
Open Science in den Humanities: Mehr Spielräume und Nachvollziehbarkeit durch Galaxy

Daniela Schneider ¹, Jan Leendertse ²

¹ Galaxy Team Freiburg, Albert-Ludwigs Universität Freiburg;

² Central Data Facility, Rechenzentrum, Albert-Ludwigs Universität Freiburg

In den Geisteswissenschaften kann der Einstieg in digitale Analysen mit etlichen Hürden verbunden sein. Fehlende Programmierkenntnisse, hohe Kosten für Spezialsoftware oder fehlende Administratorenrechte können eine Herausforderung sein. Dazu soll gute wissenschaftliche Praxis umgesetzt und gutes Forschungsdatenmanagement betrieben werden. Je mehr dieser Anforderungen zusammenkommen, desto herausfordernder wird die Umsetzung eines neuen digitalen Projektes.

Dieser Artikel untersucht, wie die Open Source Analyseplattform Galaxy an genau diesen Herausforderungen ansetzen kann. Am Praxisbeispiel eines Textvergleichs aus einer geschichtswissenschaftlichen Dissertation zeigt der Artikel, wie die in den Lebenswissenschaften entwickelte Plattform auch in den Geisteswissenschaften sinnvoll einsetzbar ist. Die Beispielanalyse, die ursprünglich manuell durchgeführt wurde und sehr zeitaufwändig war, wurde automatisiert, was die Effizienz und die Aussagekraft der Ergebnisse erheblich steigerte. Der in Galaxy erstellte Workflow ermöglicht eine transparente und reproduzierbare Analyse, wodurch die Prinzipien der guten wissenschaftlichen Praxis gefördert werden. Zudem wird gezeigt, dass auch in den Humanities eine Offenheit für Open Science möglich ist, selbst wenn die zugrunde liegenden Daten aus proprietären Quellen stammen. Dieses Beispiel verdeutlicht nicht nur die Nützlichkeit von Galaxy als Werkzeug zur Textanalyse, sondern durch das Baukastenprinzip der Plattform auch das Potenzial für die breite Anwendung in verschiedenen Fachbereichen. Durch die Implementierung von Galaxy wird es Forschenden ermöglicht, sich stärker auf ihre Kernkompetenzen zu konzentrieren und gleichzeitig die Sichtbarkeit und Nachvollziehbarkeit ihrer Methoden zu verbessern, wodurch die Herausforderungen der digitalen Analyse in den Geisteswissenschaften adressiert werden.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: <https://doi.org/10.11588/heibooks.1652.c23955> (CC BY-SA 4.0).

Keywords: Humanities, Open Science, FAIR

1 Open-Science in den Humanities?

Open Science ist als Prinzip breit zustimmungsfähig. In den Geisteswissenschaften sehen viele Forschende für sich jedoch keinen schnellen oder praktikablen Weg, dem zu folgen. Open Science wird mit Digitalisierung assoziiert, die in den Humanities vielerorts Abwehr auslöst. Darüber hinaus soll in den Projekten gute wissenschaftliche Praxis befolgt und gutes Forschungsdatenmanagement betrieben werden. Diese Anforderungen an gute wissenschaftliche Qualität und Nachvollziehbarkeit können für Forschende jedoch auch eine Hürde darstellen. Kommen zu diesen noch weitere Barrieren wie hohe Kosten für Software oder fehlende Administratorenrechte, um Programme auf dem eigenen Computer zu installieren, oder müssen erst Programmiersprachen für die Analyse erlernt werden, wird der Einstieg in digitale Analysemethoden schnell zur Herausforderung.

2 Textanalyse mit Galaxy

Am Beispiel einer Dissertation aus den Geschichtswissenschaften wird gezeigt, wie die Open Source Plattform Galaxy (The Galaxy Community u. a. 2024) als Self-Service-Tool den Einstieg in digitale Textanalyse erleichtert (Schneider und Leendertse 2025). Obwohl Galaxy in den Lebenswissenschaften entwickelt wurde, wird hier deutlich, dass die Plattform großes Potential für die Geisteswissenschaften birgt. Die Beispielanalyse wurde in der Dissertation mit bekannten geschichtswissenschaftlichen Methoden durchgeführt und war nicht genuin als digitaler Prozess geplant. Methodisch wurde die Quellenlage durch den Vergleich von Zeitungsartikeln hergestellt, die in unterschiedlichen politischen Räumen entstanden sind. Sie stimmen zum großen Teil überein, weisen aber sichtbare Unterschiede in Form von Ersetzungszeichen auf, die auf Zensureingriffe zurückgehen. Der Textvergleich wurde ursprünglich manuell durchgeführt, war sehr zeitaufwändig und repetitiv. Hier setzt Galaxy¹ als Open Source Plattform an.

