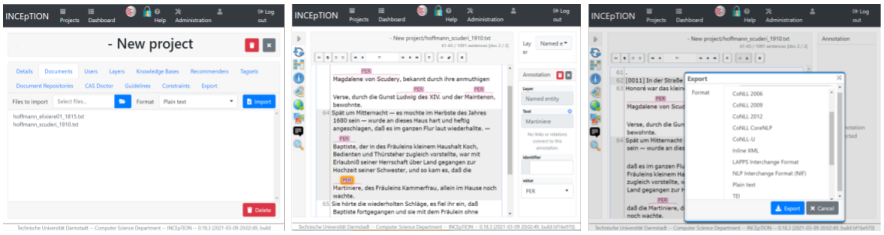
Toolbeitrag: INCEpTION

Mareike Schumacher [1]
Kristina Becker
1. Universität Regensburg

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3762
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2021-04-05 auf fortext.net
Lizenz:	CC BY SA		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 INCEpTION: Über Settings im Dashboard-Menü können die zu annotierenden Dokumente hochgeladen werden. Im Menü des Bereichs Annotation wird der jeweilige Text annotiert. Nachdem dieser finalisiert wurde, kann er in einem beliebigen Format exportiert werden.

- **Systemanforderungen:** Das Tool wird als installierbares Softwarepaket angeboten. Nutzer*innen müssen es auf dem eigenen Laptop, PC oder Server installieren.
- **Stand der Entwicklung:** INCEpTION ist 2016 als von der DFG gefördertes Projekt gestartet.
- **Herausgeber:** INCEpTION wird derzeit am Ubiquitous Knowledge Processing (UKP) Lab der Technischen Universität Darmstadt entwickelt.
- **Lizenz:** Kostenfreie Open Source-Nutzung unter Apache 2.0 Lizenz
- **Weblink:** https://inception-project.github.io/
- **Im- und Export:** Import und Export in den Formaten CoNLL-Formate, TEI XML, Plain Text (vgl. Reintext-Version), UIMA, WebAnno.
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann INCEpTION eingesetzt werden?

INCEpTION ist ein Tool zur manuellen und automatisierten Annotation und kann darum für eine Vielzahl literaturwissenschaftlicher Fragestellungen eingesetzt werden. Besonders geeignet ist es für Ansätze, bei denen vordefinierte und klar operationalisierte Kategorien genutzt werden. So kann das Tool während der manuellen Annotation lernen, welche Indikatoren im Text auf welche Weise markiert werden sollen und kann Vorschläge dafür generieren. Eine Fragestellung, die einen solchen Ansatz leiten könnte, wäre z.B. „Welche realweltlichen Orte werden in Erzähltexten des Realismus erwähnt und welche Städte und/oder Landschaften können als literarische Hotspots dieser Epoche ausgemacht werden?“.

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

Funktionen:

- Anlegen von Projekten, in denen die zu annotierenden Dokumente hochgeladen werden und eigene Annotationsschemata angelegt oder voreingestellte Kategorien genutzt werden können
- Kollaboratives Annotieren inklusive Vergleich und Korrektur zwischen den Annotator*innen
- Überlappende sowie Mehrfachannotation möglich
- (Halb-) automatisches Annotieren auf Basis von Machine-Learning-Technologien (vgl. Machine Learning)
- Abfragen (vgl. Query) (INCEpTION nutzt dazu eine Corpus Query Language oder kurz CQL)

Zuverlässigkeit: Das Tool läuft sehr zuverlässig.

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

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

INCEpTION ist ein Tool, das viele Funktionen zur Annotation bietet. Die Benutzeroberfläche ist in mehrere Module unterteilt. Für den Einstieg in die Bereiche Curation und Monitoring, sind grundlegende Kenntnisse der Konzepte nötig - das Einlesen in das Benutzerhandbuch wird hierfür empfohlen. Auch ist das Anlegen eines individuellen Tagsets (vgl. Tagset) sehr komplex und erfordert eine genauere Kenntnis des Tools.
Der Kontakt zu den Nutzer*innen ist dem INCEpTION-Team besonders wichtig und auf Anfragen wird in der Regel in kurzer Zeit reagiert. Nutzer*innen können für Support-Anfragen, Wünsche nach neuen Funktionen oder Meldungen von Fehlern Issues über GitHub einreichen. Sie können außerdem mittels Mailingliste oder Chat Fragen zum Tool stellen oder Feedback geben.

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

INCEpTION ist ein noch vergleichsweise neues Tool, wurde aber bereits in einer Reihe wissenschaftlicher Projekte eingesetzt, die zum großen Teil einen linguistischen Schwerpunkt haben. Eine Übersicht über Projekte, die mit dem Tool arbeiten stellt INCEpTION auf der Webseite bereit. In den (digitalen) Literaturwissenschaften wird es z.B. im Text Mining-Projekt MiMoText eingesetzt. In der methodenorientierten Auflistung und Beschreibung geisteswissenschaftlicher Tools „Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften“ (Frey-Endres und Simon 2021) wird auch INCEpTION gelistet, in zwei im Jahr 2019 und 2021 durchgeführten quantitativ ausgerichteten Betrachtungen der Nutzung von Tools in der internationalen Digital-Humanities-Community findet INCEpTION keine Erwähnung (Barbot u. a. 2019; Fischer u. a. 2021).

