Semantic Publishing

Yannick Lang
Matrikelnummer: 1995498
KInf-Seminar-M
Fakultät Wirtschaftsinformatik & Angewandte Informatik
Otto-Friedrich-Universität Bamberg

Abgabe: February 27, 2025

Abstract

Genuine Semantic Publishing ist ein 2017 von Tobias Kuhn und Michel Dumontier vorgestelltes Konzept, das darauf abzielt, wissenschaftliche Arbeiten maschinenverarbeitbar zu machen. Mit Hinblick auf die stetig steigende Menge an wissenschaftlichen Publikationen und der damit einhergehenden Schwierigkeit, diese zu verarbeiten, stellt sich die Frage, ob die Erstellung und Aktualisierung von Literaturübersichten durch den Einsatz von Genuine Semantic Publishing erleichtert werden kann. In dieser Arbeit wird die Eignung von Genuine Semantic Publishing für Knowledge-Graph-basierte Living Literature Reviews untersucht und mit alternativen, KI-basierten Ansätzen wie SCICERO verglichen.

Inhaltsverzeichnis

1 Einführung

Besonders in sehr dynamischen Forschungsbereichen, wie beispielsweise dem Machine Learning, ist es wichtig, stets auf dem aktuellen Stand der Forschung zu bleiben. Dies ist jedoch aufgrund der hohen Anzahl an wissenschaftlichen Publikationen, die jährlich veröffentlicht werden (Kang et al., 2024), eine Herausforderung; herkömmliche Literaturübersichten können bereits nach wenigen Monaten überholt sein. Eine Möglichkeit, um den Überblick zu behalten, sind sogenannte *Living Literature Reviews* (LLRs). Diese werden regelmäßig aktualisiert und enthalten eine Zusammenfassung der aktuellen Forschungsergebnisse zu einem bestimmten Thema.

In der vorliegenden Arbeit wird zuerst der Anwendungsfall näher erläutert. Anschließend werden zwei Ansätze zur Erstellung und Aktualisierung von LLRs vorgestellt und miteinander verglichen: *Genuine Semantic Publishing* (Kuhn and Dumontier, 2017), beruhend auf von den Autoren selbst erstellten und veröffentlichten, maschinenverarbeitbaren Informationen über Ihre Forschungsbeiträge, und *SCICERO*, ein auf Natural Language Processing und Transformer-Modellen beruhender Ansatz. Dabei wird untersucht, wie gut die beiden Ansätze in der Lage sind, den Anforderungen des intendierten Anwendungsszenarios gerecht zu werden und welche Vor- und Nachteile sie haben.

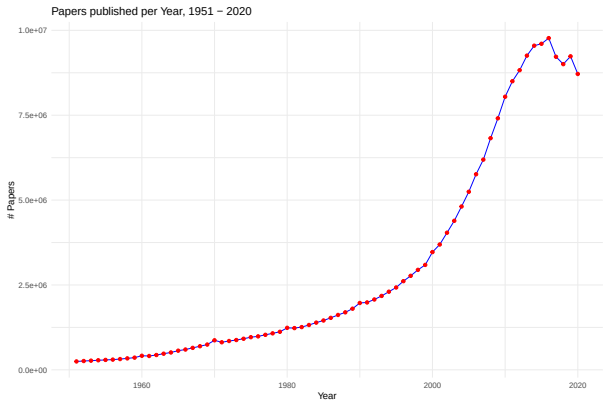

Figure 1: Anzahl der jährlich veröffentlichten wissenschaftlichen Publikationen, 1951-2020. Eigene Darstellung nach Daten von Kang et al. (2023).

2 Living Literature Reviews

Literaturübersichten sind ein wichtiges Instrument, um den Überblick über den aktuellen Stand der Forschung zu behalten. Traditionell werden diese

in Form von wissenschaftlichen Publikationen veröffentlicht, die eine Zusammenfassung der relevanten Arbeiten zu einem bestimmten Thema enthalten. Zusätzlich können auch Empfehlungen oder Metaanalysen ein Bestandteil sein, um die Ergebnisse der verschiedenen Studien zu vergleichen und zu bewerten (Elliott et al., 2017).

