

Zusammenfassung der Dissertation

„*Autonome Topic Maps*. Zur dezentralen Erstellung von implizit und explizit vernetzten Topic Maps in semantisch heterogenen Umgebungen.“ **Autor:** Lutz Maicher

Motivation – das TMweb

Als Grundidee vieler Anwendungen des Semantic Web und des Web 2.0 kristallisiert sich die dezentrale Erstellung und Veröffentlichung von sogenannten Semantic Web-Dokumente heraus. Diese bestehen aus *einigen* organisatorischen, technischen oder inhaltlichen Aussagen zu *Aussagegegenständen*, wie bestimmten Personen, Produkten oder Informationsressourcen. Das entscheidende Kriterium dieser Dokumente ist dabei die Umsetzung des *aussagegegenstandszentrierten Modellierungsparadigmas* bei ihrer Erstellung.

Entsprechend des aussagegegenstandszentrierten Modellierungsparadigmas wird für *jeden* relevanten Aussagegegenstand der realen Welt *genau* ein Repräsentant im Modell erstellt. Eigenschaften in der realen Welt werden in den Modellen als typisierte Beziehungen zwischen Repräsentanten von Aussagegegenständen dokumentiert, was zu einer *expliziten* Vernetzung innerhalb der erstellten Modelle führt. Zudem wird sichergestellt, dass Repräsentanten, die den gleichen Aussagegegenstand verkörpern, zusammengeführt werden. Dadurch können bei konsistenter Nutzung von Vokabularen Informationen aus verschiedenen, dezentral erstellten Modellen durch *zusammenführende Abfragen* integriert werden. Repräsentanten des gleichen Aussagegegenstandes, die in getrennten Dokumenten vorliegen, sind somit *implizit* vernetzt. Diese implizite Vernetzung existiert latent und wird erst evident, wenn die Modelle aufeinander treffen. In diesem Moment wird die Mächtigkeit der impliziten Vernetzung der Repräsentanten sichtbar, da alle Informationen zu einem Aussagegegenstand an einem Punkt verfügbar werden. Insbesondere Topic Maps, aber auch RDF/OWL setzen das aussagegegenstandszentrierte Modellierungsparadigma um.

Das Zusammenspiel von impliziter und expliziter Vernetzung illustriert die Emergenz einer globalen, stark vernetzten Faktenbasis, basierend auf der dezentralen, kleinteiligen und kollaborative Dokumentation von Informationen entsprechend des aussagegegenstandszentrierten Modellierungsparadigmas. Durch die Dezentralität der Erstellung dieser Faktenbasis kann ihr Wachstum durch keine Kapazitätsbeschränkung irgendeiner Organisation aufgehalten werden.

Die Etablierung und Nutzung dieser globalen Faktenbasis bedarf einer, auf bestehenden Internet-Technologien aufbauenden, Infrastruktur, die zum heutigen Zeitpunkt *noch* nicht existiert. Unter der Federführung des Autors dieser Arbeit beantragten im März 2007 mehr als 30 Wissenschaftler aus 17 Ländern bei der Europäischen Kommission die Etablierung eines Forschungsnetzwerkes für die Etablierung des in Abbildung 1 illustrierten *TMweb*, welches genau diese Infrastruktur sein wird. Im *TMweb* werden zentrale, kommerzielle und Open Access *Content-Provider*, also Topic Maps-Servern, und *Content-Consumer*, also Topic Maps-Clients, agieren. Die Consumer erfragen unter Nutzung standardisierter Austauschprotokolle Informationen zu Aussagegegenständen bei den Servern. Wenn dort entsprechende Informationen vorliegen, dann werden diese als Topic Maps-Fragmente an den anfragenden Client gesendet. Durch die implizite Vernetzung können die empfangenen Fragmente zusammengeführt werden, so dass alle Informationen zu dem angefragten Aussagegegenstand an einem Repräsentanten verfügbar sind. Da Topic Maps ein generisches Austauschformat für Informationen sind, können die Informationen durch beliebige Interfaces an den Nutzer ausgeliefert werden. Diese können bspw. webbasierte Topic Maps-Portale, aber auch Textverarbeitungssysteme oder andere Applikationen.

