).

Interoperability

Über Integrierbarkeit haben wir bereits im weiteren Sinne gesprochen. Im Bezug konkret auf Daten meint das, dass Forschungsdaten reibungslos in anderen Studien **weiterverwendet** werden können und insbesondere auch mit Daten aus anderen Quellen **kombiniert** werden können.

In diesem Kontext ist insbesondere die Nutzung von **verbindlichen Standards** relevant.

Reusability

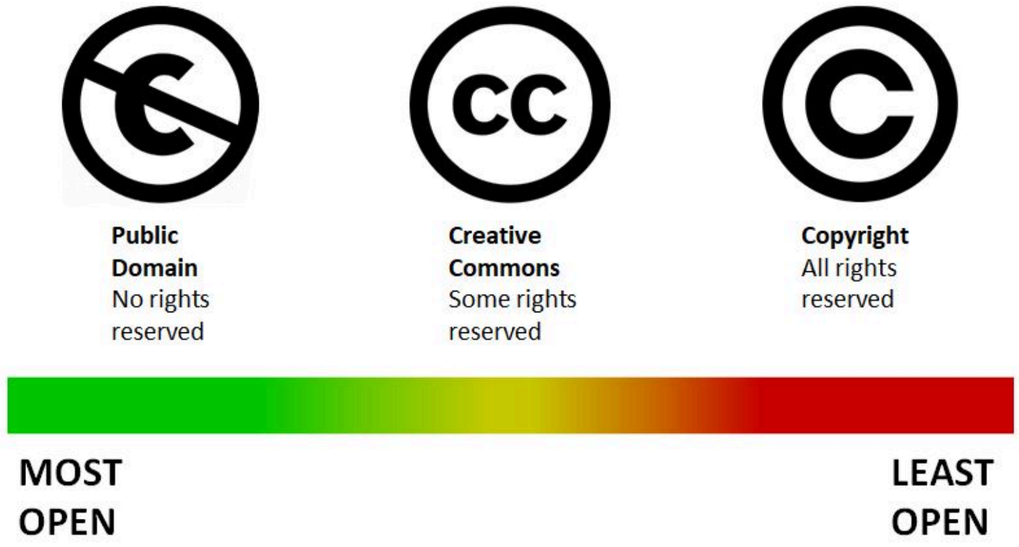
Die Wiederverwertbarkeit von Daten setzt vor Allem voraus, dass sie **offene Lizenzen** bekommen, die das Wiederverwerten erlaubt. Es gibt Datensätze, die zwar frei zugänglich sind, aber so restriktiv lizensiert sind, dass man sie in keinster Weise wiederverwenden kann. Solche Daten sind daher leider absolut unbrauchbar.

Bei Lizenzen ist zu beachten, dass Urheber:innen von Daten und Software diese **explizit** wählen müssen – sonst greifen standardmäßig Regelungen des **Urheberrechts**, die Urheber:innen alle Rechte einräumen, was jedoch die Weiterverwendung dieser Daten äußert schwierig macht!

Reusability

Bei Forschungsdaten fällt die Wahl meistens auf eine **Creative Commons (CC-BY)**-Lizenz, die grundlegend die Weiterverwendung von Daten erlaubt. Urheber:innen können hierbei modular festlegen, welche Formen der Weiterverwendung gestattet sind (z.B. ob die kommerzielle Nutzung gestattet ist oder Derivate aus den Daten erstellt werden dürfen).

Unter CC-BY-Lizenzen müssen die Urheber:innen bei jeglicher Weiter-verwendung **zitiert** werden.



Reusability

Standardisierung

Auch das Thema der Standardisierung haben wir bereits kurz angeschitten. Hierunter verstehen wir im Bezug auf Daten eine **unabhängige Vorgabe**, nach der man sich richtet, wenn man Daten bearbeitet.

Einfache Beispiele für Standards sind **Dateiformate** wie CSV oder JSON, die klar vorschreiben, wie Daten zu formatieren sind. Dadurch wird gewährleistet, dass verschiedene Programme die Daten problemlos einlesen können.

Neben formalen Vorgaben können Standards auch definieren, welche **Datenpunkte** oder **Metadaten** inkludiert werden müssen.

Referenzkataloge

Wenn wir in unserer Forschung mit Forschungsobjekten arbeiten, die als **Konstrukte klar definiert** und auch von der Forschungsgemeinschaft als **klare Einheiten** aufgefasst werden, dann lohnt es sich oft nicht, die Information, die wir über diese Objekte haben, unabhängig zusammenzutragen.

Stattdessen ist es einfacher und produktiver, auf einheitliche **Kataloge** zu verweisen, die diese Information bereits anbieten.

Ein Beispiel für einen solchen Referenzkatalog ist **Glottolog**, wo Informationen zu den Sprachen dieser Welt gesammelt werden und frei verfügbar sind.