Living Literature Reviews Besonders in sehr dynamischen Forschungsbereichen, wie beispielsweise dem Machine Learning, können Literaturübersichten zwar besonders hilfreich sein, aber auch schnell veralten. Eine Möglichkeit, um den Überblick zu behalten, sind sogenannte *Living Literature Reviews* (LLRs, auch: Living Systematic Reviews), also Literaturübersichten, die regelmäßig an den aktuellen Forschungsstand angepasst werden (Wijkstra et al., 2021). Wann solche Aktualisierungen stattfinden, kann dabei variieren, etwa in regelmäßigen Abständen oder bei wichtigen neuen Entwicklungen (Cochrane).

Das Anlegen und Aktualisieren von Literaturübersichten ist jedoch aufwändig und erfordert eine Menge manueller Arbeit. Neben dem Auffinden relevanter Publikationen müssen diese gelesen und bewertet werden, um zu entscheiden, ob sie in die Übersicht aufgenommen werden sollen. Anschließend müssen die Ergebnisse der verschiedenen Studien zusammengefasst und bewertet werden, um eine kohärente Übersicht zu erstellen. Erst dann kann die Übersicht geschrieben und veröffentlicht werden (Brack et al., 2022).

Weil die Menge an wissenschaftlichen Publikationen, die jährlich veröffentlicht werden, stetig steigt, ist es schwierig, diesen Prozess manuell zu bewältigen (Ronzano and Saggion, 2016). Andererseits ist auch eine maschinelle Unterstützung mit Hindernissen verbunden, da Publikationen traditionell in natürlicher Sprache verfasst sind und in der Regel als PDF veröffentlicht werden und daher nicht ohne Weiteres maschinell verarbeitet werden können (Brack et al., 2022).

Um den Aufwand —gerade für die Aktualisierung —zu verringern, gibt es Bestrebungen, maschinelle Unterstützung bei der Arbeit mit großen Mengen wissenschaftlicher Publikationen zu ermöglichen. So können beispielsweise Nanopublikationen (Groth et al., 2010) verwendet werden, wodurch zu einer Literaturübersicht leicht neue Publikationen hinzugefügt werden können, ohne die gesamte Übersicht neu verfassen zu müssen (Wijkstra et al., 2021). Dadurch werden auch innovativere Formen von Literaturübersichten möglich, die beispielsweise interaktiv sind und so den Nutzern erlauben, die Übersicht nach ihren eigenen Bedürfnissen zu filtern oder Veränderungen im Zeitverlauf zu betrachten (Wijkstra et al., 2021). Damit sind solche Reviews auch nicht an traditionelle Publikationsformen gebunden, sondern können neben Webseiten oder interaktiven Grafiken beispielsweise auch selbst in maschinenlesbaren Datenformaten veröffentlicht werden (Jiomekong et al., 2024).

Wissensgraphen in Living Literature Reviews Wissensgraphen könnten an mehreren Stellen in den Prozess der Erstellung von Living Literature Reviews eingebunden werden. So könnten sie beispielsweise dazu genutzt werden, Literatur ausfindig zu machen, die für die Erstellung eines Reviews relevant ist. Dabei könnten Wissensgraphen helfen, indem sie die Beziehungen zwischen verschiedenen Publikationen und Forschungsfeldern abbilden und so die Suche nach verwandten Arbeiten erleichtern. Einerseits könnte das über die Verwendung der Zitationen und Referenzen geschehen. Durch einen Fokus auf die in den Publikationen verwendeten Begriffe und Konzepte könnten Wissensgraphen aber auch dabei helfen, Publikationen zu finden, die sich mit ähnlichen Themen beschäftigen, aber nicht gegenseitig zitieren, etwa, weil den Autoren selbst die andere Arbeit nicht bekannt ist (Brack et al., 2021). Wissensgraphen können darüber hinaus auch zur Thesengenerierung und -validierung genutzt werden, um so die Ergebnisse der Literaturrecherche zu strukturieren und gegebenenfalls anzupassen, wenn neue relevante Inhalte hinzugefügt werden (Dessí et al., 2022). Schließlich können die in Wissensgraphen oft enthaltenen Herkunftsinformationen verwendet werden, um die Qualität der gefundenen Publikationen zu bewerten, indem beispielsweise die Reputation der Autoren oder der Publikationsorte berücksichtigt wird.

