

Die Semantische Kontrolle bei IUWIS

Ben Kaden, Konstantin Baierer

Ziel des Reports

Interne und externe Inhalte werden bei IUWIS in einem so genannten *Informationspool* gesammelt, nach einem Datenmodell strukturiert, nach verschiedenen Kriterien inhaltlich erschlossen und relationiert sowie für den Diskurs, vor allem in den IUWIS-Dossiers, nutzbar gemacht. Dieser Report fasst den aktuellen Stand des Grundkonzepts zum Themenkomplex „Semantische Erschließung“ bei IUWIS zusammen. Dieser Komplex umfasst einerseits die parallele Erschließung von Informationsobjekten mit einem kontrollierten Vokabular und freien *Tags* sowie andererseits die Operationalisierung dieser Erschließungsdaten für die drei Prinzipien der *Facettierung*, *Relationierung* sowie *Relevanzierung*.

In diesem informationswissenschaftlich fundierten Projekt wird mit verschiedenen Verfahren der semantischen Erschließung und Kontrolle experimentiert. Die folgende Zusammenstellung spiegelt den derzeitigen Stand der konzeptionellen Arbeit wider. Es ist nicht vorgesehen, dies im gesamten Umfang von Anfang an zu realisieren. Es wird mit einfachen Verfahren wie *Tagging* begonnen, um dann schrittweise reichere und auch automatisierbare Verfahren einzusetzen. Für weitere Arbeiten ist auch eine Visualisierung der verschiedenen Relationsformen vorgesehen.

Idee

Die *Tagsonomy* (von *tag* und *taxonomie*) erschließt die innerhalb des IUWIS-Informationspools erfassten Informationsobjekte für eine diskursnahe Nutzung. Die Auffindbarkeit der Objekte wird ebenso verbessert wie die Möglichkeiten, diese Objekte (semi-automatisiert) in den Diskurs einzubetten, beispielsweise als Materialien in die Dossiers.

Diese Option betrifft sowohl die manuelle Recherche und Verknüpfung als auch die automatisierte Erzeugung und Abbildung von semantischen Beziehungen.

Zu diesem Zweck wird die Erschließung

- a) von Informationsobjekten mit einem kontrollierten Vokabular, sowie
- b) von Informationsobjekten über freie *Personomien* (=Tags)

miteinander in einem System verschränkt und für verschiedene Anwendungszusammenhänge elaboriert.

Damit soll bei der Recherche ein hoher Recall bei hoher *Precision* abgesichert werden. Andererseits möchte IUWIS den jeweils individuellen terminologischen Zugang der Vertreter der Fachgemeinschaft (die so genannten *emergent semantics*) für die Erschließung und Auffindbarmachung integrieren. Auf diese Weise kann die terminologische Dynamik des Themengebietes möglichst zeitnah und umfassend berücksichtigt werden.

Tags, Personomien und Folksonomien

Tags gelten als intuitive und einfache Form für die Organisation von Inhalten. Tags dienen ebenso als Annotationsmittel. Da sie hypertextuell auch als Links funktionieren, und somit über Webanwendungen nutzerübergreifend geteilt werden können, spricht man auch von *Social Annotation*. Sie sind ohne großen Aufwand und mit sofortigem Nutzen anwendbar. Allerdings ist das Vokabular nicht kontrolliert, was zu einem hohen Anteil an Unschärfen bzw. eingeschränkten *Recall* führt.

Dennoch ist davon auszugehen, dass unterschiedliche *Personomien* hinsichtlich der Erschließung bestimmter Quellen Ähnlichkeiten aufweisen. Unter *Personomie* ist das von einer Person in einem *Tagging*-System verwendete Vokabular zu verstehen. Die Verbindung der *Personomien*, die in der Regel automatisch erzeugt wird, wird in ihrer Gesamtheit als *Folksonomy* bezeichnet.