5. Unterstützt INCEpTION kollaboratives Arbeiten?

Ja, es kann kollaborativ gearbeitet werden. Zudem bietet das Tool eine automatische Agreement-Berechnung an.

6. Sind meine Daten bei INCEpTION sicher?

Da die Nutzer*innen INCEpTION als Paket herunterladen, verbleiben projektbezogene Daten bei ihnen und werden nicht an das INCEpTION-Team gesendet. Manche Universitäten oder andere Institutionen betreiben eigene INCEpTION-Instanzen und bieten deren Nutzung als Service an. Nutzer*innen dieser Services sollten sich bei Fragen zur Sicherheit an die jeweiligen Betreiber*innen wenden. INCEpTION bietet außerdem die Möglichkeit, Exporte der Annotationsprojekte durchzuführen um sie als Backup außerhalb der Anwendung zu sichern. INCEpTION kann autark ohne Internetanbindung betrieben werden und ermöglicht so eine Abschottung des Tools und der Daten gegen unbefugte Zugriffe.

Personenbezogene Daten:
Bei lokalen Installationen (z.B. auf dem eigenen PC) sind beim Erstellen eines Nutzungskontos ein Name (oder Pseudonym) und ein Passwort erforderlich, welche innerhalb der Anwendung gespeichert werden. Bei lokalen Installationen werden Benutzername und Passwort lediglich lokal gespeichert. Bei einer Serverinstallation werden sie auf dem Server gespeichert. Passwörter werden verschlüsselt abgelegt.
Betreiber*innen der INCEpTION-Instanz können zustimmen, anonyme Statistiken an das INCEpTION-Team zu übermitteln (z.B. die genutzte Version von INCEpTION, Betriebssystem, Anzahl von Benutzeraccounts). Das INCEpTION-Team verwendet diese Statistik, um die Verbreitung der Software und ihrer Versionen zu verfolgen sowie die Entwicklung zu verbessern. Es wird über die Art der erhobenen anonymen Daten bei Inbetriebnahme der Instanz informiert. Dem kann bei der Erhebung direkt oder zu einem beliebigen späteren Zeitpunkt widersprochen werden.

Urheberrechtlich geschützte Daten:
Texte werden innerhalb der Anwendung hochgeladen und in einem geschützten Login-Bereich verwaltet.

Externe und weiterführende Links

- INCEpTION Webseite: https://inception-project.github.io (Letzter Zugriff: 12.06.2024)
- INCEpTION Dokumentation: https://web.archive.org/save/https://inception-project.github.io/releases/33.2/docs/user-guide.html (Letzter Zugriff: 12.06.2024)
- INCEpTION Downloadbereich: https://web.archive.org/save/https://inception-project.github.io/ (Letzter Zugriff: 12.06.2024)
- INCEpTION Übersicht über Projekte, in denen das Tool eingesetzt wird: https://web.archive.org/save/https://inception-project.github.io/use-cases/ (Letzter Zugriff: 12.06.2024)
- INCEpTION-Support: https://web.archive.org/save/https://github.com/inception-project sowie https://web.archive.org/save/https://groups.google.com/g/inception-users und https://web.archive.org/save/https://gitter.im/inception-project/Lobby (Letzter Zugriff: 12.06.2024)
- INCEpTION auf GitHub: https://web.archive.org/save/https://github.com/inception-project/inception/issues (Letzter Zugriff: 12.06.2024)
- INCEpTION auf YouTube: https://web.archive.org/save/https://www.youtube.com/channel/UC3sUTFFPYg0aWmZRag45yJw (Letzter Zugriff: 12.06.2024)
- INCEpTION Handbuch für Administrator*innen: https://web.archive.org/save/https://inception-project.github.io/releases/33.2/docs/admin-guide.html (Letzter Zugriff: 12.06.2024)
- MiMoText: https://web.archive.org/save/https://mimotext.uni-trier.de/aktuelles (Letzter Zugriff: 12.06.2024)

Bibliographie

Barbot, Laure, Frank Fischer, Yoann Moranville und Ivan Pozdniakov. 2019. Which DH Tools Are Actually Used in Research? *Weltliteratur*. 6. Dezember. https://weltliteratur.net/dh-tools-used-in-research/ (zugegriffen: 5. April 2021).

Fischer, Frank, Manuel Burghardt, Jan Luhmann, Laure Barbot, Yoann Moranville und Alireza Zarei. 2021. Die Werkbänke der Digital Humanities: Zur Rolle von Tools und Software für die Forschungsarbeit. 26. März. doi: 10.5281/zenodo.4639228, https://zenodo.org/record/4639228#.YGq-XHUzaUk (zugegriffen: 5. April 2021).

Frey-Endres, Marcel und Tobias Simon. 2021. *Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften*. Hg. von Sabine Bartsch, Evelyn Gius, Marcus Müller, Andrea Rapp, und Thomas Weitin. Bd. 2. Working Papers in Digital Philology. Darmstadt. https://tuprints.ulb.tu-darmstadt.de/17850/ (zugegriffen: 1. Februar 2023).

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.

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.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

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.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

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 (Textauszeichung) 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.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

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 of 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.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

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).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

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.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

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.