Als Kernfrage verbleibt, woher die zugrundeliegenden maschinenverarbeitbaren Daten kommen. In den folgenden Kapiteln werden hierfür zwei potentielle Lösungsansätze vorgestellt.

3 Genuine Semantic Publishing

Im Folgenden wird das von Kuhn and Dumontier (2017) vorgestellte Konzept des Genuine Semantic Publishing erläutert und dessen potentielle Einsatzgebiete für Living Literature Reviews disku-

tiert.

3.1 Motivation

Die Autoren argumentieren, dass Semantic Publishing in früheren Arbeiten oft lediglich als Anreicherung von bestehenden Publikationen mit (Meta-)Daten betrachtet wurde. Die gängige Definition von Semantic Publishing als *alles, was die Bedeutung veröffentlichter Artikel aufwertet, automatisches Auffinden von Artikeln oder Verbinden von Artikeln mit verwandtem Inhalt ermöglicht, Zugang zu Daten in verwertbarer Form verschafft oder Wiederverwendung von Daten zwischen Artikeln erleichtert* (im Original: "anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers") (Kuhn and Dumontier, 2017) weisen die Autoren jedoch als gleichzeitig über- und unterspezifisch zurück. Ihrer Einschätzung nach würde das Hinzufügen von Schlagwörtern bereits ausreichen, um unter die Definition zu fallen (ermöglicht das Verbinden inhaltlich verwandter Artikel). Andererseits könnten Publikationen, die zwar Forschungsergebnisse als semantische Repräsentationen abbilden, aber auf den Fließtext verzichten, dieser Definition nach nicht als Semantic Publications bezeichnet werden.

Um diese Schwächen zu adressieren, schlagen die Autoren vor, den gesamten Publikationsprozess zu überdenken. Um zu verdeutlichen, dass das damit Beschriebene näher an der ursprünglichen Grundidee von Semantic Publishing ist, nennen Sie dieses Konzept Genuine Semantic Publishing. Als semantisch verstehen die Autoren dabei *eine auf formaler Logik basierende Darstellung der Bedeutung des Inhalts* (im Original: "carrying a formal logic-based representation of the content's meaning").

3.2 Definition

Um die Abgrenzung zu anderen Ansätzen zu realisieren, definieren die Autoren die folgenden fünf Prinzipien, die eine Publikation erfüllen muss, um als *genuine semantic* zu gelten: (1) machine interpretable, (2) essential coverage, (3) authenticity, (4) primary component und (5) fine-grained & light-weight.

Machine interpretable Die Interpretation, nicht nur das Lesen, des Inhalts der Publikation muss durch eine Maschine möglich sein.

Essential coverage Mindestens die Kerninformationen der Publikation müssen in semantischer Repräsentation vorliegen. Zusammen mit dem Prinzip der *Machine interpretability* erlaubt dies eine automatisierte Verarbeitung der Inhalte, zum Beispiel ein Zusammenfassen und Schlussfolgern auf der Grundlage wissenschaftlicher Erkenntnisse (Kuhn and Dumontier, 2017). Dass dieser Aspekt von besonderer Bedeutung für den Einsatz in Living Literature Reviews ist, wird in Abschnitt 3.3 diskutiert.

Authenticity Die semantischen Repräsentationen müssen von Experten —im Regelfall den Autoren selbst —genehmigt, besser noch erstellt und dann veröffentlicht werden. Das soll verhindern, dass von privaten Firmen erstellte, unfreie Daten von der Definition erfasst sind.

Primary component Die semantischen Inhalte dürfen weder nachträglich veröffentlicht werden noch sollen sie ein bloßer Anhang oder von anderen Teilen der Publikation abhängig sein.

Fine-grained & light-weight Einzelne Inhalte sollen leicht ergänzt oder korrigiert werden können.

