

Toolbeitrag: Gephi

Mareike Schumacher  ¹

1. Universität Regensburg

forTEXT

Thema:	Netzwerkanalyse	DOI:	10.48694/fortext.3783
Jahrgang:	1	Ausgabe:	6
Erscheinungsdatum:	2024-08-30	Erstveröffentlichung:	2019-02-18 auf fortext.net
Lizenz:	  		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.



Der Workflow von Gephi: Tabellen- oder Netzwerkdokumente hochladen oder selbst eine Netzwerk-Tabelle anlegen, das Layout des Netzwerks wählen und anpassen, in der Vorschau Feinheiten bearbeiten und Netzwerk als PDF, SVG oder PNG Datei herunterladen.

- **Systemanforderungen:** Für MacOS, Windows und Linux offline nutzbar, Java 7 oder 8 muss installiert sein
- **Stand der Entwicklung:** Gephi befindet sich derzeit noch in der Weiterentwicklung, letztes Release war am 24.9.2017
- **Herausgeber:** Open Source Software, deren Entwicklung von einem Konsortium geleitet wird
- **Lizenz:** kostenfrei unter Open Source Lizenz nutzbar
- **Weblink:** <https://gephi.org/>
- **Im- und Export:** Eine Reihe gängiger Graph-Formate, sowie **CSV** und Datenbanken können importiert werden. Die Daten bleiben unverändert und können in denselben Formaten wieder heruntergeladen werden. darüber hinaus können die in Gephi erstellen Visualisierungen als **SVG**, **PNG** oder **PDF** Datei heruntergeladen werden.
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann Gephi eingesetzt werden?

Gephi dient zur Visualisierung (Horstmann und Stange 2024) von Netzwerken (Schumacher 2024) und kann darum besonders gut bei Fragestellungen nach netzwerkartigen Konstellationen und Strukturen eingesetzt werden. Dazu gehören Fragen wie: Welche Figuren werden innerhalb eines Werkes besonders häufig in engem Zusammenhang genannt? Welche Autoren einer Epoche haben einen ähnlichen Stil? Wie sah der briefliche Austausch von Autoren und Mäzenen in einer Epoche (z. B. im 18. Jh.) aus?

2. Welche Funktionalitäten bietet Gephi und wie zuverlässig ist das Tool?

Funktionen:

- Visualisierung von Netzwerken in unterschiedlichem Layout
- Interaktive Echtzeit-Anpassung der Visualisierung durch den Nutzer bzw. die Nutzerin
- Auswahl der gängigsten statistischen Modelle als Basis der Netzwerkanalyse
- Diachrone Darstellung von Veränderungen von Netzwerken
- Kartografie von Netzwerken
- Dynamische Filter

- Tabelleneditor
- Integration von Plugins

Zuverlässigkeit: Bei der Nutzung von Gephi können in seltenen Fällen Fehler oder auch manchmal Systemabstürze vorkommen. Die Anpassbarkeit an verschiedene Layouts, die Auswahl der Basis-Statistiken und die Option zwischen Netzwerk- und Datenansicht dynamisch zu wechseln, ermöglicht einen sehr umfassenden Einblick in die Daten- und Berechnungsbasis. Auf inhaltlicher Ebene kann Gephi darum als sehr zuverlässig bewertet werden.

3. Ist Gephi für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	–
Leichter Einstieg	–
Handbuch vorhanden	✓
Handbuch aktuell	✓
Tutorials vorhanden	✓
Erklärung von Fachbegriffen	teilweise
Gibt es eine gute Nutzerbetreuung?	✓

Gephi hat sehr viele Funktionalitäten und diese Komplexität spiegelt sich auch in der grafischen Benutzeroberfläche (vgl. **GUI**) wieder. Die Bedienung ist darum nicht unbedingt intuitiv, kann aber anhand der vorhandenen Tutorials recht leicht erlernt werden. In den Gephi-Tutorials steht die Bedienung der Applikation im Vordergrund. So wird nur selten die Bedeutung einer Benennung und die dahinter liegende Technik erklärt und stattdessen immer die Auswirkung auf die Visualisierung am Beispiel hervorgehoben.