Referenzkataloge

Viele Bezeichnungen für eine Sprache…

…aber **ein** Glottocode (axam1237)!

Alternative names

elcat:
Ahamb
Akamb
Akhamb
Axamb
naujin sdrato [nauˈtʃin sndraˈto] ‘our language’
glottolog:
Ahamb
lexvo:
Axamb [en]
multitree:
Ahamb
Axamb
Mallicolo

Cross-Linguistic Data Formats

Ein Beispiel für Standards für Forschungsdaten in der MCL ist die Initiative **Cross-Linguistic Data Formats (CLDF)**, die an unserem Lehrstuhl angesiedelt ist.

Die Kernidee ist es, Sprachdaten so einheitlich zu formatieren, dass eine Vielzahl von Quellen reibungslos kombiniert werden können. Dies ermöglicht große Kollektionen über den **Wortschatz** (*Lexibank*) oder **grammatische Merkmale** (*Grambank*) verschiedener Sprachen.

Wie bei vielen anderen Standards gibt es auch hier Computerprogramme, die die Integrität der Daten **validiert.**

Archivierung

Ein wichtiger Aspekt, um den FAIR-Kriterien gerecht zu werden, ist die **Langlebigkeit** der Daten sicherzustellen. Hierbei spielen **Archive** eine zentrale Rolle.

Als Archive werden in diesem Kontext **öffentlich finanzierte Plattformen** bezeichnet, die zusichern, Daten über einen langen Zeitraum sicher aufzubewahren und zugänglich zu machen.

Proprietäre Plattformen (Google Drive, Dropbox, ...) eignen sich hingegen nicht zur Langzeitarchivierung, da diese jederzeit **abgeschaltet** oder **verschlossen** werden können.

Versionierung

Wenn man Daten und Software zugänglich macht, sollte man auch darauf achten, dass diese sauber **versioniert** sind.

Die Idee dahinter ist simpel: Daten und Software werden in vielen Projekten konstant weiterentwickelt oder modifiziert. Werden sie in einem bestimmten Stadium für eine Analyse verwendet, möchte man den entsprechenden **Stand der Daten** in einem **Schnappschuss** festhalten – also einer Version.

Insbesondere in der Softwareentwicklung ist es elementar wichtig, die genutzten Versionen sauber zu protokollieren – sonst können ganze Programme plötzlich **nicht mehr funktionieren!**

Versionierung

Eines der wichtigsten Tools zur Versionierung, das insbesondere in der Softwareentwicklung routinemäßig verwendet wird, ist **Git.** Git wird auf Plattformen wie GitHub, GitLab oder Codeberg gehostet.

Git ermöglicht es, Dateien konstant zu verändern, ohne ältere Versionen zu verlieren. Diese können jederzeit wiederhergestellt werden – wir brauchen also keine “Hausarbeit_neu_2_final_FINAL_JETZTABERWIRKLICH.docx” mehr.

Open Science Framework

Das **Open Science Framework (OSF)** ist eine gemeinnützige Organisation, die verschiedene Services zur Verfügung stellt, um offene Forschung zu betreiben.

Dazu gehört die Möglichkeit, **Daten online** zu speichern und **anonyme Links** zu erzeugen, über die andere Personen auf die Daten zugreifen können, ohne zu sehen, wer dahinter steht. Das ist vor Allem für **anonyme Gutachten** von wissenschaftlichen Artikeln sehr nützlich.

Hierbei sind Projekte standardmäßig nicht öffentlich einsehbar, entsprechende Links erlauben aber einen einfachen Zugriff.

Zenodo

Zenodo ist einer der größten Services für das **Teilen von Forschungsdaten**, die wir in Europa haben. Es wird von der EU finanziert und ist am CERN angesiedelt. Es ist davon auszugehen, dass Zenodo noch für lange Zeit aufrecht erhalten wird.

Viele Institutionen, die **Forschungsprojekte** fördern (etwa das *European Research Council*) stellen **strenge Auflagen** für die Bereitstellung und Archivierung von Forschungsdaten. Archive wie Zenodo bieten verhältnismäßig komfortable Wege, diese Auflagen zu erfüllen.

Über sog. Communities können Daten auch einfach nach Forschungsbereich kategorisiert und durchsucht werden.

Anlaufstellen

Zum Themenkomplex Open Science, insbesondere Open Data, hat die Universität Passau die **AG Forschungsdatenmanagement** ins Leben gerufen. Sie fungiert als Bindeglied zwischen Forschungsförderung, Universitätsbibliothek und ZIM und bietet Beratung zum offenen Umgang mit Forschungsdaten und -ergebnissen an.

Im Umgang mit **personenbezogenen Daten** bietet es sich zusätzlich an, ein Gutachten der **Ethikkommission** einzuholen.

All diese Dienste stehen Forschenden sowie Studierenden zur Verfügung.