3.3 Zusammenhang mit Living Literature Reviews

Die Autoren betrachten Genuine Semantic Publishing als Möglichkeit, "komplexe Fragen zu beantworten oder interaktive Karten der wissenschaftlichen Literatur zu erstellen" (im Original: "to answer complex questions or produce interactive science maps") (Kuhn and Dumontier, 2017). Diese Vision ist eng mit dem Konzept der Living Literature Reviews verbunden, die regelmäßig aktualisiert werden und eine Zusammenfassung des aktuellen Forschungsstandes zu einem bestimmten Thema bieten. So erkennen die Autoren selbst das Potential, einen schnellen und einfachen, aber trotzdem umfassenden und zutreffenden Überblick über den aktuellen Stand der Forschung zu ermöglichen.

4 SCICERO

Ähnlich wie Kuhn and Dumontier (2017) verfolgen auch Dessí et al. (2022) das Ziel, die Sondierung, Verbindung und Analyse wissenschaftlicher Publikationen zu erleichtern. Auch Sie sind der Ansicht, dass angesichts der Menge

an wissenschaftlichen Publikationen, die jährlich veröffentlicht werden, maschinelle Unterstützung notwendig ist, um diese Ziele zu erreichen.

4.1 Motivation

Im Gegensatz zu Genuine Semantic Publishing, bei dem ganz Explizit die Autoren selbst maschinenverarbeitbare Informationen über ihre Forschungsbeiträge erstellen und veröffentlichen müssen, ist SCICERO ein extraktiver Ansatz, beruhend auf auf Natural Language Processing und Transformer-Modellen.

Gegenüber von Menschen annotierten Verfahren bietet das den Vorteil, schnell auf große Mengen an Texten angewendet werden zu können. Der Vorteil gegenüber anderen KI-basierten Ansätzen ist, dass SCICERO nicht nur für die Analyse einzelner Publikationen, sondern explizit für den Vergleich und die Verknüpfung von Publikationen entwickelt wurde. Hierbei ergeben sich für maschinelle Verfahren zwei Herausforderungen: (1) die Identifikation von relevanten Entitäten und deren Relationen innerhalb eines Textes und (2) die Verknüpfung dieser Informationen zwischen Texten. Insbesondere bei letzterem ist es wichtig, dass Entitäten zusammengefasst werden, die in verschiedenen Texten unterschiedlich benannt werden, aber dasselbe Konzept beschreiben.

4.2 Vorgehen

SCICERO besteht aus drei Schritten: (1) Extraktion, (2) Entitäts- und Relationsbearbeitung und (3) Validierung.

Extraktion Im ersten Schritt werden mehrere verschiedene Methoden zur Entitäts- bzw. Relationsextraktion angewendet. Neben klassischen Methoden des Natural Language Processing, wie zum Beispiel Part-Of-Speech Tagging, kommen hier auch auf Transformer-Modellen basierende Verfahren wie DyGIEpp (Wadden et al., 2019) zum Einsatz. Von welcher Methode eine Relation extrahiert wurde, wird als Teil der Herkunftsinformationen gespeichert.

Entitäts- und Relationsbearbeitung Anschließend werden Entitäten dedupliziert und Relationen auf eine Ontologie zurückgeführt. Um die Menge der gefundenen Entitäten zu reduzieren, kommen verschiedene Methoden zum Einsatz. Unter anderem werden Entitäten mit wenig Informationsgehalt entfernt und Abkürzungen auf ihre Langform zurückgeführt. Außerdem kommen aus dem Information Retrieval bekannte Verfahren wie das Entfernen sogenannter stop-words und die Lemmatisierung, also die Bildung von Grundformen, zum Einsatz (Ceri et al., 2013). Anschließend werden die Entitäten auf kanonische Formen zurückgeführt. Neben Transformern werden hier auch externe Quellen wie DBpedia und Wikidata verwendet, die unter anderem Informationen über Synonyme und alternative Bezeichnungen enthalten.

Auch werden die gefundenen Relationen normalisiert und inhaltlich ähnliche Beziehungen zusammengefasst.