Die in diesem Zusammenhang entstehenden so genannten *emergent semantics* basieren auf drei Aspekten:

- geteiltes implizites Hintergrundwissen (*shared knowledge*)
- gegenseitiger Einfluss von Nutzern aufeinander (*dynamic feedback*)

- Bedingungen der dahinterliegenden sozialen Netzwerke (*social paradigms*).

Der Aspekt des *Emergenten*, also des aus sich heraus und ungeplant und ungerichtet Entstehenden, bezieht sich in diesem Zusammenhang darauf, dass sich im Anschluß an das Einzelhandeln der Tags vergebenden Akteure unvorhergesehene, aber im Idealfall produktive Effekte ergeben. Ein bekanntes Beispiel ist die Steigerung der *Browsing- und Serendipity-Effekte*. Ziel des IUWIS-Tagsonomy-Systems ist es, solche Effekte so umfassend wie möglich aufzugreifen und zu nutzen.

Die Berücksichtigung freier, individueller Vokabularien in der Tagsonomy von IUWIS folgt dem Anspruch einer möglichst großen terminologischen Nähe des Erschließungsvokabulars zum tatsächlichen Vokabular der Diskursgemeinschaft.

Dabei spielt die Erfassung der Personomien und die der sich aus diesen heraus verfestigenden, für die Diskursgemeinschaft typischen Vokabularien (=Folksonomien) gleichermaßen eine große Rolle. Aus diesem Grund wird eine permanente Durchlässigkeit zwischen dem kontrollierten Vokabular(=Thementags) und dem frei verwendeten Vokabular angestrebt. (zum Verfahren vgl. unten)

Das kontrollierte Vokabular

Allerdings erweist sich die ausschließliche Ausrichtung des Angebots auf freie *Tags*, wie sie sich bei *Social Bookmarking*-Diensten wie Delicious, Connotea oder Bibsonomy findet, für die bei IUWIS angestrebte *Retrieval*-Praxis als zu kontingent. Um eine möglichst exakte Grunderschließung zu gewährleisten, die diskursnah und doch terminologisch möglichst exakt arbeitet, wird der freien Erschließung eine redaktionelle Erschließung der Informationsobjekte im IUWIS-Infopool mittels eines fixen Vokabularsystems entgegengestellt. Der Ansatz weist dabei über die *Retrieval-Leistung* traditioneller Informationssysteme hinaus und berücksichtigt den Anwendungsrahmen, der prinzipiell eine direkte Einbindung der Objekte in das Diskurssystem (vgl. den Report zu den IUWIS-Dossiers) ermöglichen soll. Ein Grundmerkmal ist die separate Erfassung des Vokabulars

zur Erschließung der Diskursthemen (=Themen) und der Diskursteilnehmer (=Akteure).

Aufgrund des Facettierungsansatzes bei IUWIS (vgl. unten) entspricht das kontrollierte Vokabular in der Abbildung Kombinationen von Klassen und Instanzen. Diese *Tag*-Kombinationen sind allerdings nicht *per se* fixiert, sondern können je nach Bedarf flexibel erzeugt werden.

Die Klassen stellen allgemeine thematische *Foci* dar – beispielsweise Recht, Gesellschaft, Wissenschaft, Bildung. Sie werden durch einen Pool von *Tag*-Instanzen, die konkrete Phänomene bezeichnen, ergänzt – beispielsweise Open Access, Medienfreiheit oder Urheberpersönlichkeitsrecht.

Zu beachten ist dabei, dass die Klassen möglichst allgemein und ihre Zahl möglichst gering gehalten werden. Der Klassenumfang soll zu Beginn der technischen Umsetzung der *Tagsonomy* nicht höher als 20 sein. Obschon er prinzipiell erweiterbar ist, soll sie möglichst nicht stark wachsen.