Galaxy bietet eine grafisch orientierte Oberfläche, die zum Experimentieren einlädt. Auf Wunsch kann das Ausprobieren aufgezeichnet, im Nachgang bereinigt und als Prozedur gespeichert werden. Solche Workflows werden so zum Ausgangspunkt weiterer Verfeinerung, sind teilbar, transparent und im besten Fall reproduzierbar. Die FAIR-Prinzipien bleiben kein schwer erfüllbarer Anspruch, sondern werden beiläufig gelebtes Prinzip. Für viele Fachbereiche gibt es bereits zahlreiche Trainingsmodule und Best-Practice-Workflows in Galaxy, die ohne Barrieren als Open Educational Resources (OER) zugänglich sind (Hiltemann u. a. 2023).

¹ <https://usegalaxy.eu/>; *Besucht am 4. Juni 2025.*

Das Beispiel zum Textvergleich chinesischer historischer Zeitungen macht dies deutlich. Statt zäher manueller Gegenüberstellung können in Galaxy zensierte und unzensierte Zeitungsartikel mit chinesischen Zeichen schnell und übersichtlich gegenübergestellt werden. Der Vergleich ist visuell aufschlussreich und kann mit Standardtools wie ‘diff’² ausgewertet werden. Dadurch können auch große Textmengen schnell analysiert werden. Technisch überraschend einfach, aber mit hoher Aussagekraft hinsichtlich der Forschungsfrage, welche Inhalte in den Zeitungen zensiert wurden. Bei der Auswertung ist weiterhin spezifische Erfahrung aus den jeweiligen Fachgebieten gefragt, besonders bei der kritischen Einordnung von Zwischenergebnissen und daraus folgend der weiteren Bearbeitung und tiefergehenden Analyse des Materials. Die Nutzung von Galaxy zeigt, wie Teile der Analyse auch in den Geisteswissenschaften teilautomatisiert werden können, so dass größere Textmengen verarbeitet werden können. So werden Ergebnisse nachvollziehbarer. Gleichzeitig bleibt für die eigentliche Auswertung mehr Zeit.

Der Workflow, die einzelnen Analyseschritte und die Daten jedes Teilschritts können dabei gemäß der guten wissenschaftlichen Praxis geteilt und wiederholt werden. Obwohl die Ausgangsdaten im Beispiel wegen ihres proprietären Ursprungs nicht zugänglich sind, kann so der Prozess geteilt werden. Das zeigt gleichzeitig, dass Open Science selbst dort möglich ist, wo Daten aus guten Gründen nicht offen sind.

3 Methodisches Vorgehen

Für einen Textvergleich auf Zeichen- und Zeilenebene eignet sich das Tool ‘diff’³, das aus der Unix-Welt stammt, aber auch für Windows portiert ist. Der Gebrauch dieses Kommandozeilentools ist nicht Teil des historischen Studiums und erfordert autodidaktische Einarbeitung. Zusätzlich eine Umgebung aufzubauen, die das Vorgehen reproduzierbar macht, erfordert weitere Kenntnisse, die nicht zum Curriculum von Historiker:innen gehören.

Der Beispielworkflow, der diesem Artikel zu Grunde liegt (Schneider 2025), zeigt, wie diese Textanalyse schrittweise im Galaxy erarbeitet werden kann. Diese Analyseplattform protokolliert das Vorgehen, das als Workflow gespeichert und verfeinert werden kann.

Das grundlegende Forschungsinteresse ist, welche Zeichen in einem chinesischen Zeitungsartikel aus Hong Kong in den 1930er Jahren zensiert wurden. Dazu werden zwei Zeitungsartikel miteinander verglichen. Da die der Arbeit zugrunde liegenden Artikel aus einer proprietären Datenbank stammen, wurden für ein Tutorial, das im Nachgang zur eigentlichen Forschung angeschlossen wurde, stattdessen mit Sprachmodellen Dummy-Texte erstellt. Mit diesen Texten kann die Analyse schrittweise nachgearbeitet werden, um das

² Das Programm ermöglicht eine automatisierte Gegenüberstellung von Texten und zeigt deren Unterschiede.

