Toolbeitrag: Protégé

Janina Jacke [1]

1. Christian-Albrechts-Universität zu Kiel

Thema:	Projektkonzeption	DOI:	10.48694/for-text.3806
Jahrgang:	1	Ausgabe:	12
Erscheinungsdatum:	30-11-2024	Erstveröffentlichung:	2020-03-09 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.

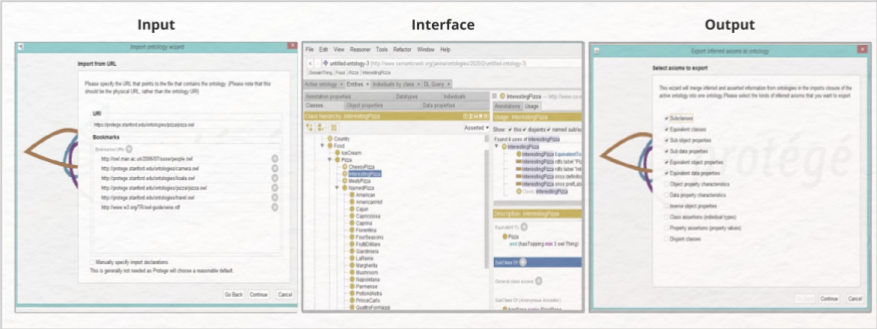

Abb.1: Der Workflow von Protégé Desktop: Vorab: Installation des Tools; Input: Import einer vorhandenen Ontologie im OWL-Format (oder Erstellung einer eigenen Ontologie ohne Import); Interface: Erstellung oder Bearbeitung von Klassen, Eigenschaften etc., ggf. Ableitung von Regeln (Axiomen); Output: Export der erstellten/bearbeiteten Ontologie oder der abgeleiteten Axiome als Ontologie im OWL-Format

- **Systemanforderungen:** Protégé Desktop kann heruntergeladen und mit Windows, Mac-Betriebssystemen oder Linux offline genutzt werden. Die plattformspezifischen Versionen enthalten bereits Java, so dass dieses nicht gesondert installiert werden muss. WebProtégé kann online mit allen großen Browsern (vgl. Browser) genutzt werden. Desktop- und Browser-Version (vgl. Webanwendung) sind kreuzkompatibel.
- **Stand der Entwicklung:** Protégé wurde 1999 erstveröffentlicht, die aktuelle Version ist v.5.5.0 aus 2019. **Herausgeber:** Das Tool wurde am Institut für Medizinische Informatik der Stanford University (anfänglich in Kooperation mit der University of Manchester) entwickelt und ist mittlerweile als Open-Source-Software verfügbar.
- **Lizenz:** kostenfrei nutzbar unter Open-Source-Lizenz (BSD 2-clause license)
- **Weblink:** https://protege.stanford.edu/
- **Im- und Export:** Im- und Export von Ontologien sind in Protégé unter anderem in den Formaten RDF/XML, OWL/XML, Turtle, OBO, LaTeX und JSON möglich.
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann Protégé eingesetzt werden?

Protégé ist ein Tool für die Erstellung komplexer Ontologien (Jacke und Gerstorfer 2024), d. h. es können Kategorien für die Beschreibung eines (beispielsweise literarischen) Gegenstandsbereichs festgelegt, untereinander verknüpft und ausführlich beschrieben werden. Das Tool unterstützt also die wissenschaftliche Modell- bzw. Theoriebildung. Die erstellten Ontologien können außerdem teilweise als Tagsets für die Annotation (Jacke 2024) von Texten verwendet werden. Mögliche literaturwissenschaftliche Fragestellungen lauten: Welche literaturwissenschaftlichen Gattungen und Genres existieren und in welcher Verbindung stehen diese zueinander? Welche Typen unzuverlässigen Erzählens gibt es – und kann ein Erzähler zugleich mimetisch und axiologisch unzuverlässig sein? Welche weiteren Kategorisierungen ergeben sich automatisch, wenn eine Textstelle in einem Gedicht als Metapher eingeordnet wird?

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

Funktion:

- Erstellung von Kategorien zur Beschreibung eines Gegenstandsbereichs
- Beschreibung von Kategorien
- Verknüpfung von Kategorien durch Festlegung von Beziehungen
- Kombination freier Ontologieerstellung mit standardisierten Komponenten
- Import bestehender Ontologien
- Automatische Konsistenzüberprüfung und Schlussfolgerungen für erstellte Ontologien
- Individualisierbare Nutzeroberfläche
- Zahlreiche Dokumentationsmöglichkeiten
- Komplexe Kollaborationsoptionen
- Plugins verfügbar, z. B. für die Visualisierung von Ontologien

Zuverlässigkeit: Protégé funktioniert zuverlässig.

