FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Analyse des Tagging-Verhaltens am Beispiel von Flickr: The Commons

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Amit Gahlay, BSc
Matrikelnummer 0625608

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl
Mitwirkung: Dipl.-Ing. Max Arends, Projektass.

Wien, 26.01.2012

(Unterschrift Amit Gahlay, BSc)

(Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Analyse des Tagging-Verhaltens am Beispiel von Flickr: The Commons

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Amit Gahlay, BSc
Registration Number 0625608

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl
Assistance: Dipl.-Ing. Max Arends, Projektass.

Vienna, 26.01.2012

(Signature of Author)

(Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Amit Gahlay, BSc
Hirschengasse 1, 1060 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Amit Gahlay, BSc)

Danksagung

Ich möchte mich bei meinem Betreuer Dieter Merkl bedanken, der mich durch seine zahlreichen inspirierenden Vorträge dazu bewegt hat, in diesem Bereich eine Diplomarbeit zu schreiben.

Ich möchte auch Max Arends für seine gute Betreuung danken. Es war wirklich eine große Freude, mit diesen Betreuern zusammen zu arbeiten.

Für die beste lebenslange Unterstützung die man sich vorstellen kann, möchte ich mich auch bei meinen Eltern, Sham und Usha Gahlay bedanken. Sie haben all dies hier überhaupt möglich gemacht und immer an mich geglaubt.

Ich möchte mich auch bei Zahra Maravandi für ihre liebevolle Unterstützung, ihre Motivation und der anregenden Diskussion bedanken.

Abstract

Social Tagging services such as *Flickr*[1] and *Del.icio.us*[2] belong, in addition to services like *Facebook*[3] and *YouTube*[4] to the stars of the Web 2.0 era. They allow users to tag different ressources with free chosen keywords. Social tagging offers an easy, intuitive and high quality way of indexing, which leads to a good description of the content and the improvement of search results, even when individual tags are classified wrong. Due to the large number of users, false information or abuse of the system is almost impossible.

Flickr created the project *The Commons* in 2009 which has two essential objectives. The first goal of *The Commons* is to show people hidden treasures in the world's public photography archives, and secondly to show how their input and knowledge can help make these collections even richer (Flickr, 2011*c*).

At "Flickr: The Commons" especially the association between the tagger and the photo is crucial. The user only tags the photo with a keyword if he can identify himself in any way with the shown object. Exactly this association relationship between the tagger and the tagged object plays a big part in this master thesis.

The first part of this thesis describes the theoretical foundations of social tagging including different tagging systems and basic knowledge of folksonomy. The second part gives a short prepartion for the practical part of this thesis by introducing "Flickr: The Commons" and different tools. Also the own developed tools *FlickrMaps* and *ClusterTags* are introduced in this part. *FlickrMaps* is a real-time mashup that combines Google Maps and Flickr. Whereas, *ClusterTags* works with a previously created database and groups keywords by geographical features.

In the last part of this thesis the data of *The Commons* is analyzed and the results are presented. Both applications FlickrMaps and ClusterTags, are used in this chapter to analyze the geographical relationships between the tagger and institutions. The application *ClusterTags* , is used to calculate the geographical influence of all cultural institutions. A globale influence was calculated for four institutions. With the application FlickrMaps the relationship between the

[1] `http://www.flickr.com/` [Zugriff am 01.11.2011]
[2] `http://www.http://www.delicious.com/` [Zugriff am 01.11.2011]
[3] `http://www.facebook.com/` [Zugriff am 01.11.2011]
[4] `http://www.youtube.com/` [Zugriff am 01.11.2011]

recording location of an image and the origin of the taggers were investigated. At this point, the investigation showed that images with an recording location have much more local tags then images without an recording location.

In addition the institutions, as well as pictures, keywords and comments are analyzed in greater details. To find popular institutions and pictures, a self definied *popularity index* is applied. This index normalizes the number of tags, comments, taggers and commentators based on their maximum values and is robust against *power users*.

The tags were examined using a frequency analysis and classified depending on the number of characters into five classes. It was shown, that the proportion of tags, which were assigned by the users, where increasing from class to class. In class 5 this value reached 90%, this means 90% of all tags were assigned by the users and only the remaining 10% were assigned by institutions. Also a correlation between the frequency of popular tags and their character-length was investigated and explained using the *Zipf-Law* (Wikipedia, 2011*i*). From the comments, keywords were extracted and compared to the tags of a picture. In this case a match of 10% was detected, this means that 10% of the extracted keywords from the comments could also be found in the tags assigned to the image. It is also shown in this chapter, that comments are very often used as an indicator for wrong information. To determine this, the comments were scanned for certain words like e.g. *false, wrong*. As a result, every fourth comment contained one of these words.

Finally a conclusion of the results and an outlook on potential further research and analysis are given.

Kurzfassung

Social Tagging-Dienste wie *Flickr*[5] und *Del.icio.us*[6] gehören, neben Diensten wie *Facebook*[7] und *YouTube*[8], zu den Stars der Web 2.0-Ära. Sie ermöglichen ihren Benutzern, diverse Ressourcen, mit frei gewählten Schlagwörtern zu versehen. Das Social Tagging bietet eine sinnvolle, intuitive und qualitativ hochwertige Verschlagwortung, welche zu einer guten inhaltliche Beschreibung der Informationsobjekte und einer Verbesserung von Suchergebnissen führt, selbst wenn einzelne Tags unpassend oder sogar als kritisch einzustufen sind. Aufgrund der großen Anzahl an Nutzern entsteht ein mächtiges Kontrollsystem, welches falsche Informationen bzw. Missbrauch des Dienstes nahezu unmöglich macht.

Durch die Initiative von Flickr entstand 2009 das Projekt „The Commons“ welches zwei Hauptziele verfolgt. Zum einen, möchte man einer breiten Masse die öffentlichen Fotoarchive dieser Welt zugänglich machen und zum anderen möchte man diese Sammlung durch deren Wissen bereichern (Flickr, 2011*c*).

Bei „Flickr: The Commons“ ist vor allem die Assoziationsfähigkeit zwischen Tagger und dem Foto ausschlaggebend d.h. der User weist dem Foto nur dann ein Schlagwort hinzu wenn er sich in irgendeiner Form mit dem gezeigten Objekt identifiziert. Genau diese Assoziationsbeziehung zwischen dem Tagger und dem abgebildeten Objekt, wird im Rahmen dieser Arbeit näher untersucht werden.

Im ersten Teil setzt sich die Arbeit mit den theoretischen Grundlagen von *Social Tagging* auseinander. Der zweite Teil dient als Vorbereitung auf den praktischen Teil. In diesen Abschnitt werden die Daten sowie die verwendeten Tools näher erläutert. Im dritten praxisbezogenen Teil der Arbeit wird anhand von „Flickr: The Commons“ die Assoziationsbeziehung zwischen Tagger und dem Objekt näher untersucht. Hier kommen neben bereits vorhandenen Tools, auch zwei eigenes entwickelte Anwendungen FlickrMaps und ClusterTags zum Einsatz. Bei FlickrMaps handelt es sich um ein Echtzeit-Mashup, welches Flickr und GoogleMaps verbindet. Dahingegen arbeitet ClusterTags mit einer vorher erstellten Datenbasis und gruppiert Schlagwörter nach geografischen Merkmalen.

[5] `http://www.flickr.com/` [Zugriff am 01.11.2011]
[6] `http://www.http://www.delicious.com/` [Zugriff am 01.11.2011]
[7] `http://www.facebook.com/` [Zugriff am 01.11.2011]
[8] `http://www.youtube.com/` [Zugriff am 01.11.2011]

Beide Anwendungen werden in diesem Kapitel zur Analyse der geografischen Zusammenhänge zwischen Tagger und den Institutionen verwendet. Im Detail wurde, ClusterTags für die Berechnung des geografischen Einflusses aller Kulturinstitutionen verwendet. Bei vier Einrichtungen konnte ein globaler Einfluss nachgewiesen werden. Mithilfe der Applikation FlickrMaps, wurde der Zusammenhang zwischen Aufnahmeort des Bildes und Herkunft der Tagger untersucht. Hier konnte auch gezeigt werden, dass Bilder mit Angabe eines Aufnahmeortes in Summe viel mehr lokale Tags aufweisen als Bilder ohne Aufnahmeort.

Neben diesem Zusammenhang wurden auch die Institutionen, Bilder sowie Schlagwörter und Kommentare im Detail untersucht. Bei den Institutionen und Bildern wurde ein eigener *Beliebtheitsindex* angewendet um die Populärsten aus der Datenbasis zu berechnen. Dieser Index normiert die Anzahl der Tags, Kommentare, Tagger und Kommentatoren auf Basis der jeweiligen Maximalwerte und ist dadurch robust gegenüber *Powerusern*.

Die Schlagwörter wurden mittels einer Häufigkeitsanalyse untersucht und abhängig von der Anzahl ihrer Zeichen in fünf Klassen eingeteilt. Bei diesen fünf Klassen wurde festgestellt, dass der Anteil der Tags welche von den Benutzern vergeben wurden, von Klasse zu Klasse immer höher wurde, bis schlussendlich 90% der Tags von den Benutzern stammen. Die restlichen Tags wurden von den Institutionen vergeben. Bei den Schlagwörtern, wurde auch ein Zusammenhang zwischen der Häufigkeit bestimmter Tags und deren Anzahl an Zeichen hergestellt und mit dem *Zipfschen Gesetz* erklärt (Wikipedia, 2011*i*).

Bei den Kommentaren wurde ebenfalls eine Häufigkeitsanalyse durchgeführt und die Anzahl der Zeichen analysiert. Weiters wurden Schlagwörter aus den Kommentaren extrahiert und mit den Schlagwörtern eines Bildes verglichen. Hier wurde eine Übereinstimmung von 10% nachgewiesen, d.h. die extrahierten Schlagwörter waren in 10% der Fälle auch in den vergebenen Tags der Bilder vorhanden. Außerdem konnte in diesem Kapitel gezeigt werden, dass Kommentare sehr häufig auf falsche Informationen hinweisen. Um dies festzustellen, wurden die Kommentare nach bestimmten Wörtern (*false, wrong*) abgesucht. Jedes vierte Kommentar enthielt einer dieser beiden Wörter.

Abschließend folgt eine Schlussfolgerung und ein Ausblick auf eine mögliche Verwendung der Ergebnisse.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

KAPITEL 1

Einleitung

1.1 Einleitung

Die Erhaltung des kulturellen Erbes ist zum wichtigsten Anliegen vieler Kulturinstitutionen geworden. Aufgrund der begrenzten räumlichen Ressourcen und steigender Anzahl an Artefakten, stehen alle Institutionen im Zwiespalt zwischen der Bewahrung alter und gleichzeitiger Präsentation neuer Kunstwerke. Um diesen Konflikt zu Umgehen haben einige Museen mit der Digitalisierung ihrer Fotoarchive begonnen. Die Digitalisierung der Bilder und Bereitstellung im Web löst diesen Konflikt zwischen Bewahrung und Präsentation.

Durch die Initiative von Flickr entstand 2009 das Projekt *The Commons*[1], welches zwei Hauptziele verfolgt. Zum einen, möchte man einer breiten Masse die öffentlichen Fotoarchive dieser Welt zugänglich machen und zum anderen möchte man diese Sammlung durch deren Wissen bereichern. Das *The Commons* - Projekt nutzt vor allem den Grundgedanken der kollektiven Intelligenz, welcher im Web 2.0 auf verschiedene Arten realisiert werden kann. Neben Wikis , Blogs, Foren und anderen Social Media werden im Web vor allem auch Social Tagging Dienste für die Zusammenarbeit genutzt. Das Social Tagging bietet eine qualitativ hochwertige und intuitive Verschlagwortung, welche meist zu einer guten inhaltliche Beschreibung der Informationsobjekte führt. Diese Beschreibung wiederum führt zu besseren Suchergebnissen, selbst wenn einzelne Tags unpassend oder sogar als kritisch einzustufen sind (Lackes, 2009; Zhu and Wu, 2009). Aufgrund der großen Anzahl an Nutzern entsteht ein mächtiges Kontrollsystem, welches falsche Informationen bzw. Missbrauch des Dienstes nahezu unmöglich macht.

Bei *Flickr Commons* könnte die persönliche Beziehung zwischen Tagger und dem Foto ausschlaggebend sein d.h. der User weist dem Foto nur dann ein Schlagwort hinzu wenn er sich in irgendeiner Form mit dem gezeigten Objekt identifizieren kann.