³ <https://pubs.opengroup.org/onlinepubs/9799919799/utilities/diff.html>, *Besucht am 27. März 2025.*

Verfahren auf ähnliche Forschungsfragen anzuwenden. Mit diesen Texten und dem Tutorial bleiben die originalen Quellen zwar weiterhin unter Verschluss, doch die Methoden werden offengelegt und können nachgenutzt werden.

Einer der Artikel enthält offensichtliche Zensurzeichen in Form von ×, welches einzelne chinesische Zeichen ersetzte. Durch intensive Recherche konnte im Original eine frühere Version desselben Artikels unzensiert recherchiert werden, der in Festlandchina veröffentlicht wurde.

Statt die beiden Texte mühevoll manuell miteinander zu vergleichen, um herauszufinden, welche Zeichen mit × ersetzt wurden, war dies mit der Plattform Galaxy automatisiert möglich. Der Workflow lässt sich dabei in verschiedene Schritte unterteilen, die mit jeweils unterschiedlichen Programmen (in Galaxy Tools), die sich zu einem Workflow miteinander verbinden lassen. Zunächst werden die Texte hochgeladen und für den Vergleich vorbereitet. Der Übersichtlichkeit halber, werden sie als ein Zeichen pro Zeile dargestellt und Leerzeichen gelöscht. ‚Diff‘ stellt beide nun identisch vorbereiteten Texte visuell gegenüber und markiert Veränderungen farblich. Auslassungen werden rot und die Einfügungen grün markiert.

184	胞	179	胞
185	言	180	言
186	之	181	之
187	- 、	182	+ 。
188	- ×	183	+ 敵
189	- ×	184	+ 寇
190	此	185	此
191	- 舉	186	+ 舉
192	- 、	187	+ ，
193	其	188	其
194	用	189	用
195	意	190	意

Abbildung 1: Screenshot des Textvergleichs mit ‚Diff‘-Tool in Galaxy.

Hier werden auch die Zeilen farblich hervorgehoben, in denen die Zensurzeichen vorkommen. Nach ihnen wird im nächsten Schritt gefiltert, sodass das Dokument nur noch × und ihre unzensierten Äquivalente enthält. Für zwei verschiedene Arten der Visualisierung werden die extrahierten Zeichen nun einmal quantifiziert und in einer Tabelle sortiert und einmal als Wordcloud grafisch ausgegeben. Der Vergleich der Artikel zeigt, dass Zeichen wie 敵 (Feind) und 寇 (Eindringling) häufig zensiert wurden. So beispielsweise auch in den Zeilen 188 und 189 der obigen Abbildung. Dies verweist auf die Zensurpolitik der britischen Regierung, die bestrebt war, anti-japanische Inhalte zu vermeiden, um mögliche

Spannungen mit Japan zu verhindern. Japaner in den in Hong Kong veröffentlichten Zeitungsartikeln als Feinde oder Eindringlinge zu bezeichnen, offenbarte eine pro-chinesische Haltung der Zeitungen. Diese zensierte die britische Hong Kong Regierung, die im Konflikt neutral bleiben wollte, um nicht selbst in den Konflikt verwickelt zu werden. Die Gegenüberstellung der Zeitungstexte im Beispieltutorial unterstreicht diese Erkenntnisse.

4 Erschließbarkeit der Ergebnisse durch Galaxy

Analysen werden parametrisierbar und schaffen eine neue Vergleichbarkeit von Ergebnissen. Daten und Inhalte von Plattformen, wie beispielsweise Zenodo, können in Galaxy importiert und in einem Workflow algorithmisch analysiert werden. Die Ergebnisse sind wiederum in einem frei wählbaren Repositorium (Zenodo, FreiData, etc.) als FAIRes Digitales Objekt (FDO), beispielsweise RO-Crate, speicherbar. FDO sind auf Export und maschinenlesbare Weiterverarbeitung ausgelegt.