Die Anzahl der Instanzen, also der *Tags*, die das kontrollierte Vokabular bilden, soll anfangs nicht höher als 1000 sein. Dies bezieht sich sowohl auf die Akteure als auch auf die Themen. Allerdings ist hier angesichts der vorhersehbaren Diskursentwicklung ein größeres Wachstum zu erwarten. Das Wachstum wird von vornherein in der Struktur berücksichtigt. Da das Primäranliegen von IUWIS in der Unterstützung der Diskurse liegt, ist eine Erweiterung des Vokabulars generell angestrebt. Die Erfassung von freien *Tags* und der Integrationsworkflow von relevanten *Tags* in dem kontrollierten Pool ermöglichen einen möglichst fließenden Übergang ohne terminologischen Verlust.

Der Pool der kontrollierten Thementags wird zur besseren Übersicht in einer Ontologie erfasst und gepflegt. Langfristig ist wahrscheinlich, dass die ontologische Struktur auch in die Navigationsstruktur einfließen kann.

Leitprinzipien der semantische Kontrolle

Mithilfe des IUWIS-*Tagsonomy*-Ansatzes wird es u.a. möglich, das diskursive Geschehen auf IUWIS semantisch aufbereitet auf einer Metaebene zu operationalisieren. Der Vorteil eines solchen Verfahrens liegt

1. in einem präziseren Zugriff auf die einzelnen Diskurselemente (bzw. Informationsobjekte) für die Nutzer,
2. einer optimierten Verknüpfbarkeit von Aussagen und
3. elaborierten Visualisierungen der Diskurslinien, die nicht zuletzt diskursanalytische Erkenntnisse generieren. Diese können in die Weiterentwicklung von Plattformen für die Wissenschaftskommunikation einfließen.

Der Ansatz von IUWIS ist demnach mehrdimensional konzipiert. Die Tagstrukturen sind sowohl Rechercheinstrument als auch potentielles Analysewerkzeug. Damit trägt das Projekt dem wachsenden Bedarf an proaktiven Webanwendungen Rechnung.

Die Leitprinzipien bei der Entwicklung des Modells zur Semantischen Kontrolle bei IUWIS lassen sich daher mit den drei Prinzipien

- Facettierung
- Relationierung
- Relevanzierung

beschreiben.

Facettierung

Die Facettierung kommt hauptsächlich bei der Organisation und Spezifizierung von Themen zum Einsatz. Bisherige Erfahrungen sowohl mit kontrolliertem Vokabular als auch besonders Folksonomien weisen auf ein hohes Maß an Unschärfe aufgrund der aus dem *Information Retrieval* bekannten Probleme wie Homonymie oder Polysemie hin. Über eine zusätzliche den Kontext beschreibende Facettierung soll dem entgegen gewirkt werden.

Die bisherigen Evaluierungen der Facettierung ergaben jedoch, dass ein die gesamte Vielfalt erfassen wollendes Konzept für die praktische Umsetzung zu komplex ist. Das für die Umsetzung geplante Verfahren versucht daher einen Mittelweg zwischen der angestrebten Präzision und dem Erfassungsaufwand.

Kategorisierung

Als grundsätzlich sinnvoll erweist sich eine Kategorisierung der Verwendungsentention von *Tags*. Im Gegenstandsbereich wird zwischen Akteuren und Themen differenziert. Weitere aus *Folksonomien* bekannte Funktionstags werden in das Metadatenmodell ausgelagert und so weit als möglich automatisiert erfasst. Bezugsangaben wie Ort und Zeit können ebenfalls differenzierter erfasst werden. Die Kombination einzelner Metadaten und *Tags* ermöglicht die differenzierte Erfassung von Ereignissen.

Thematische Facettierung

Auf der Themenebene können *Tags* zusätzlich weitgehend flexibel facettiert werden. Hier können *Tags* als Klassen auftauchen, denen andere *Tags* je nach Blickwinkel als Instanzen zugeordnet werden. So kann beispielsweise das Phänomen Internet je nach inhaltlicher Ausrichtung des Informationsobjektes aus der Klasse *Wirtschaft*, der Klasse *Gesellschaft* oder auch der Klasse *Recht* erfasst werden.