[1] http://www.flickr.com/commons

Die Erforschung dieser Verbindung bringt für Kultureinrichtungen große Vorteile mit sich. Einerseits können Artefakte, welche ein hohes Interesse durch die *Usercommunitiy* aufweisen festgestellt werden und anderseits kann mit sehr einfachen Mitteln, Werbung gezielt an potenzielle Besucher gebracht werden.

Genau diese persönliche Beziehung zwischen dem Tagger und dem abgebildeten Objekt, soll im Rahmen dieser Diplomarbeit näher untersucht werden.

1.2 Problemstellung und erwartetes Resultat

Um diese Beziehung genauer durchleuchten zu können, müssen verschiedene Abhängigkeiten berücksichtigt werden. Beispielsweise ist ein Motiv für das Tagging, die geografische Abhängigkeit zwischen dem Tagger und dem Bild.

Wobei hier auch verschiedene Varianten zu unterscheiden sind. Einerseits könnte ein direkter Zusammenhang zwischen Aufnahmeort und dem User bestehen d.h. das Foto wurde in unmittelbarer Umgebung des Taggers aufgenommen bzw. das Foto ist in einem naheliegenden Museum ausgestellt. Andererseits existiert auch ein indirekter Zusammenhang, welcher sich nicht sofort auf eine geografische Abhängigkeit zurückführen lässt z.B. Tourismus.

Neben der geografischen Abhängigkeit soll auch der Einfluss von Kulturinstitutionen ermittelt werden. Dieser Einfluss hängt von der geografischen Verteilung der Tagger rundum die Institutionen ab. Gibt es viele Tagger aus anderen Ländern oder ist das Interesse an dieser Kulturinstitution nur lokal?

Auch die Häufigkeit der verwendeten Schlagwörter soll untersucht werden. Bei dieser Analyse soll geklärt werden, ob häufige Tags generelle bzw. grobe Aspekte eines Bildes beschreiben. Wohingegen, nicht so häufige Tags eher dazu verwendet werden spezielle bzw. detaillierte Eigenschaften eines Objektes aufzuzeigen. Auch die Fragestellung der chronologischen Ordnung von detaillierten bzw. generellen Tags kann mithilfe einer Häufigkeitstabelle geklärt werden.

Neben den Schlagwörtern steckt auch in den Kommentaren sehr viel Information über die gezeigten Objekte. Durch Anwendungen von Indexverfahren sollen gezielt Informationen aus den Kommentaren extrahiert werden. Bei diesen Verfahren werden alle Wörter, bis auf Stoppwörter, in den Index aufgenommen. Danach wird mittels Stemming ein gemeinsamer Wortstamm gebildet, welcher als Basis für sämtliche Analysen dient.

1.3 Aufbau der Arbeit

Diese Arbeit gliedert sich in einen theoretischen und praktischen Teil. Im theoretischen Teil werden allgemeine Informationen im Bezug auf *Social Tagging* erläutert. Danach werden verschiedene Werkzeuge und Tools im Bereich Tag-Analyse und Textmining beschrieben. Unter Berücksichtigung dieser Grundlagen wird dann im praktischen Teil die eigentliche Untersuchung

der Beziehung zwischen Tagger und dem getaggten Objekt durchgeführt. Vor dem praktischen Teil werden die zwei selbst entwickelten Tools *FlickrMaps* und *ClusterTags* erläutert und für die empirische Analyse verwendet. Ausgehend von den Ergebnissen und Vergleichen werden schließlich die Hypothesen bestätigt oder verworfen.

KAPITEL 2

Grundlagen Social Tagging

2.1 Social Tagging

2.1.1 Einleitung

Der Mensch nutzte schon immer die Sprache um seinen Mitmenschen Dinge zu beschreiben. Um die gesprochenen Wörter auch unabhängig vom Redner an andere zu übermitteln und sie für die Zukunft festzuhalten, entwickelte sich in Laufe der Geschichte die Schrift. Mithilfe der Schrift wurde es auf einmal möglich eine Nachricht zu erstellen und über verschiedenste Medien zwischen Sender und Empfänger zu übertragen (GMW, 2008). Auch Notizen sind nichts anderes als kleine Nachrichten, welche uns an einen späteren Zeitpunkt an etwas erinnern sollen.

Aufbauend auf der Idee, Notizen direkt an die Objekte anzubringen für welche sie gedacht sind, wurde 1968 das *Post-it* erfunden. Das *Post-it* fand mithilfe von Web 2.0 und Social Tagging ihren Weg in die digitale Welt (GMW, 2008).

2.1.2 Web 2.0 und Social Tagging

> Web 2.0 is the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them. This is what I've elsewhere called harnessing collective intelligence. (O'Reilly, 2006)

In den letzen 15 Jahren hat sich das Web von einem Werkzeug für Wissenschaftler am CERN, zu einem globalen Informationsraum für mehrere Milliarden Menschen entwickelt. Das *Web 2.0* brachte die große Veränderung des Internets mit sich. Der Begriff *Web 2.0* wurde 2004 zum ersten Mal von Dale Dougherty und Craig Cline verwendet und bezeichnet keine Neuentwicklung des Webs, vielmehr repräsentiert er eine Sammlung von Methoden bzw. Techniken, welche die Evolution des Internets mit sich brachte (Sen et al., 2006). Typische Beispiele für Web 2.0 Techniken sind z.B. RSS, Wikis, Weblogs, Social Networking und Social Tagging. Ein

wesentlicher Unterschied zwischen Web 1.0 und Web 2.0 ist die dynamische Komponente. In Web 1.0 wurden Informationen meist von einer Quelle bezogen und waren nur durch den *Administrator* veränderbar. In Web 2.0 werden Informationen nicht nur von einer Quelle, sondern von mehreren Quellen und das in Echtzeit bezogen (z.B. RSS). Weiters hatten in Web 1.0 die Nutzer keinerlei Möglichkeit den Inhalt einer Seite zu verändern. Sie waren nur Konsumenten des Inhalts (Top-Down-Prinzip) (Lewis, 2006). In Web 2.0 wird der Benutzer aktiv beteiligt und dadurch auch zum Produzenten des Inhalts gemacht (Buttom-Up-Prinzip). Viele Web 2.0 Dienste unterstützen diese Einbeziehung der Nutzer und werden deswegen auch als „Social Software" bezeichnet. In Web Terminologie ist Social Tagging eine nicht hierarchische freie Vergabe von Schlagwörtern mit deren Hilfe eine Ressource im Internet beschrieben wird.

Einer der Pioniere von Social Tagging ist Joshua Schachter, der Gründer von *Deli.cio.us* [1], einem Social Bookmarking Service (Kammergruber, 2009). Clay Shirky, Professor an der New York University, wird vom Guardian [2] folgendermaßen zitiert:

> Clay Shirky [...] studied tagging and advised Delicious. He describes Schachter as, the first person to figure out the **social value** of labeling. Any one person's labels are **messy, inconsistent and partial**, and are therefore much less valuable than formal classification systems. However, if there is a way to aggregate those labels, and therefore their value, they become more valuable than formal systems, because they are **robust, socially accurate and cheap.** (Norton, 2006)

Diese relativ gut komprimierte Definition beinhaltet alle wesentlichen Aspekten von Social Tagging. Tags haben im Vergleich zu klassischen Ordnungssystemen (Ontologien, Thesauri) einen erheblichen Nachteil. Sie werden meistens chaotisch (messy) vergeben und haben einen Mangel an Struktur und Konsistenz. Social Tagging weist auf den ersten Blick sehr große Schwächen bei der Klassifikation von Objekten auf. Aggregiert man jedoch Tags zusammen entsteht ein oft unterschätzter sozialer Wert (social value), welcher das gemeinschaftliche Indexieren in Summe wertvoller als formelle Ordnungssysteme macht. Betrachtet man die Verteilung der Schlagwörter über eine Menge von Benutzern so ergibt sich eine Verteilung die dem Potenzgesetz folgt (Cattuto et al., 2007).

Die breite Akzeptanz von Social Tagging lässt sich auf einige Aspekte zurückführen (Lackes, 2009):

1. intuitive, freie Verschlagwortung

2. sehr große Anzahl an Nutzern (mächtiges Kontrollsystem) dadurch wird ein Missbrauch des Dienstes nahezu unmöglich

3. einfache und kostengünstige Erstellung von Tags

[1] `http://www.delicious.com/` [Zugriff am 1.11.2011]

[2] `http://www.guardian.co.uk/media/2006/jan/26/newmedia.technology1` [Zugriff am 2.11.2011]

2.1.3 Tagging von Ressourcen

Betrachtet man Tagging aus der technischen Sicht, so ist das Taggen eines Objektes mit dem Erzeugen einer Instanz auf Datenbankebene verbunden. Diese Instanz ist in der Regel ein n-Tupel folgender Form (GMW, 2008):

Definition 1: Ein Tag-Instanz ist ein n-Tupel bestehend aus:

- $(Objekt_j, Beschriftung_a, Benutzer_x, \ldots.)$

Das Objekt stellt die Ressource im Web da, welche mittels URL eindeutig identifiziert werden kann. Neben Bildern können im Internet auch Webseiten, Dokumente und multimediale Ressourcen mit Tags versehen werden.
Das Vergeben von Tags (Label) erfolgt in den meisten Fällen ohne vorgeschriebene Regeln. Einige Systeme verwenden Recommender Systeme um dem Benutzer Schlagwörter vorzuschlagen (Leimstoll and Stormer, 2007).
Der Parameter User identifiziert den Tagger und muss immer mitgeführt werden, um die Mehrfachvergabe von Tags pro User und Objekt zu verhindern. Im Gegensatz zu hierarchischen Ordnungssystemen wo ein Objekt genau einem Label zugeordnet wird, ist beim Tagging die Zuordnung zu mehreren Labels von signifikanter Bedeutung. Beispielsweise würde man dieses Dokument in klassischen Ordnungssystemen in den Ordner Diplomarbeit legen. Im Gegensatz dazu vergibt man beim Taggen dieser Arbeit mehrere Schlagwörter wie z.B. Diplomarbeit, Social Tagging, TU Wien.
Diese Mehrfachzuordnung gilt sowohl für Dokumente als auch für Schlagwörter. Ein Dokument kann durch mehrere Schlagwörter beschrieben werden aber ein Schlagwort kann auch mehrere Dokumente beschreiben. Aus technischer Sicht ist ein Tag ein Paar bestehend aus Label und eine Menge an Objekten auf die es verweist (GMW, 2008):

Definition 2: Ein Tag ist ein Paar aus:

- $(Beschriftung_a, Objekt_j \mid \exists(Objekt_j, Beschriftung_a, Benutzer_x))$

Ein Tag (Kante) verbindet alle Objekte (Knoten), welche mit dem Tag versehen wurden, miteinander. Aus dieser Verbindung resultiert ein sogenannter Hyphergraph[3]. In Abbildung 2.1 ist der gesamte Vorgang des Social Taggings abgebildet. Die beiden Definitionen können hier mit den Punkten a und b der Abbildung gleichgesetzt werden.

Das Endergebnis des *Social Taggings* ist eine Folksonomie, die sich aus allen vergebenen Tags zusammensetzt und sehr häufig als *Tag Cloud* (Siehe Kapitel 2.3.1) dargestellt wird. Vollständigkeitshalber wird hier auch die Definition der Folksonomie angegeben, im Detail wird auf diesen Begriff im Kapitel 2.1.9ff eingegangen.

(Schmitz et al., 2006) gibt eine formale Definition von Folksonomien an:

[3] Ein hypergraph G kann als paar (V,E) definiert werden, wobei V die Menge der Knoten und E die Menge der Hyperkanten entspricht. Jede Hyperkante stellt die Relation zwischen mehreren Knoten her: E $\subseteq$ u, v, ... $\in 2^V$ (Hyperkanten sind ungerichtet)

Definition 3: Eine Folksonomie ist ein Tupel $\mathbb{F}$:= *(U, T, R, Y, $\prec$)* wobei:

- U,T und R endliche Mengen sind, deren Elemente man *Benutzer, Tags bzw. Ressourcen* nennt.
- Y ist eine ternäre Relation zwischen diesen Mengen, d. h Y $\subseteq$ U$\times$T$\times$R, deren Elemente Tags *Assignments* heißen.
- $\prec$ ist eine benutzerspezifische Unter-/Obertag-Relation, d.h $\prec \subseteq$ U$\times$T$\times$T, die Is-A Relation genannt wird.