Validierung Im letzten Schritt werden die extrahierten Relationen validiert. Hier soll einerseits sichergestellt werden, dass sie konsistent mit anderen Relationen sind, die das System als korrekt annimmt, und andererseits, dass die Entitäten einer Relation dem —gemäß der verwendeten Ontologie —erwarteten Typ entsprechen (z.B. Material, Aufgabe oder Metrik). Für die erste Validierungsform wird die Tatsache ausgenutzt, dass die Wahrscheinlichkeit, dass eine Relation korrekt ist, mit der Anzahl der Artikel korreliert, aus denen die Relation extrahiert wurde. Relationen mit hohem Support, die also aus vielen Artikeln stammen, werden als korrekt angenommen und zum fine-tuning eines Transformer-Modells verwendet. Für Relationen mit wenig Support entscheidet dieses Modell dann, ob sie konsistent mit den als korrekt angenommenen Relationen sind. Schlussendlich erfolgt eine regelbasierte Validierung, bei der die Entitäten einer Relation auf ihre Typen überprüft werden, beispielsweise kann eine Methode ein Material benutzen (`<methodX usesMaterial materialY>`), nicht aber umgekehrt. Die verfügbaren Typen sind dabei in einer Ontologie festgelegt.

4.3 Zusammenhang mit Living Literature Reviews

Die Autoren haben SCICERO explizit für die Generierung von Wissensgraphen zu wissenschaftlichen Forschungsfeldern entwickelt. So haben Sie aus 6.7 Millionen wissenschaftlichen Publikationen aus dem Bereich der Informatik den Wissensgraphen *CS-KG* (Dessí et al., 2022) erstellt. Mit einem ähnlichen Ansatz hatten die Autoren zuvor bereits einen Wissensgraphen zu dem Subfeld der Künstlichen Intelligenz erstellt (Dessì et al., 2020) und damit die Anwendbarkeit auch für

kleinere Felder demonstriert.

Je nach intendiertem Forschungsfeld kann die verwendete Ziel-Ontologie angepasst werden. Statt der auf Informatik spezialisierten Computer Science Knowledge Graph Ontology können beliebige Alternativen —beispielsweise *Gene Ontology* oder *Mathematics Subject Classification* (Dessí et al., 2022) —verwendet werden. Entsprechend erfordert dies dann auch Anpassungen an den Extraktionsmodulen und Validierungsregeln.

5 Vergleich

Verwendete Technologien Sowohl Genuine Semantic Publishing als auch SCICERO verwenden das vom World Wide Web Consortium (W3C) entwickelte Resource Description Framework (RDF) zur Modellierung der semantischen Daten. Während Genuine Semantic Publishing auf die Beschreibung der Ergebnisse durch die Autoren setzt, gegebenenfalls auch ganz ohne einen begleitenden Text, verwendet SCICERO Natural Language Processing und Transformer-Modelle zur Extraktion semantischer Informationen aus bestehenden, textbasierten Publikationen.

Beide Ansätze verwenden Ontologien, um die extrahierten Informationen zu strukturieren. SCICERO verwendet dabei in der Originalform eine explizit für Informatik ausgelegte Ontologie. Wie im vorangegangenen Kapitel beschrieben, ist aber mit gewissen Anpassungen auch die Verwendung einer anderen Ontologie möglich, beispielsweise für einen enger begrenzten Teilbereich der Informatik oder auch für komplett andere Forschungsfelder. Genuine Semantic Publishing ist währenddessen größtenteils agnostisch gegenüber der Wahl konkreter Technologien, solange diese dazu beiträgt, die fünf Forderungen zu erfüllen. Entsprechend empfehlen die Autoren keine spezifische Ontologie, nennen aber Beispiele wie CiTO (Citation Typing Ontology, (Shotton, 2010)) und SKOS (Simple Knowledge Organisation, (Miles et al., 2005)).

Nachvollziehbarkeit Beide Ansätze enthalten Herkunftsinformationen (im Original: *provenance*) über die Forschungsergebnisse. Bei Genuine Semantic Publishing beziehen sich diese Informationen auf die Autoren selbst, die die semantischen Daten erstellt haben. Dadurch soll ein Anreiz geschaffen werden, die Daten korrekt und vollständig zu erstellen (Kuhn and Dumontier, 2017). Außerdem stellt das sicher, dass die Intention der Autoren korrekt wiedergegeben wird, da diese die Daten direkt anlegen. Dadurch entfällt der Umweg über potentiell weniger eindeutigen Text in natürlicher Sprache, der anfälliger für Missverständnisse oder Unklarheiten ist.