Eine noch granularere Erschließung von Inhalten wird durch die Kombinierbarkeit von Ober- und Unterinstanzen erreicht. So kann bei Bedarf einer Instanz eine weitere als Spezifizierung zugewiesen werden. Auch hier wird eine flexible Facettierbarkeit ermöglicht.

Prinzipiell sollen alle im kontrollierten Vokabular erfassten *Tags* auf dieser Basis frei kombiniert werden können. Gleiches gilt für die erfassten Akteure.

Relationierung

Beim Verfahren der Relationierung werden Informationsobjekte auf der Grundlage bestimmte Regeln aufeinander bezogen. Diese Relationierung ist in der Darstellung invers.

Die Relationierung kann sowohl formal wie auch diskursiv bestehen. Beispiele für eine formale Relation sind ein Einzelvortrag als Bestandteil einer Konferenz oder ein Aufsatz als Teil eines Sammelbandes. Weiterhin ermöglicht die *Tagsonomy* eine thematische Relationierung.

Formale Relationen

Die formale Relationierung wird nicht über die *Tagsonomy* selbst, sondern über das Datenmodell realisiert. Die Realisierung wird hauptsächlich Enthaltenseins-Beziehungen umfassen.

Diskursive Relationen

Die diskursiven Relationen dienen der Abbildung einer argumentativen Qualität. Obschon sich auch hier ein granulares Verfahren anbietet, werden aus Komplexitätsgründen die Basisrelationen

- zitiert (wertfreier Bezug)
- stützt (beurteilender Bezug, positiv)
- kritisiert (beurteilender Bezug, negativ)
- falsifiziert (Widerlegung durch eindeutige Beweisführung)
- verifiziert (Bestätigung durch eindeutige Beweisführung)

vorgesehen. Der diskursive Charakter wird, so die Annahme, besonders die ersten drei Formen umfassen.

Die Zuweisung der diskursiven Relationen erfolgt damit analog zu den freien *Tags* nicht absolut, sondern jeweils durch die Einzelnutzer. Der daraus entstehende und möglichst transparent angegebene Häufigkeitswert (*15 Nutzern meinen, der Inhalt dieses Beitrags stützt die Thesen dieses Beitrags*) dient zur Objektivierung der Einschätzung.

Thematische Relationen

Da jedes Informationsobjekt sowohl mit *Tags* aus dem kontrollierten Vokabular ausgezeichnet ist, als auch von den Nutzern mit freien *Tags* erschlossen werden kann, lassen sich die Objekte relativ präzise aufgrund thematischer Übereinstimmungen in Beziehung setzen. Anhand der Übereinstimmungen im Erschließungsvokabular ist es beispielsweise möglich, Schnittmengen verschiedenen Grades zu ermitteln und diese entsprechend sortiert beim *Retrieval* als Empfehlungen bzw. thematisch verwandte Objekte auszugeben.

Gleiches ist selbstverständlich auch auf der Grundlage der Akteur-*Tags*, zeitlicher, räumlicher und anderer Angaben bzw. in Kombination möglich. Auch

die Kombination mit den diskursiven Relationen ist denkbar. Am Ende könnten demnach Objekte ermittelt werden, die in einem bestimmten Zeitraum zu einem konkreten Thema unter Berücksichtigung eines bestimmten Ereignisses mit einem spezifischen diskursiven Bezug auf andere Objekte erstellt wurden. Die angestrebte flexible Kombinierbarkeit der Filtermerkmale soll dabei einen denkbar exakten Zugriff auf bestimmte Argumentationslinien im Diskurs ermöglichen.

Relevanzierung

Mittels der Relevanzierung soll die Möglichkeit geboten werden, Informationsobjekte innerhalb des Erschließungskontextes thematisch zu gewichten. Dies erfolgt über die differenzierende Zuweisung jeweils eines Haupt-*Tags* und weiterer *Tags*. Diese Zuweisung erfolgt für Themen und Akteure.