Die Personomie P_u eines Users u$\in$U ist die Beschränkung von $\mathbb{F}$ auf u, d. h P_u := (T_u, R_u, I_u, $\prec_u$) wobei I_u := (t,r) $\in$ T$\times$R|(u,t,r) $\in$ Y, T_u := $\pi_1(I_u)$ und R_u := $\pi_2(I_u)$ gilt; dabei bezeichnet π_i die Projektion auf die i-te Dimension. Wenn man die *Is-A-Relation* nicht betrachten möchte, kann man die Folksnomie als ein Quadrupel $\mathbb{F}$:= *(U, T, R, Y)* notieren. Diese Struktur ist in der formalen Begriffsanalyse (Ganter and Wille, 1999) *triadischer Kontext* bekannt. Eine äquivalente Sicht ist die eines tripartiten ungerichteten Hypergraphen $G = (V, E)$, wobei $V = U \bigcup T \bigcup R$ die Menge der Knoten und $E = u, t, r \mid (u, t, r) \in Y$ die Menge der Hyperkanten sind (Schmitz et al., 2006).

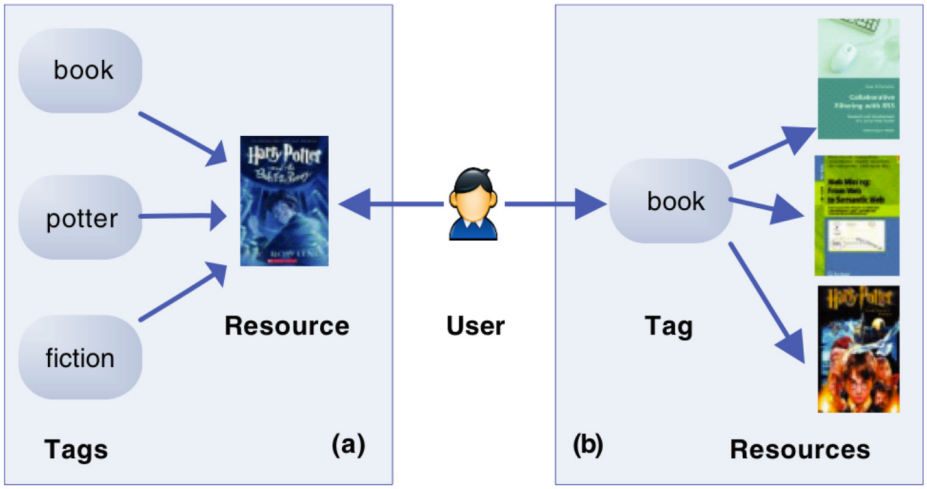


Abbildung 2.1: Social Tagging: (a) der User findet eine Ressource im Internet (z.B. ein Buch) und versieht dieses mit Schlagwörtern; (b) User verwendet eine Sammlung von Schlagwörtern (z. B. Tagclouds) um Content zu filtern; (Zhang and Liu, 2010)

2.1.4 Arten von Tags

Tags sind meistens einfache Schlagwörter, die als Metadaten verwendet werden und beschreiben oft den Inhalt oder die Art und Weise wie eine Ressource charakterisiert wird. Nach Golder und Huberman (Golder and Huberman, 2005) gibt es sieben Arten von Tags:

1. **Beschreibung des Inhalts / Identifying what (or who) it is about**:
   Diese Art von Tags nehmen direkten Bezug auf das Themengebiet einer Ressource. Sie beschreiben kurz und prägnant den Inhalt einer Arbeit. In den meisten Fällen handelt es sich bei diesen Tags um Nomen. Beispielsweise könnte diese Arbeit mit den folgenden Tags versehen werden: *social tagging, fickr the commons, tag clustering, tagging verhalten*.

2. **Beschreibung der Art / Identifying what it is:**
   Diese Schlagwörter werden dazu verwendet um die Art einer Ressource zu kennzeichnen. Weitverbreitet sind diese Tags vor allem bei Social Bookmarking Diensten wie *Del.icio.us*, da hier Webressourcen mit Schlagwörtern versehen werden und im Vorhinein nicht klar ist um welche Art von Ressource es sich eigentlich handelt.
   Diese Arbeit würde man mit dem Tag *diplomarbeit* kennzeichnen.

3. **Besitzer (oder Autor) eines Objektes / Identifying who owns it:**
   Diese Tags stellen den Bezug zum Urheber bzw. Autor des Dokuments her. Wobei es sich hier neben einzelnen Personen auch um Organisationen oder Personengruppen handelt.

   Auf diese Arbeit treffen in diesem Fall folgende Tags zu: *tu wien, gahlay amit* .

4. **Kategorie Verfeinerung / Refining Categories:**
   Diese Tags werden dazu verwendet um die Kategorie eines Dokuments zu „verfeinern“oder um detailliert auf eine bestimmte Facette einer Ressource einzugehen. Diese Schlagwörter besitzen allein genommen keine Aussagekraft, aber betrachtet man alle Tags zusammen so lässt sich das Dokument ganz präzise kategorisieren. Ein gutes Beispiel für solche Schlagwörter sind Zahlen, welche erst mit einer Mengen bzw. Maßangabe eine relevante Bedeutung bekommen (z. B. 10kg, 100liter).

   Refining categories Tags für diese Arbeit sind neben *diplomarbeit* auch *tu wien, 066926* [4]*, isis*[5].

5. **Qualität oder Charakteristika / Identifying Qualities or Characteristics:**
   Ein Benutzer verwendet diese Tags um seine subjektive Meinung über eine Ressource im Web auszudrücken. Speziell für Fotosharing Plattformen wie z. B. Flickr sind diese Tags von großer Bedeutung, da sie dem Benutzer ermöglichen Charakteristika von Fotos zu beschreiben. Bei anderen Ressourcen kennzeichnen Tags dieser Klasse den Dokumentinhalt durch Adjektive wie zum Beispiel *schön, gut*.

[4] Studienkennzahl für Masterstudium Wirtschaftsinformatik
[5] Institute of Software Technology and Interactive Systems

6. **Selbstreferenzierung / Self Reference:**
   Diese Art von Tags werden vom Benutzer verwendet um eine Beziehung zwischen dem Dokument und ihm selbst darzustellen. Sie bestehen meistens aus einem Possessivpronomen (mein, dein, etc.) und einer Beschreibung zum Typ des Dokuments z. B. meineArbeit , meinBlog.

7. **Arbeitsorganisation / Task Organizing:**
   Hiermit sind Schlagwörter gemeint die dem Benutzer dabei helfen Aufgaben wie Lesen oder Schreiben mit einem Dokument zu verbinden. In der englischen Sprache beginnen diese Tags meistens mit *to* z. B. toRead oder toWrite.

Anhand dieser Tag-Kategorien wird erkennbar, dass einige Tags einen Mehrwert für die Allgemeinheit besitzen, wohingegen andere lediglich dem Tagger selbst einen Nutzen verschaffen. Die ersten 4 Arten von Tags sind für die Allgemeinheit bestimmt und bilden einen sozialen Mehrwert. Die letzten 3 Arten sind für den persönlichen Gebrauch des Taggers gedacht.

2.1.5 Ausführungen von Tags

Im folgenden Punkt werden die verschiedenen Ausführungen von Schlagwörtern, welche sich im *World Wide Web* durchgesetzt haben, näher erläutert.

- **Maschinenlesbare Tags**
  Ein *Machine* oder *Triple* Tag kann ebenso wie ein normaler Tag von einem User geschrieben werden. Diese Tags enthalten ein spezielles Format um zusätzliche Informationen an ein Objekt anzuheften und dadurch diese auch für Maschinen verständlich zu machen. Durch diese zusätzliche Information wird die Maschinenlesbarkeit der Tags ermöglicht. Der Aufbau eines *Machine* Tag ähnelt einem *RDF*[6]-Element.

  Ein sehr gutes Beispiel für eine Webseite, welche *Machine*-Tags von *Flickr* auswertet ist *Upcoming.org*[7]. Bilder auf Flickr, welche zu einer Veranstaltung gehören, können durch den Machine-Tag „upcoming:event=428011“ gekennzeichnet werden. Danach sucht der Server von *Upcoming.org*, nach den Tags welche mit „upcoming: “ beginnen und kann dadurch Videos und Fotos einer Veranstaltung zuordnen.

- **Geo Tags**
  Eine der häufigsten Formen von *Machine*-Tags stellen *Geo Tags* dar. Vor allem bei *Flickr* und anderen Fotosharing Plattformen, spielen diese eine sehr große Rolle, da sehr viele Kameras bereits bei den Bildern die Ortskoordinaten dazu speichern (Flickr, 2011*b*). Beim *Geotagging* werden folgende drei Tags zu einer Ressource hinzugefügt:

  [-] Tag *geotagged* um zu Kennzeichnen, dass das Objekt Ortsinformationen enthält

  [-] Tag *geo:lat=12.345678* Breitengrad

[6] Resource Description Framework - `http://www.w3.org/RDF/` [Zugriff am 19.01.2012]
[7] `http://www.upcoming.org` [Zugriff am 17.11.2011]

[-] Tag *geo:lon=23.455511* Längengrad Eine Software muss am Beginn nur nach dem ersten Tag suchen und kann dadurch feststellen, ob es zu diesem Objekt Längen- und Breitengrad Koordinaten gibt (Smith, 2008).

- **Hash Tags**
  Hash Tags kommen vorwiegend bei *Twitter*[8] zum Einsatz und können eigentlich mit Stichwörtern verglichen werden. Diese Tags werden mit dem Doppelkreuz (# in Englisch *hash*) eingeleitet, worauf sich auch der Name dieser Tags zurückführen lässt (z. B. #twitter). Der größte Unterschied zu anderen Tags ist, dass diese Art von Tags in der Nachricht selbst enthalten sind (Wikipedia, 2011*c*).

2.1.6 Motivation

Ob ein Tagging-Service erfolgreich ist, hängt von der Motivation der Anwender und den Anreizen, die ein System bietet, ab. In diesem Abschnitt soll die grundlegende Frage warum Anwender überhaupt Schlagwörter vergeben, geklärt werden. Nach (Marlow et al., 2006) kann man die Motivationen der User, in zwei Klassen (*Organisation* und *Sozial*) unterteilen. Bei der ersten Klasse verschafft sich der Nutzer eine strukturierte Ablage für Ressourcen. Zum anderen gibt es auch User, die sozial motiviert sind und ihre Meinung zu einer Ressource mittels Tags wiedergeben wollen. (Marlow et al., 2006) untersuchten einige Social-Tagging-Systeme und stellten dabei folgende Motivationsfaktoren fest:

- **Zukünftige Suche / Future Retrieval**
  Das Kennzeichen von Ressourcen zur späteren Verwendung, zum Beispiel für interessante Unterlagen für eine Diplomarbeit oder ein gutes Lied. Hier spielen Tags der Kategorie *Arbeitsorganisation* (Siehe Kapitel 2.1.4) eine wichtige Rolle.

- **Beitragen und Inhalte teilen / Contribution and Sharing**
  Hier ist die ausschlaggebende Motivation, Ressourcen für andere Benutzer im Web bereitzustellen. Ein Beispiel hierfür ist das Taggen von Fotos für Freunde und Bekannte.

- **Aufmerksamkeit gewinnen / Attract Attention**
  Um mehr Aufmerksamkeit auf eine Ressource zu ziehen, versehen Benutzer gerne ihre Dokumente mit Tags, welche sehr häufig genutzt werden, auch wenn diese das Dokument gar nicht beschreiben.

- **Teilnehmen und Wettkampf / Play and Competition**
  Einige *Poweruser* sehen es als Motivation, ihre eigenen Tags besonders beliebt zu machen, indem sie soviele Objekte wie möglich mit ihren Tags versehen. Auf Flickr beispielsweise gibt es sogar eine eigene Gruppe „squad circle“, welche ihre Mitglieder auffordert alle Bilder mit runden Formen auf ein Quadrat zuzuschneiden und mit dem Tag „squad circle“ zu kennzeichnen. Der Anreiz besteht hier bei der Manipulation des Systems und der Populäritätssteigerung ihres eigenen Begriffs und somit ihrer Ressource.

[8] `http://www.twitter.com` [Zugriff am 17.11.2011]

- **Selbstpräsentation / Self Presentation**
  Hier möchte man durch die Annotierung den Bezug zwischen Nutzer und der Ressource herstellen. Zum Beispiel gibt es bei *Last.fm* den Tag „seen live“ wo die Nutzer angeben können welchen Künstler sie schon Live in einem Konzert erlebt haben. Es werden hier Tags der Kategorie *Selbstreferenzierung* verwendet (Siehe Kapitel 2.1.4).