Bei SCICERO sind neben den Artikeln, aus denen die Informationen extrahiert wurden, auch die Modelle und Methoden, die zur Extraktion verwendet wurden, Teil der Herkunftsinformationen. Wie bei allen KI-basierten Ansätzen kann es aber natürlich auch hier zu Fehlern oder Halluzinationen kommen. Diese können an mehreren Stellen im Prozess auftreten. Dass in der Extraktionsstufe beispielsweise Entitäten oder Relationen erkannt werden, die so gar nicht im Text enthalten sind, ist ein ganz natürlicher Aspekt von SCICERO, weswegen ja auch am Schluss noch eine Validierung erfolgt, in der inkorrekte Daten erkannt und entfernt werden sollen. Aber auch diese Validierung ist natürlich nicht perfekt, und es kann auch hier zu Fehlern kommen, indem entweder korrekte Daten verworfen oder inkorrekte aufgenommen werden. Genauso können in der Harmonisierung der Daten Fehler auftreten, wenn beispielsweise zwei Entitäten, die eigentlich das gleiche Konzept beschreiben, nicht zusammengeführt werden oder zwei unterschiedliche Entitäten irrtümlich zusammengeführt werden. In der Evaluation von SCICERO finden die Autoren zwar heraus, dass bessere Werte als mit anderen Ansätzen erreicht werden, aber auch hier gibt es mit Precision-Raten von ca. 75% noch deutliches Verbesserungspotential.

Realisierbarkeit Eines der Hauptargumente, die bei dem beschriebenen Anwendungsszenario gegen Genuine Semantic Publishing sprechen, ist die Realisierbarkeit. Während SCICERO jederzeit schon eingesetzt werden könnte, um Wissensgraphen für Forschungsfelder zu erstellen —wie die Autoren ja auch mit ihrem Wissensgraphen zum Feld der Informatik zeigen (Dessí et al., 2022) —ist Genuine Semantic Publishing erst einmal eine Zukunftsvision. Neben dem Aufwand, den Autoren zukünftig in die Formalisierung ihrer Ergebnisse stecken müssten, erfordert dieser Ansatz in erster Linie einen Paradigmenwechsel im gesamten Publikationsprozess. Um das zu schaffen, müsste also eine heterogene Menge an Beteiligten —allen voran natürlich Wissenschaftler und Verleger —sich darauf einigen, diesen Weg zu beschreiten. Hilfreich wäre es sicherlich, wenn sich hierfür eine

Standard-Ontologie etablieren würde, die von allen Beteiligten genutzt werden könnte. Außerdem wäre auch eine Infrastruktur an Tools und Services notwendig, die die Autoren bei der Erstellung der semantischen Daten unterstützen. Auf Nanopublikationen setzende Experimente konnten zwar schon die prinzipielle Bereitschaft von Autoren zeigen, sich auf solche derartige Neuerungen einzulassen (Bucur et al., 2023), ob das aber auch für eine flächendeckende Umstellung reicht, ist angesichts des begrenzten Umfangs der Studie weiterhin fraglich.

Auch dann bleibt noch das Problem der bereits existierenden Publikationen, die nicht für maschinelle Verarbeitung ausgelegt sind. Für diese müsste dann eine zusätzliche Lösung, etwa eine rückwirkende Anreicherung relevanter Literatur mit semantischen Daten oder Methoden wie SCICERO, gefunden werden.

Konsistenz Ein weiterer potentieller Nachteil von Genuine Semantic Publishing ist die Konsistenz der erstellten Daten. Da die Autoren selbst für die Erstellung der semantischen Daten verantwortlich sind, könnte es zu Inkonsistenzen kommen, wenn beispielsweise verschiedene Autoren unterschiedliche Ontologien verwenden oder die gleichen Konzepte unterschiedlich benennen. Hier wären entweder standardisierte Ontologien notwendig oder eine (teil-)automatisierte Harmonisierung der Daten, wie sie beispielsweise SCICERO durchführt.