4. Wie etabliert ist Gephi in den (Literatur-)Wissenschaften?

Gephi ist in den digitalen Geisteswissenschaften sehr etabliert und wird in vielen Veröffentlichungen genannt. Weniger digital arbeitende Geisteswissenschaftler*innen setzen Gephi derzeit eher selten ein.

5. Unterstützt Gephi kollaboratives Arbeiten?

Nein, Gephi sieht nicht vor, dass Visualisierungen kollaborativ erstellt werden.

6. Sind meine Daten bei Gephi sicher?

Ja. Gephi wird auf dem eigenen Rechner offline ausgeführt. Es findet keine Angabe personenbezogener Daten statt. Es werden keine kompletten Texte bei Gephi eingespeist, sondern nur bearbeitete Datenwerte. Die Nutzung von Gephi ist also auch aus urheberrechtlicher Sicht unproblematisch.

Externe und weiterführende Links

- Gephi-Webseite: <https://web.archive.org/save/https://gephi.org/> (Letzter Zugriff: 01.08.2024)
- Gephi auf GitHub: <https://web.archive.org/save/https://github.com/gephi/gephi> (Letzter Zugriff: 01.08.2024)

Bibliographie

- Eder, Maciej. 2017. Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities* 32, Nr. 1: 50–64. doi: 10.1093/llc/fqv061, (zugegriffen: 13. November 2018).
- Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, <https://fortext.net/routinen/methoden/textvisualisierung>.
- Schumacher, Mareike. 2024. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, <https://fortext.net/routinen/methoden/netzwerkanalyse>.

Glossar

- Annotation** Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch **Machine-Learning-Verfahren** durchgeführt wird. Ein klassisches Beispiel ist das automatisierte **PoS-Tagging** (Part-of-Speech-Tagging), welches oftmals als Grundlage (**Preprocessing**) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.
- Browser** Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.
- Commandline** Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (**GUI**) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.
- CSV** CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.
- GUI** GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der **Commandline** zu umgehen.
- HTML** HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von **Webbrowsern** dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche **Metainformationen** enthalten, die auf einer Webseite selbst nicht ersichtlich sind.
- Lemmatisieren** Die Lemmatisierung von Textdaten gehört zu den wichtigen **Preprocessing**-Schritten in der Textverarbeitung. Dabei werden alle Wörter (**Token**) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.
- Machine Learning** Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.
- Markup (Textauszeichnung)** Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch **Auszeichnungssprachen** wie **XML** implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.
- Markup Language** Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. **HTML**, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch **Annotationen** durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.
- Metadaten** Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch **Annotationen** bzw. **Markup** sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.
- Named Entities** Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.
- OCR** OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

- PDF** PDF steht für *Portable Document Format*. Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein **OCRter** Scan oder ein am Computer erstellter Text.
- POS** PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- **Tagging** beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist ein wichtiger **Preprocessing**-Schritt, beispielsweise für die Analyse von **Named Entities**.
- Preprocessing** Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden **lemmatisiert**.
- SVG** SVG steht für *Scalable Vector Graphics* und ist ein freies, standardisiertes Dateiformat, das Bilddateien bezeichnet, die als 2D-Vektorgrafiken größenunabhängig reproduziert werden können. Bei SVG-Dateien wird im Gegensatz zu anderen Bildgrafiken somit die Auflösung der Abbildung beim Vergrößern nicht schlechter. Es basiert auf den Strukturen von **XML** und wird dazu verwendet, Bilddaten zu repräsentieren.
- TEI** Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch **XML** und **Markup**).
- Type/Token** Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.
- XML** XML steht für *Extensible Markup Language* und ist eine Form von **Markup Language**, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das **TEI-XML**.