- **Meinungsaustausch / Opinion Expression**
  Hier haben Benutzer das Bedürfnis, ihre Meinung über ein Objekt anderen mitzuteilen und benutzen Tags der Klasse *Qualität oder Charakteristika* z. B. funny, boring oder well-writen.

Die Motivationsfaktoren nach (Marlow et al., 2006) wurden in (Ames and Naaman, 2007) erweitert. In dieser Erweiterung wurde neben zahlreichen Social Tagging-Systemen auch die Fotosharing Plattform *Flickr* analysiert, wobei hier die Funktion Bilder direkt mittels Handy hochzuladen und mit Tags zu versehen, genauer betrachtet wurde. In (Ames and Naaman, 2007) wird ein Modell für die Tagging-Motivation beschrieben, welches sich in zwei Dimensionen gliedert. Durch die erste Dimension Sozialität (*Sociality*) wird beschrieben, ob die Ressource für den Eigen- bzw. Fremdgebrauch annotiert wurde. Die zweite Dimension Funktion (*Function*) baut auf der in (Marlow et al., 2006) getroffenen Unterteilung der Motivation in die Klassen organisational und sozial auf.

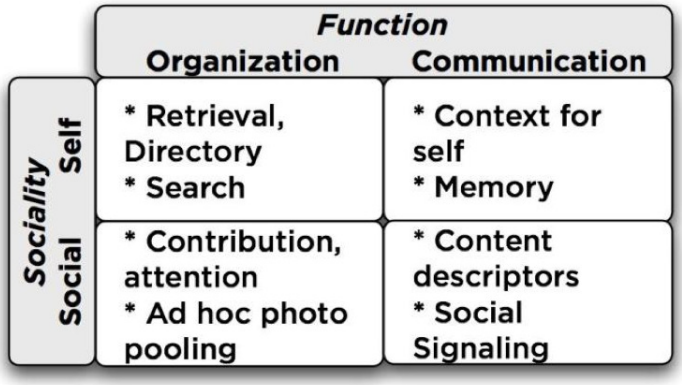


Abbildung 2.2: Motivationsfaktoren für Social Tagging nach (Ames and Naaman, 2007)

- **Self/Organization: Suche und Wiederauffindung / Search and Retrieval**
  Entspricht dem in (Marlow et al., 2006) erwähnten Motivationsfaktor *Future Retrieval*, wobei dieser hier nochmals um die Organisation von Ressourcen für den Selbst bzw. Fremdzweck erweitert wurde. Diese Dimension behandelt die Suche und die Organisation der Ressourcen für den Selbstzweck.

- **Self/Communication: Erringung und Kontext / Memory and Context**
  Bei diesem Punkt soll die Ressource durch die Vergabe von Tags einem anderen Kontext zugewiesen werden. Dies kann beispielsweise mit der Beschriftung von Bildern auf der

Rückseite verglichen werden, damit man auch zu einem späteren Zeitpunkt den Kontext (Person oder Ort) des Bildes wiedererkennt (Selbstzweck).

- **Self/Organization: Öffentliche Suche und Foto Sammlungen / Public Search and Photo Pools**
  Der Punkt *social/organization* in Abbildung 2.2 gibt die Motivation an, Ressourcen mit Tags zu annotieren damit sie später auch durch andere User (Fremdzweck) auffindbar sind. Zu dieser Sektion werden auch die Motivationsfaktoren *Contribution and Sharing*, *Attact Attention* und *Play and Competition* zugeordnet.

- **Social/Communication: Kontext und Signalisieren / Context and Signaling**
  Bei diesem Punkt wird auch der Kontext eines Bildes festgelegt aber diesmal nicht für den Eigenzweck sondern für andere. Beispielsweise wenn man seinen Freunden ein Fotoalbum auf Flickr zur Verfügung stellt und mit Tags den Kontext der Bilder beschreibt. Also wo wird das Bild aufgenommen, welche Personen sind auf diesem Bild zu sehen, wann wurde das Bild aufgenommen etc. Außenstehende können quasi auf die Tags zurückgreifen und die Geschichte des Bildes sofort nach zu vollziehen.

2.1.7 Social Tagging Systeme

Tagging-Systeme wie *Del.icio.us*, *Flickr*, etc. sind seit dem Aufkommen von Web 2.0 nicht mehr aus dem Internet wegzudenken (Siehe Tabelle 2.1). Sie bieten im Gegensatz zu traditionellen Klassifikationsmodellen erhebliche Vorteile. Der Tagging Prozess ist mit sehr geringem Aufwand verbunden und die eigentliche Arbeit wird auf die Benutzer des Systems aufgeteilt. Da es bei den meisten Systemen keine Regeln oder vorgeschriebenes Vokabular gibt, spiegeln Tags nicht nur fachliche Kompetenz sondern auch die Wahrnehmung und den sozialen und kulturellen Hintergrund der Benutzer wider (Begelman et al., 2006; Smith, 2008). Dieser Aspekt hat den Vorteil, dass die im System verwendeten Schlagwörter immer die aktuelle Sprache der User widerspiegeln. Das Vokabular ist, wenn genügend User das System verwenden, am aktuellen Stand und verändert sich mit der Sprachgewohnheit der Benutzer mit.

Alle Systeme bieten die Möglichkeit an, eigene User-Profile anzulegen. Diese Profile enthalten alle Daten und Beziehungen die durch das Tagging entstanden sind. Durch die Analyse dieser Profile werden Cluster generiert die dem Anwender bei der Suche unterstützen. Bei einigen *Social Tagging*-Plattformen wie beispielsweise *Flickr* und *Del.icio.us* sind auch Empfehlungssysteme im Einsatz, die dem Benutzer aufgrund seiner letzten Aktivitäten neue Ressourcen vorschlagen.

Im Großen und Ganzen fördern viele *Social Tagging*-Plattformen den Einstieg zwischen Suchen und Stöbern. Wer oft in diesen System unterwegs ist, entdeckt häufig unbekannte Ressourcen.

Ein weiterer wichtiger Aspekt sind die geringen Kosten und das unmittelbare Feedback, durch Betrachtung der anderen Tags bzw. anderen Ressourcen.

Social Tagging ist häufig in Systemen zu finden, welche folgende Ziele verfolgen (Weghuber, 2009; Smith, 2008):

- zur Verwaltung von Ressourcen
  - private Ressourcen (Dokumente z. B. in *GoogleDocs*, Mails in z. B. *Gmail*)
  - öffentliche Ressourcen (Webseiten z. B. auf *Del.icio.us*)
- zur Kategorisieren von Objekten
  - Sammeln digitaler Objekte (Fotos auf *Flickr*)
  - Verbesserung von Prozessen (z. B. Taggen von Produkten auf *Amazon*)

Plattform	**Beschreibung**
Del.icio.us `www.delicious.com`	Social Bookmarking Seite: Beschlagwortung von Webressourcen im Internet. Erzeugung einer kollektiven Linksammlung um die Navigation im Web zu erleichtern. (Siehe Kapitel 2.1.7.1)
Flickr `www.flickr.com`	Foto-Sharing: Online-Plattform zum Verwalten von digitalen Bildern, mittlerweile auch Videos. Taggen von Fotos um den Suchvorgang und die Kategorisierung zu verbessern.

Viddler
`www.viddler.com`
YouTube
`www.youtube.com`
	Video-Sharing: Ermöglicht das Veröffentlichen von Videos. Annotieren von Videos mittels *Global* und *Time* Tags (Siehe Kapitel 2.1.7.4)

Amazon `www.amazon.com`	E-Commerce-Plattform: User versehen Produkte mit Tags und erleichtern so dem System ähnliche Produkte schneller zu finden.
Last.fm `www.last.fm`	Musik-Datenbank: Verwaltung von Musik, durch Verschlagwortung von Songs, Alben und Künstlern.
GMail `www.gmail.com`	E-Mail-Dienst: Kennzeichnen von Mails mit Labels führt zur bessern Organisation.

CiteULike
`www.citeulike.com`
Bibsonomy
`www.bibsonomy.com`
	Social Bookmarking: Erweiterung von Social Bookmarking System - Taggen von wissenschaftlichen Dokumenten im Internet.

Tabelle 2.1: Beispiele für Social Tagging-Systeme.

In den nächsten zwei Abschnitten werden die Social-Tagging-Dienste *Del.icio.us*, *Flickr*, *Viddler* und *YouTube* näher erläutert.

2.1.7.1 Social Bookmarking

Beim Social Bookmarking wird dem Benutzer die Navigation im Internet durch kollektive Linksammlungen erleichtert (Golder and Huberman, 2005). Links werden nicht mehr lokal im Browser sondern auf einer Plattform gesammelt, wo sie auch durch andere Benutzer verwendet und mit Tags annotiert werden (Wikipedia, 2011*b*).

Social Bookmarking hat folgende Funktionsweise:

- Es wird ein zentraler Server zur Verwaltung von Hyperlinks verwendet
- Hyperlinks werden vom Benutzer mit Schlagwörtern versehen und so richtig kategorisiert

Einer der ersten Dienste dieser Art war *Del.icio.us*. Der Entwickler Joshua Schachter startete diesen Dienst Ende 2003 (Delicious, 2012). *Del.icio.us* ermöglicht seinen Benutzern persönliche Linksammlungen (Lesezeichen) anzulegen und mit Schlagwörtern oder Tags zu versehen (Abbildung 2.3). Diese Lesezeichen sind im Allgemeinen öffentlich sichtbar, jedoch kann man auch private Linksammlungen definieren, welche dann nur für bestimmte Benutzergruppen zugänglich sind (Hammond et al., 2005).

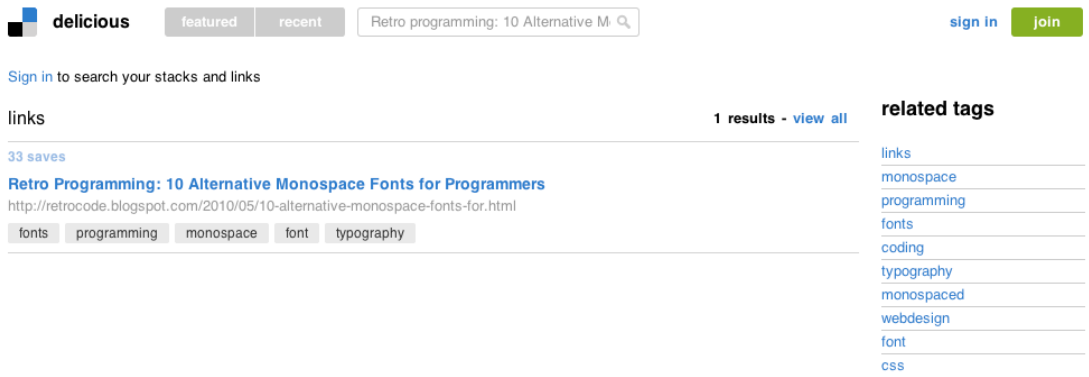


Abbildung 2.3: Screenshot von *Del.icio.us*. Jedes Lesezeichen hat einen Titel (z. B. *Retro Programming..*), eine URL und ist mit Tags (z. B. *fonts, programming, monospace, font, typography..*) annotiert

Eine Erweiterung von Social Bookmarking-Systemen stellen Dienste wie *CiteULike* oder *Bibsonomy* dar. Diese richten sich an eine wissenschaftlich orientierte Zielgruppe und bieten die Möglichkeit einer weitaus differenzierteren Beschreibung der Nachweise (Hotho et al., 2010). Neben wissenschaftlichen Publikationen bieten diese Webseiten auch andere elektronische Publikationsformen wie Zeitschriftenartikel und Bücher an.

Fast alle Systeme bilden anhand der vergebenen Tags Userprofile, mit deren Hilfe dann User mit ähnlichen Interessen berechnet werden. In Abbildung 2.4 sieht man einen Screenshot vom Social-Bookmarking Dienst *Bibsonomy*. Auf der linken Seite werden die Suchanfragen beantwortet und rechts werden anhand eines Userprofiles verschiedene Relationen wie beispielsweise *similar users*, meine *Tagcloud*, meine *Sphäre* (Freundschaftsliste) dargestellt.