Auf dieser Grundlage soll die Präzision des *Retrievals* dahingehend erhöht werden, dass Nutzer ihre Recherche entweder auf Objekte konzentrieren, die für das jeweilige Themengebiet besonders relevant sind oder aber eine erarbeitete Trefferliste nach Relevanz nachsortieren können. Hier sollen die Haupt-*Tags* je nach Sucheinstieg für die Anzeigesortierung besonders gewichtet werden.

Die Relevanzierung dient darüberhinaus auch dazu, mögliche Schnittmengen bei der thematischen Relationierung von Informationsobjekten stärker zu gewichten.

Freie Tags

Freie *Tags* können von den Nutzern ohne jede syntaktische oder semantische Vorgabe jedem im Infopool erfassten Informationsobjekt zugewiesen werden. Sie werden als *Personomien* erfasst. Daraus folgt, dass die Relation zwischen dem Tag, dem getaggtten Informationsobjekt sowie dem *taggenden* Nutzer auf Dauer erfasst ist. Damit wird es möglich,

- zu jedem Nutzer eine nutzerspezifische *Tagcloud* zu generieren,
- über den Abgleich von *Personomien* bei gewissen Ähnlichkeiten eine Art „Interessennähe“ von freilich beschränkter Aussagekraft zu postulieren,

- das Auftreten und die Verbreitung bestimmter *Tags* im Diskursverlauf zu analysieren, um damit Rückschlüsse über die Struktur des Diskurses zu ziehen.

Selbstverständlich können diese Dienste und Analysemöglichkeiten nur unter Beachtung datenschutzrechtlicher Vorgabe implementiert werden.

Die dabei erfassten Angaben sind:

- Wann,
- wer,
- welches *Tag* aus
- welchem Vokabular (in diesem Fall: „Freie Tags“)
- welchem Informationsobjekt

zugeordnet hat/ist.

Autoextraktion von Tags als Sonderfall freier Tags

Eine weitere Möglichkeit, zusätzliches Erschließungsvokabular zu gewinnen, liegt in der Autoextraktion von Stichwörtern aus den inhaltsbeschreibenden Metadaten der Informationsobjekte. So sollen Titelstichwörter extrahiert und als freies Vokabular eines simulierten (Dummy)Nutzers „IUWIS-Auto-Tags“ erfasst werden. Dies ermöglicht eine technische Verarbeitung dieser Zusatzangaben innerhalb des bestehenden technischen Ansatzes.

Darüberhinaus werden Autoren- bzw. Herausgeberschaftsangaben als *Akteurtags* (=assoziierte Akteure) erfasst.

Übernahme von freien Tags in die Tagsonomy

Ein automatisiertes *Monitoring* des quantitativen Auftretens von freien *Tags* dient darüber hinaus zur permanenten Aktualisierung des kontrollierten Vokabulars.

Wenn das Auftreten eines *Tags* in freier Verwendung eine zuvor definierte quantitative Schwelle überschreitet, also z.B. 20-mal auftritt, wird es automatisch in eine dynamisch erzeugte Kandidatenliste aufgenommen, die regelmäßig redaktionell gesichtet und bearbeitet wird. Je nach Entscheidung

kann der Redakteur auf dieser Grundlage den bestehenden Korpus des kontrollierten Vokabulars bedarfsnah erweitern.

Zudem besteht die Möglichkeit, sowohl zwischen freien *Tags* als auch zwischen freien *Tags* und dem kontrollierten Vokabular Äquivalenzrelationen zu hinterlegen. Über diese können verschiedene Ansetzungsformen, Sprachvarianten, Schreibweisen, etc. als synonym definiert und auf eine bestimmte Hauptansetzung festgelegt werden, ohne dass den jeweiligen Nutzern eine bestimmte Ansetzung für ihre *Personomie* vorgeschrieben wird.

Die Definition von Hauptansetzungen auf einer Äquivalenzbasis ermöglicht es weiterhin, gezielt eingängige Bezeichnungen für die Ausgabe z.B. in *Tagclouds* festzulegen und somit die semantische Tiefe der Erschließung bei begrenzter Darstellungskomplexität aufrecht zu erhalten.