Ein Spezialfall im *TMweb* stellen die Semantic Web-Dokumente dar, welche hier als Published Topic Maplets bezeichnet werden. Diese publizierten Dokumente werden entweder durch Crawler aufgefunden oder können direkt bei zentralen Services registriert werden. Aufgabe dieser Services ist zum einen die Indexierung der registrierten Topic Maplets, um die in den Modellen genutzten Terme global verfügbar zu machen, und zum anderen werden diese Services als Content-Provider agieren, um die in den Topic Maplets vorliegenden Informationen über standardisierte Schnittstellen verfügbar zu machen.

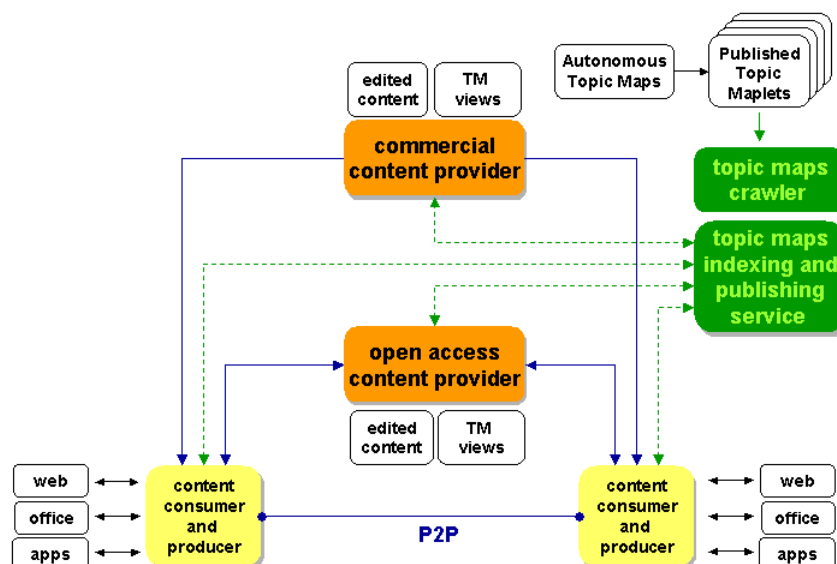


Abbildung 1: TMweb - The Future of Topic Mapping

Es ist offensichtlich, dass erst durch das *TMweb* eine Infrastruktur zur Verfügung gestellt werden wird, welche eine sinnvolle Nutzung von Semantic Web-Dokumenten erlauben wird. Es ist jedoch ebenso ersichtlich, dass gerade die Menge des verfügbaren Content (und hierzu gehören insbesondere auch die veröffentlichten Topic Maplets) ein kritisches Erfolgskriterium für die Etablierung des *TMweb* sein wird.

Problemstellung und Lösungskonzept

Im Rahmen dieser Arbeit wird mit *Autonomen Topic Maps* (ATM) eine Konzeption entwickelt, mit deren Hilfe die dezentrale Erstellung qualitativ hochwertiger, vernetzter Dokumente in semantisch heterogenen Umgebungen ermöglicht wird. Eine ATM ist eine Topic Map, die eine durch einen generischen Interpreter ausführbare Modellierungsmethoden formalisiert, durch deren Ausführung Beobachtungen zu einer spezifischen Domäne in implizit und explizit vernetzten Topic Maps adäquat dezentral dokumentiert werden. ATMs können leicht verteilt werden und in Interpretern für beliebige Anwendungskontexte genutzt werden. Das Konzept der ATMs erlaubt somit die Skalierung der Erstellung von Topic Maplets, um die kritische Masse an Content für das *TMweb* zu erzeugen. Eine ATM ist folgendermaßen definiert:

Eine ATM ist eine Topic Map, die die notwendigen Informationen beinhaltet, dass gegeben ein generischer Interpreter, in beliebigen Situationen bestmöglich **implizit vernetzte Modellinstanzen desselben Modelltyps** erstellt werden. Autonome Topic Maps legen alle Informationen über die angewandte Modellierungsmethoden offen und dokumentieren diese somit. Die erstellten Modellinstanzen sind Topic Maps. Sie sollten die zur Erstellung genutzte ATM referenzieren.