Galaxy bietet Open Educational Resources (OER) – ebenso Arbeitsfeld von Open Science – im Galaxy Training Network (GTN), das bereits über 400 Tutorials enthält, die über offene Lizenzen bereitgestellt werden. Neben den praktischen Tutorials bietet es Beispieldatensätze, E-Learning-Material, Videos, Aufzeichnungen und komplette Schulungsworkshops, die Forscher:innen weltweit zum Selbststudium zur Verfügung stehen. Das Schulungsnetzwerk wird von einer globalen Community unterstützt, die neuen Nutzer:innen den Einstieg erleichtert und in kollaborative Arbeitsformen einführt. Durch diesen Ansatz haben verschiedene Wissenschaftsfelder Galaxy für sich entdeckt. Erste Analyseprojekte aus den Humanities finden sich bereits auf der Plattform.⁴ So versucht Galaxy Forschende dabei zu unterstützen, sich auf das Kerngeschäft ihrer Arbeit zu fokussieren: die Forschung und Untersuchung/Analyse von Material und Daten. Reproduzierbarkeit und die Grundlage für gutes Forschungsdatenmanagement sind automatisches „Nebenprodukt“ der Analyse und verringern so die Hürden, diese Erwartungen an gute, FAIRe Forschung tatsächlich umzusetzen. Technisch gibt es sehr viele bereits integrierte Programme, die kostenlos, ohne Installation und ohne Programmierkenntnisse im Browser abrufbar sind. Dies ist auch in den Geisteswissenschaften von großem Interesse für die Forschenden.

⁴ Liste von Digital Humanities Tutorials im Galaxy Training Network: <https://training.galaxyproject.org/training-material/topics/digital-humanities/>; *Besucht am 9. Juli 2025 durch die Herausgeber*innen.*

5 Fazit

Zusammenfassend zeigt das Fallbeispiel, wie Galaxy als Open Source Plattform digitale Analysen in den Geisteswissenschaften unterstützen kann. Da Galaxy ohne Installation im Internet-Browser zugänglich ist und Analysetools mit wenig Einarbeitungsaufwand zugänglich macht, verringert sie die eingangs erwähnte Hürde für digitale Textanalysen. Durch das Baukastenprinzip ist Galaxy zudem sehr flexibel einsetzbar, und kann leicht für das eigene Forschungsvorhaben angepasst werden. Die einzelnen Schritten innerhalb eines protokollierten Workflows werden automatisch mit den Zwischenergebnissen dokumentiert. Die Ergebnisse lassen sich auch kleinschrittig teilen. Auch der Workflow als Ablauf der Analyse ist teilbar und reproduzierbar. Damit wird über die Analyse auch das methodische Vorgehen für andere erschließbar. Sie können als Ro-Crate Format exportiert und in weitere Repositorien leicht übernommen werden. Das Beispiel eines algorithmisch unterstützten Textvergleichs mit ungewöhnlichem Zeichensatz zeigt, wie FAIRes Forschungsdatenmanagement in den Humanities praktischen Nutzen hat.

Danksagungen

Wir bedanken uns für die finanzielle Unterstützung durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Nationalen Forschungsdateninfrastruktur – NFDI 46/1 – 501864659164 – NFDI4BioImage. Darüber hinaus wurden wir von DataPLANT 442077441 durch die Nationale Forschungsdateninitiative (NFDI 7/1), de.NBI-RBC und de.KCD unterstützt.

Beiträge

Daniela Schneider und Jan Leendertse haben beide gleichwertig als Co-Autoren den obigen Text verfasst.

Interessenkonflikte

Keine.

Literaturverzeichnis

- Hiltemann, Saskia, Helena Rasche, Simon Gladman, Hans-Rudolf Hotz, Delphine Larivière, Daniel Blankenberg, Pratik D. Jagtap u. a. 2023. „Galaxy Training: A powerful framework for teaching!“ Herausgegeben von Francis Ouellette. *PLOS Computational Biology* 19 (1): e1010752. <https://doi.org/10.1371/journal.pcbi.1010752>.
- Schneider, Daniela. 2025. *Text-Mining Differences in Chinese Newspaper Articles*. Tutorial. Besucht am 21. Mai 2025. <https://gxy.io/GTN:T00530>.
- Schneider, Daniela und Jan Leendertse. 2025. *Open Science in the Humanities: Galaxy as a Flexible and yet Reproducible Platform*. Poster. Heidelberg. Besucht am 21. Mai 2025. <https://doi.org/10.11588/HEIDOK.00036288>.
- The Galaxy Community, Linelle Ann L Abueg, Enis Afgan, Olivier Allart, Ahmed H Awan, Wendi A Bacon, Dannon Baker u. a. 2024. „The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update“. *Nucleic Acids Research* 52 (W1): W83–W94. <https://doi.org/10.1093/nar/gkae410>.