Da bei SCICERO hingegen dasselbe System für die Extraktion und Harmonisierung der Daten aller Publikationen verwendet wird, ist die Konsistenz der Daten hier gewährleistet. Andererseits sind —gerade durch den Einsatz von Transformer-Modellen —auch Fehler in der Extraktion möglich, die dann in allen extrahierten Daten auftreten. Außerdem ist dadurch auch die benötigte Rechenleistung ein Vielfaches höher als bei manuellen Ansätzen.

Weitere Herausforderungen SCICERO ist darauf ausgelegt, sehr große Mengen an Texten zu verarbeiten. Ob es bei geringeren Datenmengen, wie sie beispielsweise in einem spezifischen Forschungsfeld auftreten, zu Einschränkungen —etwa in der Genauigkeit der Validierung —kommt, ist nicht klar.

Bei beiden Ansätzen besteht die Gefahr, dass die semantischen Daten nicht alle für das Living Literature Review relevanten Daten enthalten. Da sich aber sowohl SCICERO als auch Genuine Semantic Publishing lediglich auf die Erstellung von Wissensgraphen konzentrieren, die dann als Grundlage für die Literaturübersicht dienen, ist die weitere Verarbeitung zur fertigen Übersicht aus dem Graphen für diese Arbeit nicht relevant.

6 Fazit

Aus technischer Sicht spricht nichts gegen die Vision des Genuine Semantic Publishings. Dennoch müsste der gesamte Publikationsprozess verändert werden, traditionelle, textbasierte Prozesse müssten abgelöst werden durch solche, die den Fokus auf die Erkenntnisse und deren präzise Repräsentation legen. Da ein solcher Wandel, der der Mitarbeit aller Forschenden bedarf, vorerst unrealistisch erscheint, sind die Vorteile des Verfahrens in der Praxis irrelevant.

Für den beschriebenen Anwendungsfall bleibt Genuine Semantic Publishing also prinzipiell geeignet, aber eher als eine Vision der Zukunft. Um sofort, oder auch nur in naher Zukunft, etwas Derartiges zu erreichen, führt wohl kein Weg an KI-basierten Ansätzen wie SCICERO vorbei, die auch mit den großen, bereits bestehenden Datenmengen umgehen können, die nicht extra mit dem Ziel erstellt wurden, für Maschinen oder Programme verarbeitbar zu sein.

Insbesondere angesichts der Tatsache, dass bei KI-basierten Ansätzen wie SCICERO immer die Möglichkeit besteht, dass Fehler gemacht werden, ist es wichtig, dass die Ergebnisse dieser Ansätze regelmäßig überprüft und gegebenenfalls korrigiert werden. Hier könnten Human-in-the-Loop Ansätze eine zentrale Rolle spielen und so die —insbesondere die Effizienz betreffenden —Vorteile von KI-basierten Ansätzen mit der Genauigkeit von manuellen Ansätzen verbinden. In solchen Ansätzen könnten menschliche Experten die Ergebnisse der KI-Systeme überprüfen und gegebenenfalls korrigieren, um so die Qualität der Ergebnisse zu sichern.

Tsaneva et al. (2024) zeigen, dass Human-in-the-loop Ansätze auch in der Praxis funktionieren können, allerdings mit Effizienzeinbußen gegenüber dem vollautomatisierten System. Zusätzlich zu den bereits genannten Vorteilen, wie der höheren Qualität der Ergebnisse, könnten solche Ansätze auch dazu beitragen, die Akzeptanz von KI-Systemen zu erhöhen, indem sie die Kontrolle über die Ergeb-

nisse wieder in die Hände der Menschen legen. Andererseits zeigen Sie auch, dass selbst mit Large Language Models (LLMs) als Experten (LLMs-in-the-Loop) die Qualität der Ergebnisse gegenüber dem Ausgangswert gesteigert werden kann, auch wenn die Präzision dieses Ansatzes (85%) nicht an die der menschlichen Experten (93%) heranreicht (Tsaneva et al., 2024). Die besten Ergebnisse können demnach mit einer Kombination aus LLMs-in-the-Loop und menschlichen Experten erreicht werden, bei der die menschlichen Experten nur diejenigen Fakten prüfen, bei denen die automatisierten Validierungsmechanismen —einerseits die Transformer, wie in Kapitel **??** beschrieben, und andererseits die LLMs —zu unterschiedlichen Ergebnissen kommen. Dadurch kann mit minimalem menschlichen Aufwand eine hohe Qualität der Ergebnisse erreicht werden.