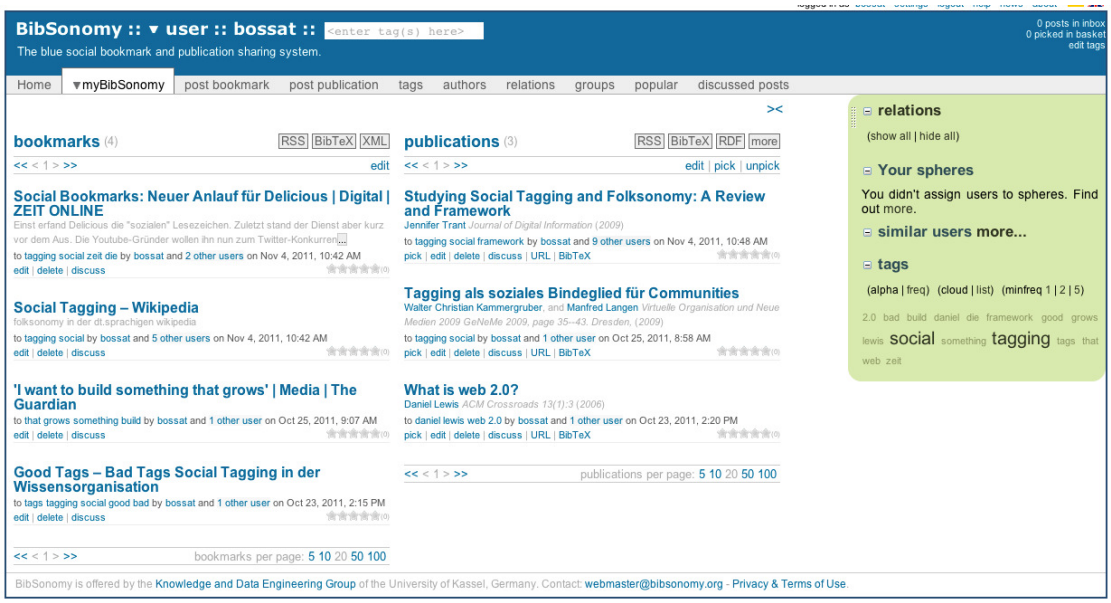


Abbildung 2.4: Screenshot von *http://www.bibsonomy.org/*

2.1.7.2 Social Tagging - Multimedia

Auch bei der Kategorisierung von Multimedia Ressourcen im Internet werden Tags verwendet. Einer der ersten Plattformen in diesem Bereich war *Flickr*. Dieser Dienst ist seit Ende 2002 im Internet zugänglich und nutzt Social Tagging zur Organisation von Fotos (Flickr, 2011*c*). Nach dem Erfolg von *Flickr*, wurden auch andere digitale Medien wie Video und Musik durch Social Tagging organisiert.

In den nächsten drei Kapiteln werden drei Vertreter aus diesem Bereich vorgestellt.

2.1.7.3 Flickr

Flickr wurde ursprünglich 2002 in Kanada entwickelt und hat ca. 40 Millionen registrierte Benutzer und gehört damit zu den erfolgreichsten Social Tagging Plattformen (Cha et al., 2009). Es ist eine Online-Plattform zum Verwalten von digitalen Bildern, mittlerweile auch Videos. Einer der grundlegenden Aspekte warum *Flickr* so erfolgreich ist, ist die Vernetzung zwischen den Benutzern. Benutzer bilden Gruppen (*Communities*) um so über die neuesten Aktivitäten anderer informiert zu bleiben. Um Fotos eines Users zu präsentieren verwendet *Flickr* einen *Fotostream*, welcher ähnlich einem Webblog aufgebaut ist. Jedes veröffentlichte Bild stellt dabei einen (Webblog)-Eintrag dar und wird in chronologischer Reihenfolge ausgegeben. Um Bilder systematisch zu organisieren, bietet *Flickr* neben dem *Fotostream*, auch Alben (*Set*) und Sammlungen (*Collections*) an. Wobei sich ein *Set* an traditionelle Fotoalben anlehnt und *Collections* eine Gruppierung zusammenhängender *Sets* darstellen.

Eine weitere Form der Organisation, welche auch sehr wichtig für den Suchvorgang ist, stellt die Indizierung der Annotationen dar. Neben Informationen aus dem Bildtitel und der Kurzbeschreibung werden auch *Notes* indexiert. Mittels *Notes* ist es möglich auf einem Bild bestimmte Details zu markieren. Um spezielle Objekte auf einem Bild zu finden, kann dann nach diesen *Notes* gesucht werden. In Abbildung 2.5 ist ein Bild mit einer *Notes*-Markierung zu sehen.

Die wichtigste Form der Organisation stellen jedoch die Schlagwörter dar, welche sowohl bei der Kategorisierung als auch bei der Suche einen sehr hohen Stellenwert darstellen. Betrachter können ein Bild mit maximal 75 Tags annotieren, sofern der Bildbesitzer *Fremd Tagging* erlaubt. Auf Flickr werden neben *Raw-Tags* auch *Machine-Tags* verwendet.

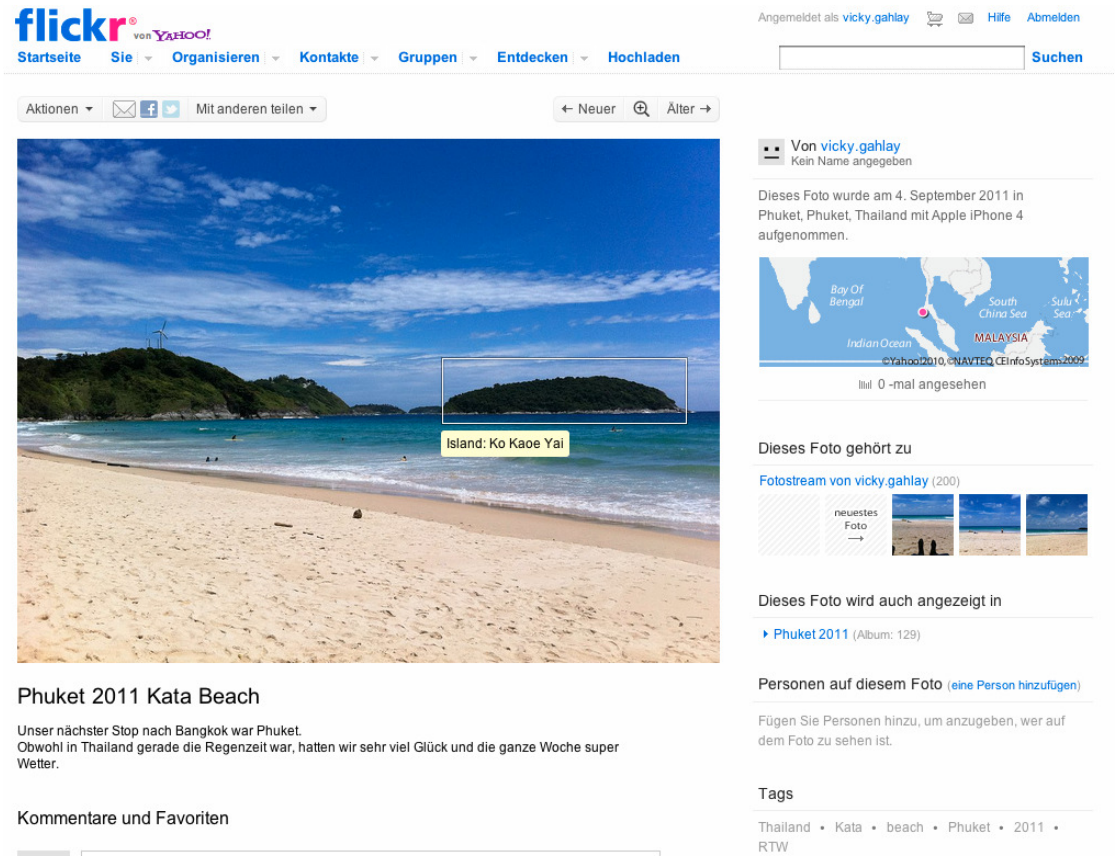


Abbildung 2.5: Screenshot von *http://www.flickr.com*

2.1.7.4 Viddler und YouTube

Viddler und *YouTube* ermöglicht seinen Benutzern, Videos in hoher Qualität zu veröffentlichen (Ding et al., 2011). Wie auf den meisten Videoportalen erhält man auf der Startseite eine Übersicht über die interessantesten Videos (Neue, meist betrachtete, meisten Kommentare etc.). Wie bei *Flickr* legen beide Plattformen sehr großen Wert auf Social Networking. Neben der Mög-

lichkeit Gruppen zu gründen, können auch private Nachrichten zwischen Benutzern versendet werden. Jeder User und jede Gruppe erhält auch eine eigene Profilseite, wo neueste Videos chronologisch aufgelistet sind. Auf beiden Portalen ist es möglich Videos mit Tags zu versehen. Auf *Viddler* können sogar die zeitlichen Momente eines Videos mit eigenen Tags und Kommentaren (*Time Tags*, *Time Comments*) versehen werden, welche dann beim Abspielen sichtbar werden. YouTube unterstützt nur sogenannte *Global Tags* und *Global Comments*, welche sich auf das ganze Video beziehen (Müller et al., 2010).

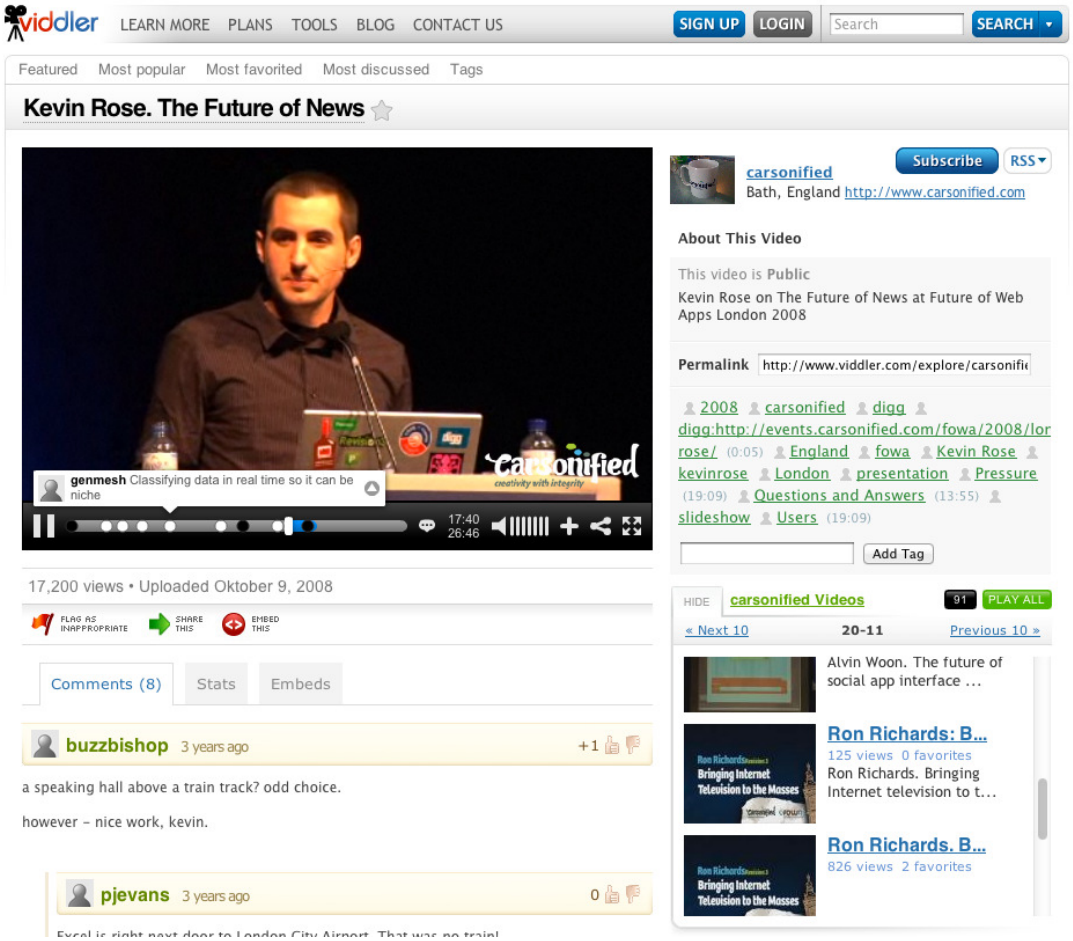


Abbildung 2.6: Screenshot von einem Video auf *http://www.viddler.com* mit *Time Tags* und *Global Tags*

2.1.8 Ausprägung von Tagging-Systemen

Die in Kapitel 2.1.7 erwähnten Systeme basieren zwar alle auf Social Tagging, aber unterscheiden sich bei der Implementierung einzelner Funktionen bzw. Regeln. Diese Regeln beschreiben die Interaktion der Benutzer, Ressourcen und Tags und sind maßgeblich für den Erfolg eines Systems verantwortlich. In (Marlow et al., 2006) wurden zahlreiche Tagging-Dienste analysiert und können anhand von 7 Dimensionen kategorisiert werden:

- **Tagging Rechte / Tagging Rights**
  Einer der wichtigsten Aspekte eines Tagging-Systems sind die Benutzerrestriktionen. Neben *Self-Tagging*, welches als extreme Form nur dem Benutzer der die Ressource online gestellt hat, das Taggen erlaubt (z.B. Technorati), gibt es auch noch *Free-for-All*-Tagging. Letzteres ermöglicht allen Benutzern das Taggen (z.B. Del.icio.us oder Yahoo! Podcasts) von Ressourcen. Zwischen diesen beiden Extremen existieren einige abgeschwächte Varianten, wo nur bestimmte Benutzergruppen wie Freunde, Familie (Flickr private Alben) die Ressourcen mit Schlagwörtern versehen dürfen.