Es besteht die Frage, warum für die Erzeugung implizit vernetzter Topic Maplets für das *TMweb* die üblicherweise propagierte Nutzung eines kontrollierten Vokabulars bzw. einer gemeinsamen Ontologie nicht ausreichend ist, sondern durch diese Arbeit das Konzept der Autonomen Topic Maps eingeführt wird. Die Begründung für die Nutzung von Ontologien erscheint plausibel: ausgehend von der Annahme, dass Ontologien gemeinsame Vokabulare für Aussagen über Aussagegegenstände definieren, wird üblicherweise geschlossen, dass die Autoren bei Nutzung des gemeinsamen Vokabulars *identische* Beobachtungen über die Welt mit *identischen* Aussagen der verfügbaren Sprache ausdrücken. Ein Seitenblick auf die natürliche Sprache lässt jedoch Skepsis aufkommen, da es zumindest dort ein gegebenes Vokabular erlaubt, Beobachtungen der Welt durch eine Vielzahl verschiedener, syntaktisch und semantisch korrekter Sätze zu explizieren. Dieser Interpretationsspielraum von Vokabularen wird *Semantic Gap* in Ontologien genannt: die Nutzung *einer* Ontologie impliziert die Möglichkeit der Erstellung von Instanzen *einer Vielzahl* unterschiedlicher Modelltypen. Es muss somit sichergestellt werden, dass Vokabular *und* Modellierungsmethode verteilt werden.

Die zweite zentrale Problemstellung ist die Sicherstellung der impliziten Vernetzung, sowohl auf Typ- als auch auf Individuenebene. Die dabei entstehenden Probleme sind offensichtlich: so können mit dem durch die Ontologie gegebenen Vokabular gleiche Beobachtungen unterschiedlich beschrieben werden (*Synonymie*) bzw. unterschiedliche Beobachtungen gleich beschrieben werden (*Homonymie*).

Die Auflösung von Homonymie und Synonymie ist kein Gegenstand des Integrationsmodells von Topic Maps, da davon ausgegangen wird, dass die Beobachtungen über die Identität der Aussagegegenstände korrekt, d. h. entsprechend der Intention des Integrationsmodells, dokumentiert wurden. Die Beseitigung von Homonymie und Synonymie ist somit Bestandteil des *Erstellungsprozesses* von Topic Maps. Dies bedeutet, dass die Modellierungsmethoden das Auftreten von Homonymie und Synonymie unterbinden sollten.

Die *Offenlegung der Modellierungsmethode* ist somit ein probates Mittel, um dem Ziel qualitativ hochwertiger, implizit vernetzter Modelle näher zu kommen. Der nächste Schritt ist die *Formalisierung der Modellierungsmethode*, so dass diese durch generische Interpreter ausführbar wird. Im Rahmen dieser Arbeit wird mit Modelling Workflow Patterns (MWP) eine solche Infrastruktur für beliebige Modellierungsmethoden geschaffen. Eine ATM ist eine als Topic Map repräsentierte MWP-Beschreibung, die Topic Maps erzeugt.

Der Einsatz von MWP-Beschreibungen löst jedoch nicht Problematik der Homonymie und Synonymie auf Individuenebene in terminologisch dynamischen Domänen, da die offengelegten Modellierungsmethoden nicht mit *ex ante* unbekanntem Individuen umgehen können. In diesem Fall ist es notwendig, dass das Auftreten eines neuartigen Aussagegegenstandes, z. B. einer bestimmten Person, gehandhabt wird. Hierfür kann das *aussagegegenstandszentriertes Retrieval* genutzt werden. Besonders wichtig hierbei ist der *Impact of Semantic Handshakes*, da durch die Dokumentation und Offenlegung von lokalen Integrationsentscheidungen, d. h. der synonymen Nutzung mehrerer Terme für die Referenzierung eines Aussagegegenstandes, das Problem der Synonymie global weitgehend aufgelöst werden kann. Es wird in dieser Arbeit weiter gezeigt, dass auf die Nutzung zentraler Ontologien (auf Individuenebene) verzichtet werden kann, wenn jedes Topic mind. zwei (populäre) Terme nutzt, um den repräsentierten Aussagegegenstand anzuzeigen. Dieser bottom-up Ansatz ohne zentrale, ordnende Instanz ist orthogonal zu dem üblichen Weg der Definition global gültiger, kontrollierter Vokabulare.