Literatur

Arthur Brack, Anett Hoppe, and Ralph Ewerth. 2021. Citation recommendation for research papers via knowledge graphs. In *Linking Theory and Practice of Digital Libraries*, pages 165–174, Cham. Springer International Publishing.

Arthur Brack, Anett Hoppe, Markus Stocker, Sören Auer, and Ralph Ewerth. 2022. Analysing the requirements for an open research knowledge graph: use cases, quality requirements, and construction strategies. *International Journal on Digital Libraries*, 23(1):33–55.

Cristina-Iulia Bucur, Tobias Kuhn, Davide Ceolin, and Jacco van Ossenbruggen. 2023. Nanopublication-based semantic publishing and reviewing: a field study with formalization papers. *PeerJ Comput Sci*, 9:e1159.

Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, and Silvia Quarteroni. 2013. *The Information Retrieval Process*, pages 13–26. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cochrane. Guidance for the production and publication of cochrane living systematic reviews: Cochrane reviews in living mode. https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf.

Danilo Dessí, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2022. Cs-kg: A large-scale knowledge graph of research entities and claims in computer science. In *ISWC 2022: 21st International Semantic Web Conference*, volume 13489, pages 678–696. Springer, Cham.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, Enrico Motta, and Harald Sack. 2020. Ai-kg: An automatically generated knowledge graph of artificial intelligence. In *The Semantic Web – ISWC 2020*, pages 127–143, Cham. Springer International Publishing.

Danilo Dessí, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2022. Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain. *Knowledge-Based Systems*, 258:109945.

Julian H Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, et al. 2017. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of clinical epidemiology*, 91:23–30.

Paul Groth, Andrew Gibson, and Jan Velterop. 2010. The anatomy of a nanopublication. *Information Services and Use*, 30(1-2):51–56.

Azanzi Jiomekong, Sören Auer, and Allard Oelen. 2024. Linked open literature review using the neuro-symbolic open research knowledge graph. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1015–1018, New York, NY, USA. Association for Computing Machinery.

Huquan Kang, Luoyi Fu, Russell J. Funk, Xinbing Wang, Jiaxin Ding, Shiyu Liang, Jianghao Wang, Lei Zhou, and Chenghu Zhou. 2024. Scientific and technological knowledge grows linearly over time.

Huquan Kang, Xinbing Wang, Luoyi Fu, Jiaxin Ding, Shiyu Liang, Jianghao Wang, Lei Zhou, and Chenghu Zhou. 2023. Publication data for article "knowledge does not explode but increases linearly over time".

Tobias Kuhn and Michel Dumontier. 2017. Genuine semantic publishing. *Data Science*, 1(1-2):139–154.

Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. Skos Core: Simple knowledge organisation for the Web. *International Conference on Dublin Core and Metadata Applications*, 2005.

Francesco Ronzano and Horacio Saggion. 2016. Knowledge extraction and modeling from scientific publications. In *Semantics, Analytics, Visualization. Enhancing Scholarly Data*, pages 11–25, Cham. Springer International Publishing.

David Shotton. 2010. Cito, the citation typing ontology. *Journal of Biomedical Semantics*, 1(1):S6.

Stefani Tsaneva, Danilo Dessí, Francesco Osborne, and Marta Sabou. 2024. Enhancing scientific knowledge graph generation pipelines with llms and human-in-the-loop. In *Sci-K@ISWC*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Michel Wijkstra, Timo Lek, Tobias Kuhn, Kasper Welbers, and Mickey Steijaert. 2021. Living literature reviews. In *Proceedings of the 11th Knowledge Capture Conference*, K-CAP '21, page 241–248, New York, NY, USA. Association for Computing Machinery.