  Diese unterschiedlichen Restriktionen spielen auch bei der Indexierung von Tags eine wichtige Rolle. Es besteht ein sehr großer Unterschied von welcher Personengruppe ein Tag zu einem Foto hinzugefügt wurde. Tags von Fotografen, Freunden oder Familie stellen möglicherweise das Bild in einem ganz anderen Zusammenhang dar, wie Tags von einem fremden User.

  Neben den Restriktionen wer Schlagwörter hinzufügen darf, gehört zudem auch das Löschen von Tags zur Tagging-Rights Kategorie. Hier gibt es von System zu System sehr große Unterschiede. Fotosharing Plattformen wie Flickr z.B. erlauben das Entfernen der Tags nur durch den Uploader. Bei anderen Social Tagging-Systemen hingegen, darf nur der User der den Tag erstellt hat, diesen auch wieder löschen (Last.fm).

- **Tagging Unterstützung / Tagging Support**
  Bei diesem Begriff geht es um die Beeinflussung bzw. Unterstützung des Benutzers bei der Vergabe von Schlagwörtern. Hier unterscheidet man zwischen *blind*, *viewable* und *suggestiv* Tagging. Beim „Blind Tagging“ darf der Benutzer, während des Taggingvorganges nur seine eigenen Schlagwörter sehen (Del.icio.us), wohingegen beim *Viewable Tagging* alle Tags ersichtlich sind (e.g.Yahoo Podcasts or Flickr). Bei der letzten Variante „Suggestiv Tagging“ werden dem Benutzer durch ein Recommender-System Tags vorgeschlagen. Diese Vorschläge werden auf verschiedenste Arten generiert. Neben bereits vorhandenem Tags und deren Synonymen werden auch andere kontextbezogene Informationen verwendet. Im Gegensatz zu den anderen Arten ist *Suggestiv Tagging* beim Indexierungsvorgang wesentlich schneller. Hier kommt auch die *Social Proof* Hypothese von Cialdini zur Anwendung ((Cialdini and Truus, 2005)): Menschen tendieren dazu, die Handlungen und Handlungsweisen von Personen in ihrer direkten Umgebung nachzuahmen. Aus diesem Grund bildet das „Suggestiv Tagging“ nach (Marlow et al., 2006) konvergente Folksonomien, da die Benutzer bereits verwendete Tags wiederverwenden.

- **Aggregation**
  Auch bei der Aggregation von Tags gibt es zwei verschiedene Varianten. Hier differenziert

man zwischen „Bag-Model“ und „Set-Model“-Aggregation. Der Unterschied zwischen diesen beiden Verfahren liegt bei der Handhabung von Mehrfachnennung von Schlagwörtern. Bei der „Bag-Model“-Aggregation ist die Mehrfachnennung von Tags erlaubt und fließt in die Gewichtung mit ein. Sehr häufige Schlagwörter werden als erstes und nicht so häufige als letztes aufgelistet. Dieser Ansatz kommt beispielsweise bei der Social-Bookmarking-Plattform *Del.icio.us* zum Einsatz.

Bei der „Set-Model“-Aggregation ist die Mehrfachnennung von Schlagwörtern nicht erlaubt. Alle Tags werden gleich gewichtet und es können nur Tags zu einer Ressource hinzugefügt werden, welche noch nicht vergeben wurden. Diese Variante wird beispielsweise von *Flickr* verwendet.

- **Art der Ressource / Type of Object**
  Bei dieser Dimension geht es um den Typ der Ressource. Im Internet werden folgende Ressourcen sehr häufig verwendet: Webseiten (*Del.icio.us*, Bilder (*Flickr*), Video (*YouTube*), Musik (*grooveshark*, *Last.fm*) oder Dokumente (*CiteULike*). (Marlow et al., 2006) vermutet das sich die Tags von verschiedenen Ressourcen sehr stark unterscheiden.

- **Herkunft der Ressource / Source of Material**
  Dieser Punkt behandelt die Herkunft der Ressourcen. Bei einigen Systemen wie *Flickr* oder *YouTube* ladet der User selbst das Material hoch, wohingegen bei anderen Plattformen wie *Last.fm* das System die Ressource zur Verfügung stellt.

- **Ressourcen Konnektivität / Ressource Connectivity**
  Ressourcen können auch ohne Tags zueinander in Relation stehen. Diese Beziehung bezeichnet man als Ressourcen Konnektivität, wobei hier auch zwischen *linked*, *grouped* und *none* Konnektivität unterschieden wird. Beispielsweise können Webseiten durch direkte Links miteinander verbunden werden (*linked*). Bei *Flickr* ist es möglich Bilder nach Themengebieten in Gruppen zusammenzufassen (*group* connectivity). Kann keine Relation zwischen den Ressourcen hergestellt werden so bezeichnet man dies als *none connectivity*.

- **Soziale Konnektivität / Social Connectivity**
  Ähnlich dem letzten Punkt können auch Benutzer zueinander in einer Relation stehen. Durch die Verwendung von Social Networking bilden die User Netzgemeinschaften wie Gruppen oder Freundschaftslisten. Auch hier unterscheidet man zwischen *linked*, *grouped* und *none*.

Dimension	Main categories	Summary of potential implications
Tagging Rights	self-tagging, permission-based, free-for-all	Nature and type of resultant tags; role of tags in system
Tagging Support	blind, suggestive, viewable	Convergence on folksonomy or overweighting of tags
Aggregation model	bag, set	Availability of aggregated statistics
Object type	textual, non-textual	Nature and type of resultant tags
Source of material	user-contributed, system, global	Different incentives, nature and type of resultant tags
Resource connectivity	links, groups, none	Convergence of similar tags for linked resources
Social connectivity	links, groups, none	Convergence on localize folksonomy

Abbildung 2.7: Dimensionen des Designs von Taggingsystemen (Marlow et al., 2006)

Tabelle 2.7 zeigt eine Zusammenfassung der sieben Dimensionen des Designs von Taggingsystemen nach (Marlow et al., 2006).

2.1.9 Folksonomien

Dieser Begriff wurde 2004 von Gene Smith geprägt und setzt sich aus den Begriffen *Folk* und *Taxonomy* zusammen. Thomas Vander Wal definiert den Begriff Folksonomie auf folgende Weise (Wal, 2007):

> Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrival. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information. The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object as well. The people are not so much categorizing as providing a means to connect items and to provide their meaning in their own understanding.

Eine Folksonomie ist also das Ergebnis der persönlichen freien Verschlagwortung (*Free Tagging*) von Informationen und Objekten und entsteht aus dem Akt des Taggens. Wobei mit *Free Tagging* die freie Zuordnung von Begriffen zu einer Ressource im Internet gemeint ist (Golub et al., 2009). Durch diese Zuordnung bildet sich somit eine Art Ordnung bzw. eine Struktur, welche aber keineswegs mit Ontologie oder Taxonomie verglichen werden kann. Eine Folksonomie unterscheidet sich durch folgende Punkte von Taxonomie bzw. Ontologien (Albrecht, 2006):

- Tagging erfolgt durch freies nicht kontrolliertes Vokabular der Benutzer

- Es herrscht keine hierarchische Struktur (keine Eltern-Kind Beziehung), da keines der Tags einem anderen übergeordnet ist. Einzige Relation zwischen den Tags sind der Benutzer und das getaggte Objekt selbst.

2.1.9.1 Arten von Folksonomien

Nach (Vander Wal, 2005) gibt es zwei Arten von Folksonomien - breite (*broad*) und enge bzw. schmale (*narrow*) Folksonomie. Bei einer breiten Folksonomie (Siehe Abbbildung 2.8a) annotieren sehr viele Benutzer die selbe Ressource mit Schlagwörtern. Ein User (*Content Creator*) erstellt die Ressource und veröffentlicht sie im Internet. Danach ordnen verschiedene Benutzer (Gruppe A-F) dem Objekt ein Schlagwort (Tag 1-5) hinzu und verwenden diese Tags auch für die Organisation. Bei der *Broad*-Folksonomie ist das mehrfach Taggen der selben Ressource erlaubt.

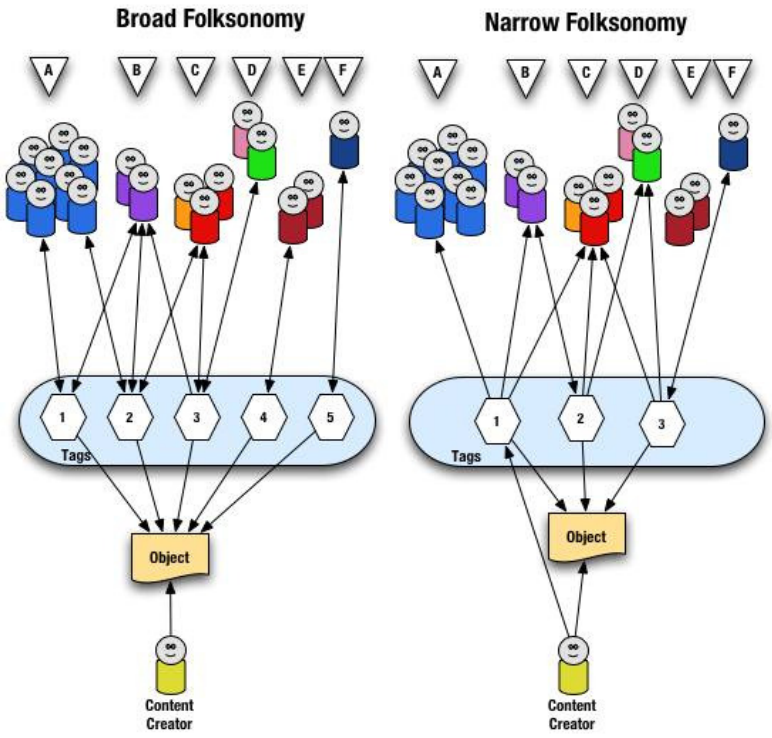
In einer *Narrow*-Folksonomie darf nur eine kleine Anzahl an Personen Objekte mit Tags versehen (Siehe Abbildung 2.8b). In diesem Fall werden bereits durch den *Content Creator* einige Schlagwörter vorgegeben. Nur wenige Benutzer (Gruppe B und F) annotieren diese Objekte mit weiteren Begriffen. In engen Folksonomien sind Tags direkt mit der Ressource verknüpft und sind sehr gut auf Objekte anwendbar, welche nicht durch einfache Textsuche gefunden werden können (z. B. Fotos, Musik).

2.1.9.2 Vor- und Nachteile von Folksonomien

Neben den ökonomischen Vorteilen von Folksonomien und Social Tagging (schnelle und intuitive Anwendbarkeit, geringe Kosten), können weitere Vor- und Nachteile von Folksonomien aufgeführt werden.