Problematisch bleibt allein die homonyme Nutzung von Termen auf Individuenebene. Durch die Trennung zwischen Stadien- und Integrationsrepräsentanten kann diese Homonymität nicht vermieden werden, jedoch bei deren Entdeckung eine eindeutige Disambiguierung in den Modellen global durchgesetzt werden.

Beitrag und Aufbau der Arbeit

Der wissenschaftliche Beitrag dieser Arbeit ist:

- in Kapitel A* die Entwicklung des Konzepts der Autonome Topic Maps, als Verfahren zur Dokumentation von qualitativ hochwertigen, implizit und explizit vernetzten Dokumenten in verteilten, heterogenen Umgebungen, und dessen Einbettung in die Idee des *TMweb*,
- in Kapitel B* die umfassende, deutschsprachige Zusammenfassung des derzeitigen Entwicklungsstandes der Topic Maps-Technologien,
- in Kapitel C* die Diskussion der Semantik in Topic Maps-Technologien, und dabei insbesondere der Diskurs zur Möglichkeit der Auflösung von Homonymie und Synonymie auf Individuenebene in dynamischen Domänen,
- in Kapitel D* die exakten *Spezifikation* von Autonomen Topic Maps, basierend auf der Architektur aus Petrinetz-Datenmodell und Petrinetz-Prozessmodell, und die Erstellung der Referenzimplementierung *fluidS*, als generischer ATM-Interpreter,
- in Kapitel E* die exemplarische Demonstration der Anwendung von ATMs im Kontext der dezentralen Dokumentation von Metadaten mit dem Dublin Core-Vokabular durch die *DC4TM-ATM*.

In Kapitel B werden die bestehenden Topic Maps-Technologien detailliert eingeführt. Diese Arbeit hat den Anspruch, einen Beitrag zur Etablierung einer *deutschsprachigen* Topic Maps-Terminologie zu leisten. Nach einer einleitenden Übersicht über die historische Entwicklung der Topic Maps-Technologie wird in B.1 das Topic Maps-Datenmodell vorgestellt. Nach der Einführung bestehender Austauschformate für Topic Maps in B.2 wird in Abschnitt B.3 das Topic Maps-Referenzmodell beschrieben. Dieses Informationsmodell ist das Fundament des theoretischen Diskurses über die Semantik in Topic Maps-Technologien in Kapitel C. Die Diskussion bestehender Ansätze zur Abfrage und Beschränkung von Topic Maps in den Abschnitten B.4 und B.5 wird benötigt, da Modellierungsmethoden als Sequenz von Abfrage- und Modifikationsphasen betrachtet werden können, welche Modelle erzeugen, die bestimmten Beschränkungsphrasen unterliegen. In B.6 werden die Unterschiede zwischen Topic Maps und RDF/OWL herausgearbeitet. Der Abschnitt B.7 stellt Austauschprotokolle für Topic Maps vor, auf denen die Kommunikation im TMweb basieren wird. Im abschließenden Abschnitt B.8 werden bestehende Topic Maps-Engines kurz beschrieben.