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

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

Protégé ist nur lose an die traditionelle Literaturwissenschaft angebunden – primär ist das Tool auf die Erstellung von Ontologien in naturwissenschaftlichen Disziplinen ausgelegt. Generell wird in den Literaturwissenschaften nur punktuell mit Ontologien bzw. Taxonomien gearbeitet, beispielsweise im Rahmen narratologischer Untersuchungen oder formaler Gedichtanalyse. Dabei spielen allerdings in Protégé verfügbare Optionen wie die Festlegung differenzierter Beziehungen zwischen Kategorien oder das Ziehen logischer Schlüsse auf Basis einer Ontologie in literaturwissenschaftlichen Kontexten gemeinhin keine Rolle.
Obwohl Protégé eine anpassbare grafische Nutzeroberfläche aufweist, ist die Verwendung des Tools für DH-Einsteiger*innen nicht intuitiv: Vorausgesetzt wird Wissen über die Ontologie-Beschreibungssprache OWL (Web Ontology Language), die beispielsweise festlegt, welche Elemente Ontologien aufweisen können (Klassen, Relationen, Attribute, Regeln etc.).
Für die Desktop- und die Web-Variante von Protégé sind Handbücher vorhanden, allerdings setzen diese bereits einiges Vorwissen voraus und sind darüber hinaus – gemessen an den umfangreichen Funktionen von Protégé – recht knapp. Dies gilt insbesondere für das Handbuch für WebProtégé. Dieses verweist zudem teilweise auf Vorträge, die sich auf frühere Protégé-Versionen beziehen, und einige verlinkte Tutorials existieren nicht mehr. Beides lässt darauf schließen, dass das Handbuch nicht gründlich aktualisiert wird. Es sind dennoch einige aktuelle Tutorials verfügbar, die allerdings nicht systematisch die in Protégé zur Verfügung stehenden Funktionen abdecken. Handbuch und Tutorials für die Desktop-Variante sind dagegen systematischer und aktuell.
Fachbegriffe werden in Protégé selbst und in Handbuch bzw. Tutorials teilweise erläutert. Diese Erklärungen setzen jedoch in der Regel auf einer höheren Ebene an, als es für literaturwissenschaftliche Einsteiger*innen in die DH nötig wäre.
Das Protégé-Projekt bietet unterschiedliche Varianten der Nutzerbetreuung an. Zuvorderst zu nennen ist hier eine Mailingliste, über die Fragen zum Tool gestellt werden können, die dann wiederum von anderen Nutzer*innen oder dem Protégé-Entwickler-Team beantwortet werden. Frühere Anfragen, die über diese Liste verschickt worden sind, können im Archiv der Liste eingesehen werden. Dort lässt sich allerdings auch erkennen, dass nicht jede gestellte Frage beantwortet wird. Zudem werden kostenpflichtige Protégé-Kurse und Beratung angeboten.

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

Protégé ist mit ca. 300.000 registrierten Nutzer*innen das am meisten genutzte Tool zur Ontologieerstellung und kommt insbesondere in naturwissenschaftlichen Kontexten zum Einsatz. In den Literaturwissenschaften, auch in den digitalen, findet es bisher kaum Anwendung.

5. Unterstützt Protégé kollaboratives Arbeiten?

Ja, Protégé bietet zahlreiche und komplexe Kollaborationsmöglichkeiten – ebenso wie facettenreiche Optionen der Dokumentation und der Versionierung, die kollaboratives Arbeiten ebenfalls erleichtern. So können Projekte bzw. Ontologien in unterschiedlichen Modi mit anderen registrierten Nutzer*innen geteilt werden. Außerdem werden alle Änderungen an Ontologien gespeichert und können eingesehen werden. Dabei haben Nutzer*innen die Möglichkeit, auch ältere Versionen einer Ontologie herunterzuladen. Alle Elemente der Ontologien können darüber hinaus mit Notizen unterschiedlicher Kategorien versehen werden – beispielsweise mit Kommentaren, Vor- oder Ratschlägen–, die die Kommunikation zwischen Teamkolleg*innen vereinfachen und effizient machen.

6. Sind meine Daten bei Protégé sicher?

Ja. Für den Download der Desktop-Version von Protégé ist es erforderlich, einen Namen und eine Projektbeschreibung anzugeben. Dabei können allerdings beliebige Angaben getätigt werden. Für die Nutzung der Web-Version müssen Name und E-Mail-Adresse angegeben sowie ein Passwort vergeben werden. Die E-Mail wird allerdings nicht (beispielsweise im Rahmen eines Bestätigungsvorgangs) überprüft. Die in WebProtégé erstellten Ontologien sind nur in einem gesicherten Login-Bereich verfügbar. Persönliche und in Protégé erstellte Daten werden nicht weitergegeben – sie werden aber ggf. analysiert, um die angebotenen Dienstleistungen aufrechtzuerhalten und zu verbessern.

Wenn Sie in Protégé bereits existierende Ontologien importieren möchten, sollten Sie darauf achten, dass Sie dazu berechtigt sind, diese Ontologien zu verwenden.

Externe und weiterführende Links

- Protégé Ontology Editor: https://web.archive.org/web/20241106125431/https://protege.stanford.edu/ (Letzter Zugriff: 06.11.24)

Bibliographie

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Jacke, Janina und Dominik Gerstorfer. 2024. Methodenbeitrag: Entwicklung von Kategoriensystemen. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 10. Projektkonzeption (29. November). doi: 10.48694/fortext.3805, https://fortext.net/routinen/methoden/entwicklung-von-kategoriensystemen.

Lohmann, Steffen, Stefan Negru, Florian Haag und Thomas Ertl. 2016. Visualizing Ontologies with VOWL. *Semantic Web* 7, Nr. 4: 399–419.

Maroto, Nava und Amparo Alcina. 2009. Formal description of conceptual relationships with a view to implementing them in the ontology editor Protégé. *Terminology* 15, Nr. 2: 232–257.

Noy, Natalya F. und Deborah L. McGuinness. 2001. Ontology Development 101. A Guide to Creating Your First Ontology. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf (zugegriffen: 6. März 2020).

Schärfe, Henrik. 2004. Narrative Ontologies. *Knowledge Economy Meets Science and Technology - KEST2004*: 19–26.

Song, Dezhao, Christopher G. Chute und Cui Tao. 2012. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies. In: *Proceedings of AMIA Summit on Translational Science*. http://swat.cse.lehigh.edu/pubs/song12a.pdf (zugegriffen: 9. März 2020).

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.

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 (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.

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.

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.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online“ genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

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.