Nach (Ebersbach et al., 2008) ist ein wichtiger Vorteil das dynamische Vokabular, welches sich mit der aktuellen Sprachgewohnheit der Benutzer mit verändert. Auch die leichte Skalierbarkeit, ohne zusätzlichen Aufwand, stellt im Gegensatz zu Ontologien einen erheblichen Mehrwert dar. Der Zufall hat in Folksonomien einen hohen Stellenwert, der Benutzer kann ausgehend von Tags und Objekten zu anderen Benutzern Verbindungen aufbauen und durch diese Verbindung auch andere Ressourcen entdecken. Neben diesen Vorteilen erleichtert die Anpassungsfähigkeit, Veränderbarkeit und Dynamik die Entwicklung und Pflege von bestehenden Systemen. Auch das Suchen in Social Tagging-Systemen ist viel einfacher und intuitiver als die Recherche mit ausgeklügelten Instrumenten der Informationssuche. In (Peters and Stock, 2008) werden die Vorteile von Folksonomien folgendermaßen zusammengefasst:

- spiegeln die Sprache der Nutzer authentisch wider
- erlauben verschiedene Interpretationen
- ermöglichen eine günstige Form der Inhaltserschließung
- sind Term-quellen für die Entwicklung und Pflege von Ontologien und kontrolliertem Vokabular
- geben die Qualitätskontrolle an die Nutzer weiter



(a) Broad-Folksonomie **(b)** Narrow-Folksonomie

Abbildung 2.8: Arten von Folksonomien (Vander Wal, 2005)

- sind die einzige Möglichkeit, Masseninformation im Web zu erschließen
- erlauben konkretes Suchen und Browsing, berücksichtigen Neologismen
- tragen dazu bei, Communities zu identifizieren, geben eine Basis für Empfehlungssysteme
- sensibilisieren Nutzer

Als Nachteil wird in (Peters and Stock, 2008) festgestellt, dass gerade das Fehlen eines kontrollierten Vokabulars durch die Gleichrangigkeit aller Begriffe große Schwierigkeiten mit sich bringt, wie schon durch das Ausbleiben eines Synonymabgleichs deutlich wird. Auch die Nachteile von Folksonomien werden in (Peters and Stock, 2008) zusammengefasst :

- verschiedene Levels der Indexierung, Vermischung von Sprachen, versteckte paradigmatische Relationen bleiben ungenutzt
- fehlende Trennung von formalen bzw. bibliographischen Tags, Spam-Tags, nutzer-spezifische Tags und andere uneindeutige Schlagworte sowie gerade bei der Verschlagwortung von nicht-textuellen Inhalten die Durchmischung von Bildbeschreibung, Herkunftsangaben und Zuschreibungen.

2.2 Folksonomien vs Ontologien

Folksonomien weisen meistens keine Struktur auf, sie sind schwach bis unstrukturiert. Diese fehlende Semantik erschwert den Such- und (Wieder-)Findungsprozess von Ressourcen. Im Gegensatz zu Ontologien unterscheiden sich Folksonomien durch folgende Faktoren (Golder and Huberman, 2005; Guy and Tonkin, 2006):

- **Schreibweise**
  Schlagwörter werden entweder falsch (Rechtschreibung) oder auf unterschiedliche Art und Weise geschrieben. Beispielsweise: Abkürzungen WWI (World War I), Verwendung des Plural „Noodle(s)“

- **Mehrsprachig**
  Verschiedene Benutzer verwenden oft unterschiedliche Sprachen um dieselben Objekte zu beschreiben z. B. Meer, Ocean, Mar (Spanisch).

- **Polysemie**
  Die Polysemie bezeichnet in den Sprachwissenschaften, ein Wort, welches für verschiedene Begriffe steht (Mehrdeutigkeit). Dies kann dazu führen, dass der Benutzer bei der Suche mit ungenauen Ergebnissen konfrontiert wird z. B. Bank, Kreditinstitut oder Sitzgelegenheit.

- **Homonymie**
  Ähnlich der Polysemie, jedoch hat hier ein Begriff mehrere verschiedene Bedeutungen. Beispielsweise bezeichnet der Begriff *Hund* ein Tier, wird er jedoch auf einem Bild, wo ein Mensch zu sehen ist, versehen, so wird das als „Schimpfwort“ interpretiert. Dies kann genauso zu falschen Suchergebnissen führen.

- **Synonymie**
  Bezeichnet die Gleichheit der Bedeutung von verschiedenen Wörtern. Zum Bespiel haben die Tags *Zehn*, *10*, *X*, *ten* die gleiche Bedeutung. Bei einigen Social-Tagging Systemen werden Ressourcen mit sehr vielen synonymen Tags versehen um eine effektive Suche zu ermöglichen.

- **Unterschiedliche Abstraktionsebenen**
  Nach (Tanaka and Taylor, 1991) werden aufgrund unterschiedlicher Intentionen Tags auf anderen Abstraktionsebenen definiert z. B. ist eine mit dem Tag *Sitzbank* annotierte Ressource nicht mit dem Begriff *Bank* zu verwechseln

Abhängig vom Einsatzgebiet, können alle diese Faktoren gleichzeitig als Vorteile bzw. Nachteile von Folksonomien gesehen werden. Beispielsweise kann der Benutzer durch die Verwendung der Mehrsprachigkeit ein Objekt genauer indexieren, aber genau diese unterschiedliche Schreibweise kann auf der anderen Seite auch den Wiederfindungsprozess erschweren.

Mit Ontologien ist es möglich Relationen zwischen einzelnen Elementen auf unterschiedliche Abstraktionsebenen abzubilden. Auch unterstützen Ontologien *Synonome* und trennen Homonyme und mehrsprachige Begriffe. Ontologien sind aber, sobald es um die Skalierung geht sehr

schwer zu handhaben. Es existieren zwar Werkzeuge zur Verwaltung von Ontologien sind aber für die meisten Nutzer viel zu kompliziert.

Taxonomien sind für die Wissenschaft von sehr großer Bedeutung, da sie den Umgang mit neuen Objekten erleichtern und Aussagen über bereits klassifizierte Objekte möglich machen. Naturwissenschaftliche Disziplinen verwenden Taxonomien z. B. für die Klassifikation von Lebewesen. (Golder and Huberman, 2005) demonstrieren dieses Konzept anhand von Katzenarten:

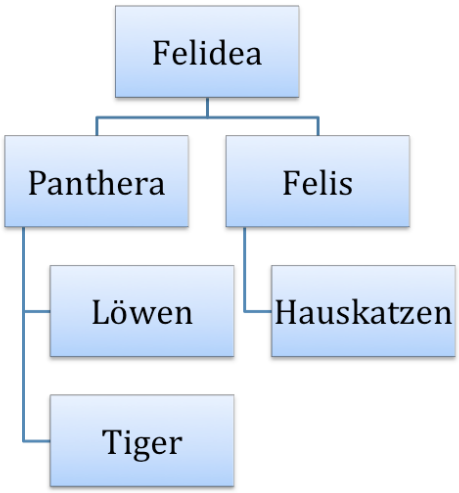
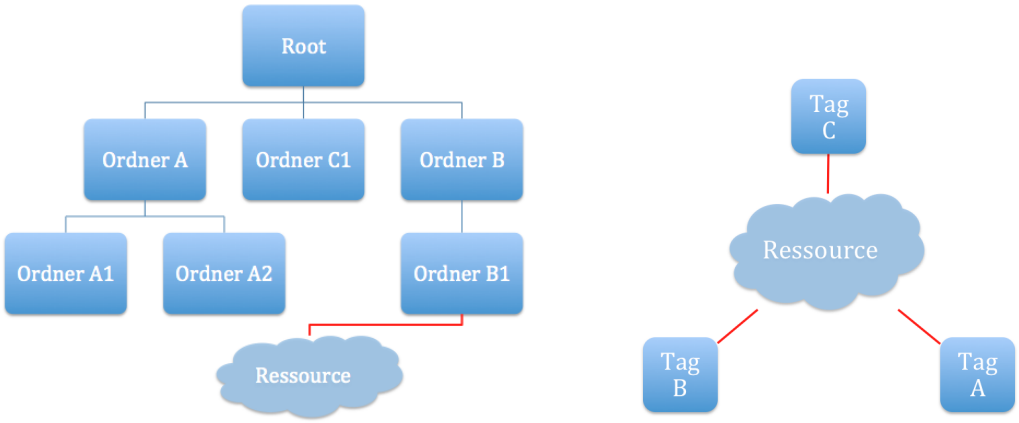


Abbildung 2.9: Hierarchische Darstellung von Katzenarten

In der Biologie haben sich eigene Begriffe (z. B. *Art*, *Gattung*, *Familie*) entwickelt um den Rang einer Systematik besser zu beschreiben. *Löwen* und *Tiger* sind Arten und gehören zur Gattung *Panthera*. *Hauskatze* ist auch eine Art und gehört zur Gattung *Felis*. Wobei *Panthera* und *Felis* wiederum unter der Familie *Felidae* zu finden sind. Diese Art von Zuordnung wird auch als hierarchisch und exklusiv bezeichnet. Genauer gesagt, ist jedes Element einer bestimmten Kategorie zugeordnet, wobei diese Kategorien wiederum einer übergeordneten, allgemeineren Kategorie angehören. Social Tagging ist im Gegensatz dazu weder exklusiv noch hierarchisch, was sich in vielen Fällen als Vorteil erweisen kann. Beispielsweise möchte ein Biologe/eine Biologin einen Artikel über Katzenrassen in seinem hierarchischem Ordnungssystem einbringen. So gibt es für ihn mehrere Auswahlmöglichkeiten (Abbildung 2.10a).



(a) Hierarchisches Ordnungssystem **(b)** Nicht-Hierarchisches Ordnungssystems

Abbildung 2.10: Hierarchisches vs. Nicht-Hierarchisches Ordnungssystem

In einem klassischen Ordnungssystem existieren die Kategorien bereits vor der Einordnung einer Ressource. Hier muss ausgehend vom Wurzelknoten (*Root*) der Artikel in den entsprechenden Ordner gelegt werden. Es ist nicht möglich einer Ressource mehrere Kategorien zu zuweisen (außer der Benutzer kopiert das Dokument in mehrere Ordner). Dies bringt auch den Nachteil der hierarchischen gegenüber der nicht-hierarchischen Gliederung mit sich. Beim Social Tagging hat der Benutzer die Möglichkeit sein Dokument mit mehreren Tags zu kategorisieren (Abbildung 2.10b). Diese Tags müssen weder alphabetisch noch nach einer bestimmten Struktur vergeben werden.

Trotz dieser Einschränkung sind hierarchische Systeme bei der Organisation deutlich effizienter, da die Objekte einen fixen Platz zugewiesen bekommen. Jedoch kann beim nicht-hierarchischen effektiver nach Elementen mit mehreren Kategorien gesucht werden. Letzteres wird nicht mehr als Suchvorgang sondern als *Filtern* bezeichnet, da aus allen vorhandenen Tags die zutreffenden herausgefiltert werden.

2.2.1 Power Law in Tagging-Systemen

Das Power Law oder Potenzgesetz ist eine Verteilung, die dadurch charakterisiert ist, dass wenige Elemente in einer hohen Frequenz auftreten und die meisten Elemente nur selten. Diese Verteilung tritt in sehr vielen Bereichen auf. Beispielsweise fand Vilfredo Pareto, ein italienischer Ökonom des 19 Jahrhunderts, heraus, dass 80 % des Vermögens auf 20 % der Bevölkerung verteilt ist. Diese Art der Verteilung bezeichnet man heute auch als *Paretoprinzip* (Wikipedia, 2011*f*).

Das Power Law gilt für viele Bereiche von Social Tagging wie die Anzahl der Tags pro Benutzer, die Nutzungshäufigkeit der Schlagwörter eines Users oder auch die Aktivität einer Benutzergruppe. Letzteres wurde von (Shirky, 2003) näher untersucht. In sozialen Netzwerken, bilden sich automatisch Elite-Gruppen, welche für die Mehrheit des Datenverkehrs verantwortlich sind. Am Beginn gibt es diese Gruppenbildung nicht, aber sobald das System wächst, können sich nicht mehr alle Benutzer an allen Konversationen beteiligen. Manche Gruppen werden

immer populärer und bilden schließlich die Elite. Die Verteilung nähert sich dann einer Potenzverteilung (Power-Law) an.

In Abbildung 2.11 ist die Anzahl an Tags pro URL dargestellt. Es ist erkennbar, dass wenige URLs mit sehr vielen Tags versehen sind. Um genau zu sein besitzen die ersten vier URLs 80% der Tags, danach nähert sich die Kurve gegen Null.

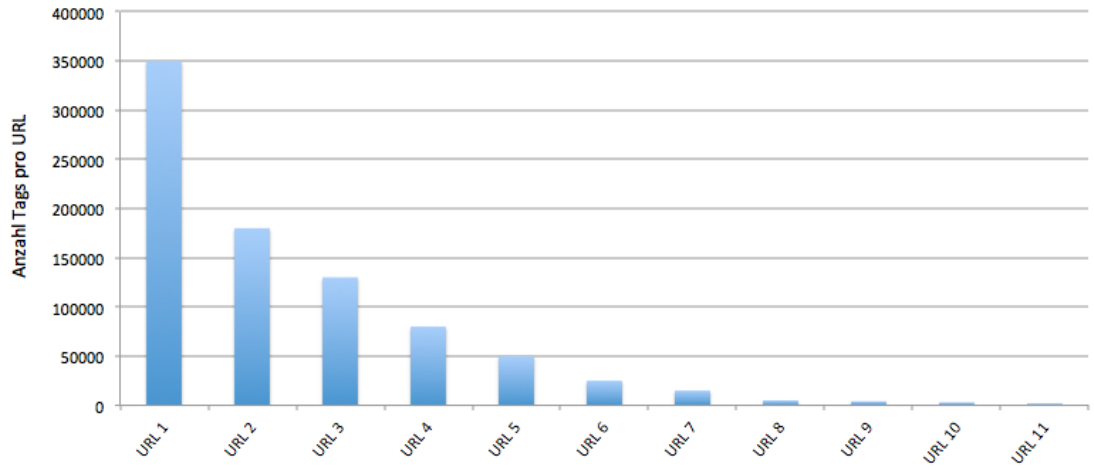


Abbildung 2.11: Verteilung Tags pro URL *Del.icio.us* (Vander Wal, 2005)

2.3 Visualisierung von Folksonomien

Im Web finden sich zahlreiche Applikationen mit denen sich Folksonomien visualisieren lassen. Neben Schlagwörtern und deren Beziehungen werden auch sehr häufig die sozialen Kontakte im Detail betrachtet.