In Kapitel C wird die Semantik in Topic Maps diskutiert und Implikationen für deren qualitative Erstellung in dezentralen, terminologisch heterogenen Umgebungen bei Sicherstellung der impliziten Vernetzung entwickelt. In C.1 wird herausgearbeitet, welcher Verantwortungsbereich der semantischen Domäne eines topic maps-verarbeitenden Informationssystems durch die Topic Maps-Standards abgedeckt wird und welcher Bereich in der Verantwortung der entsprechenden Anwendungen liegt. Es zeigt sich, dass die einzige Funktionalität im Verantwortungsbereich der Standards die Sicherstellung des Zustands ist, dass Repräsentanten im Modell, bei denen Gleichheit des Aussagegegenstands vorliegt, als zusammengeführt zu betrachten sind. Dies wirft die Fragestellung nach der Gleichheit des Aussagegegenstandes eines Repräsentanten auf, der in Abschnitt C.2 diskutiert wird. In Abschnitt C.3 wird anschließend die Dokumentation von Topic Maps bei terminologischer Heterogenität erörtert. Im abschließenden Abschnitt C.4 wird mit Hilfe von Simulationen der Einfluss von *Semantic Handshakes* aufgezeigt. Diese lokalen Integrationsentscheidungen führen zu terminologischer Harmonisierung in Abwesenheit zentraler, ordnender Instanzen.

In Kapitel D wird die Konzeption der formalisierten, ausführbaren Modellierungsmethoden, den Modelling Workflow Patterns, spezifiziert. In Abschnitt D.1 wird aufgezeigt, dass eine Modellierungsmethode als Prozess zu formalisieren ist. Mit Hilfe von Petrinetzen könne beliebige Prozesse repräsentiert werden, zudem können beliebige Petrinetze durch einige grundlegenden Bausteinen definiert werden. In der Literatur existiert kein Ansatz zur Repräsentation von Petrinetzen in Topic Maps, so dass dies im Kontext dieser Arbeit entwickelt wird. In D.2 werden die notwendigen, terminologischen Festlegungen für Petrinetze getroffen und in Abschnitt D.3 wird deren Grundprinzip beschrieben. In D.4 wird ein Petrinetz-Datenmodell spezifiziert. Dieses Datenmodell erlaubt es, beliebige Petrinetze zu repräsentieren. Die entsprechende Prozesssemantik wird durch Prozessmodelle spezifiziert, wobei ein spezifisches Prozessmodell in Abschnitt D.5 eingeführt wird. Die in D.6 spezifizierte isomorphe Abbildung zwischen dem Petrinetz- und dem Topic Maps-Datenmodell erlaubt die Repräsentation beliebiger Petrinetze in beliebigen Topic Maps-Austauschformaten. In D.7 wird an praxisorientierten Beispielen gezeigt, wie MWP-Beschreibungen als Topic Maps zu erstellen sind. In D.8 wird die entwickelte Konzeption diskutiert. Im abschließenden Abschnitt D.9 wird die im Rahmen der Arbeit entwickelte Referenzimplementierung *fluidS* vorgestellt.

In Kapitel E wird die Nutzung von ATMs am Beispiel der dezentralen Dokumentation von Metadaten-Topic Maplets mit dem Dublin Core-Vokabular demonstriert. Durch die in E.2 spezifizierte Abbildung zwischen dem TMDM und dem Metamodell von Dublin Core kann dieses Vokabular konsistent in Topic Maps genutzt werden. Mit *DC4TM* wird in E.2 eine Modellierungsmethode für die Nutzung des DC-Vokabulars in Topic Maps entwickelt. Durch *DC4TM* wird die Nutzung des Vokabulars auf der Typebene standardisiert, adäquate Gegenstandsanzeiger auf Individuenebene werden, wie in den Abschnitten E.3 und E.4 gezeigt, entweder durch kontrollierte Vokabulare (in statischen Domänen, z. B. Sprachen, Orte und Daten) oder durch aussagegegenstandszentriertes Retrieval (in dynamischen Domänen, z. B. Personen) zur Verfügung gestellt. In Abschnitt E.5 wird die Modellierungsmethode durch die *DC4TM-ATM* als Autonome Topic Map formalisiert.

Im abschließenden Kapitel F wird der Beitrag der Arbeit zur dezentralen Erstellung qualitativ hochwertiger, implizit und explizit vernetzter Topic Maps in semantisch heterogenen Umgebungen diskutiert.