2.3.1 Tagcloud

Die bekannteste und am weitesten verbreitete Visualisierung einer Folksonomie sind *Tag-* bzw. *Wordclouds*. Sie stellen meistens alphabetisch sortiert die beliebtesten Tags zu einer Ressource dar (Abbildung 2.12). Die Häufigkeit der Tags wird durch die Größe der Darstellung des Wortes visualisiert. Diese Visualisierung ist sehr einfach gehalten, wird aber in den meisten Social-Tagging-Systemen verwendet um einen Überblick über die meist verwendeten Tags zu geben (Voss, 2006).

Bevor eine *Tagcloud* gebildet werden kann, ist es wichtig Wörter auf ihre Grundform zu reduzieren (*Stemming*) z. B. Wikis auf Wiki etc. Danach kann die Schriftgröße einer *Tagcloud* berechnet werden, welche bei kleineren Dokumenten direkt proportional zur Häufigkeit der Wörter ist. Für größere Dokumente und Häufigkeiten muss eine Normierung nach folgender Formel vorgenommen werden (Wikipedia, 2011*h*):

Definition 4: $s_i = \left\lceil f_{\max} \cdot \frac{t_i - t_{\min}}{t_{\max} - t_{\min}} \right\rceil$ für $t_i > t_{\min}$; sonst $s_i = 1$

- s_i: anzuzeigende Schriftgröße
- $f_{\max}$: maximale Schriftgröße
- t_i: Häufigkeit des betreffenden Schlagwortes
- $t_{\min}$: Häufigkeit, ab der ein Schlagwort angezeigt werden soll
- $t_{\max}$: Häufigkeit des häufigsten Schlagwortes

Abbildung 2.12: *Word Cloud* dieser Arbeit

2.3.2 Mindmap

Mittels einer Mindmap wird ein Themengebiet und deren Beziehung zu anderen Themen visualisiert. Mindmaps sind eigentlich beschriftete Baumdiagramme, welche durch zusätzliche Informationen Element näher erläutern (Wikipedia, 2011*e*). Diese Art der Visualisierung setzt bereits am Beginn eine vernetzte Struktur ein und eignet sich dadurch sehr gut, um Abhängigkeiten zwischen unterschiedlichen Elementen darzustellen. Ausgehend von einem zentralen Thema (Mitte der Mindmap), verlaufen Hauptäste zu anderen Themen, welche wiederum auf Unterkapitel verzweigen. Das Einsatzgebiet von Mindmaps ist sehr groß. Neben der Planung von wirtschaftlichen Anwendungen, dienen sie auch oft zur Darstellung kognitiver Vorgänge. Weiters werden sie auch sehr häufig bei der Sammlung von Ideen (Brainstorming), beim Lernen bzw. beider Prüfungsvorbereitung sowie Protokollierung von Besprechungen verwendet (Dalamagas et al., 2010). Bei Folksonomien werden Mindmaps zur Darstellung der Tag-Relationen untereinander verwendet.

Die nachfolgende Abbildung zeigt eine Mindmap von Wikipedia, welche mit WikiMindMap[9] erstellt wurde:

[9] `http://www.wikimindmap.com/` [Zugriff am 30.11.2011]

Abbildung 2.13: Mindmap von Wikipedia zum Begriff *Social Tagging*.

2.3.3 Tag Soup

Eine *Tag Soup* in dieser Form eine *Delicious Soup*[10] ist ein interaktives Tool und zeigt welche Tags sehr häufig in Kombination mit anderen Tags auftreten. Diese Beziehung wird dann in Tag-Gruppen zusammengefasst, welche als Kugeln dargestellt werden. Die Nähe zwischen den Kugeln ist ein Indikator dafür, dass diese Tags sehr häufig gemeinsam vorkommen. Umso näher die Kugeln zueinander positioniert sind, umso häufiger kommen diese Schlagwörter im selben Kontext vor. Dahingegen spiegelt die Größe der Elemente die Gesamthäufigkeit des Schlagwortes wider.

In Abbildung 2.14 ist die *Personomie* Linksammlung eines *Del.icou.us*-User abgebildet. In dieser Darstellung wurde der Tag (Kugel) *Cool* als Hauptkugel ausgewählt und somit beziehen sich alle Berechnungen auf diese Kugel. Dieser Tag wurde von diesem User 155 mal vergeben. Die Tags (Kugeln) *Cool, Design, Art* werden sehr nah aneinander dargestellt, da sie sehr häufig in Kombination miteinander vergeben wurden. Auch wird aus dieser Abbildung sofort ersichtlich, welcher Tag im Gesamten am häufigsten vergeben wurden. In diesem Fall ist die Kugel mit der Beschriftung *Design* am größten und somit wurde dieser Tag am häufigsten vergeben.

[10] `http://www.zitvogel.com/deliciousoup/` [Zugriff am 1.12.2011]

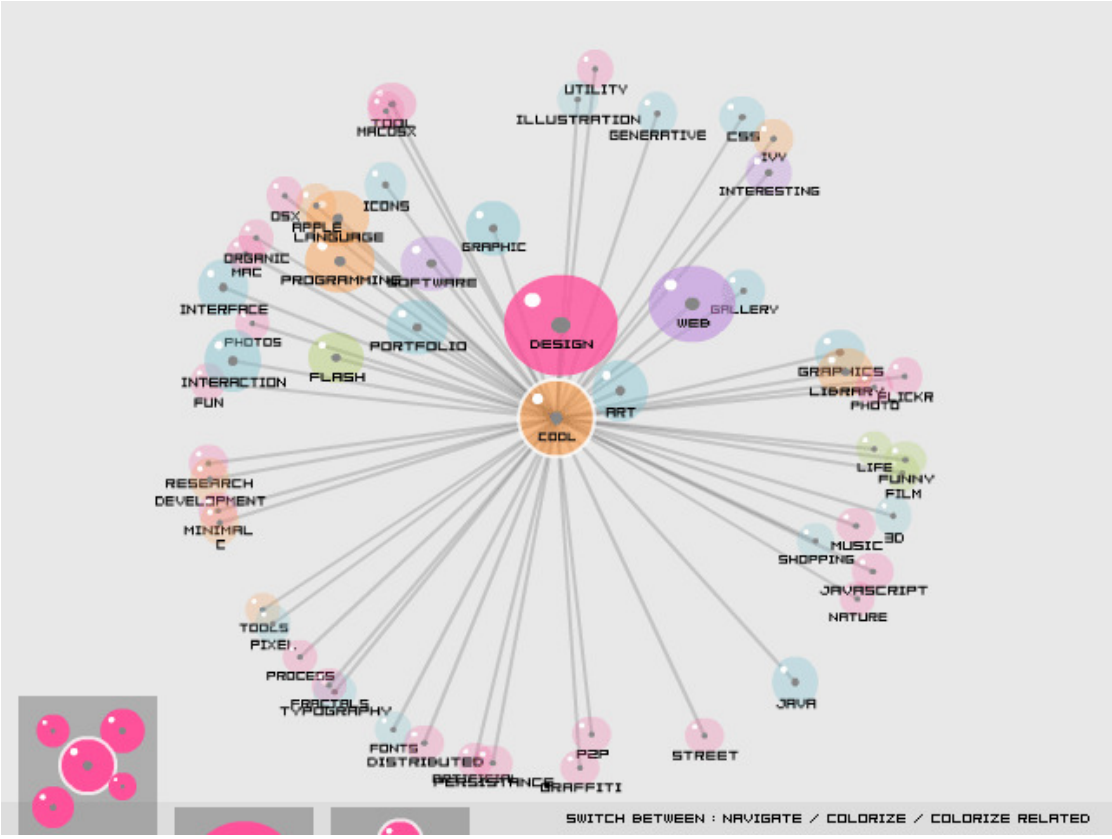


Abbildung 2.14: Visualisierung der *Personomie* (Linksammlung) eines *Del.icou.us*-User mittels *Delicious Soup*

2.3.4 Darstellungen mittels Graphen

Ein Graph eignet sich sehr gut um die Relationen zwischen Personen (*Social Graph*) bzw. Ressourcen (*Touch Graph*) zueinander darzustellen. In den folgenden Unterkapiteln werden die zwei am weitest verbreiteten Ausführungen der graphentheoretischen Darstellung von Folksonomien näher erläutert.

2.3.4.1 Touch Graph

Bei dieser Darstellung werden die Beziehungen zwischen den Tags als Graphen dargestellt. In Abbildung 2.15 werden die Folksonomien zu Social Software und Web 2.0 von *Del.icio.us* verwendet und deren Relation mittels *Touch Graph* visualisiert.

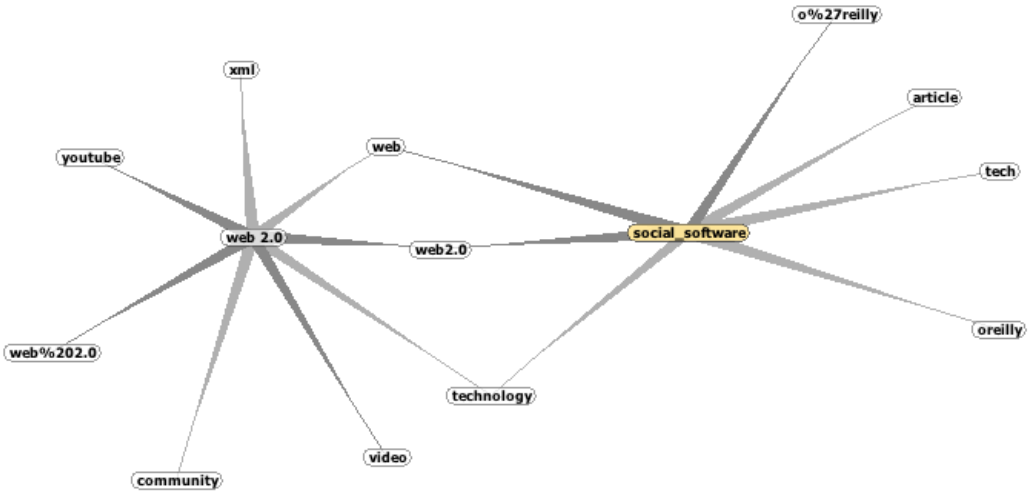


Abbildung 2.15: Visualisierung der Folksonomie der Tags *Web 2.0* und *Social Software* mittels *Touch Graph*

Aus diesem Graphen ist sofort ersichtlich, welche Tags gemeinsam sehr häufig vergeben wurden. Beispielsweise wurde der Tag *Web 2.0* sehr oft in Kombination mit den Tags *YouTube*, *Communitiy* etc. vergeben. Auch der Begriff Social Software wurde mit den *Refinning Tags Orelly*, *Articel* verfeinert. Das Bindeglied zwischen den beiden Folksonomien sind die Schlagwörter *Web*, *Web 2.0* und *Technologie*. Der Tag *Web 2.0* ist in dieser Abbildung zwei mal vorhanden, da er sowohl ein Bindeglied als auch eine Folksonomie darstellt.

2.3.4.2 Social Graph

In dieser Darstellung werden die Beziehungen zwischen den Usern näher beleuchtet. Diese Art der Darstellung wurde durch *Facebook* sehr populär und stellt die Verbindungen zwischen den einzelnen User aus graphentheoretischer Sicht dar.

In Abbildung 2.16 ist der *Social Graph* meines Facebook Benutzers dargestellt. Dieser Graph wurde mit der *Social Graph*[11]-Applikation auf Facebook erstellt und stellt die Verbindung zwischen allen Freunden aus der Freundschaftsliste dar. Der in Abbildung 2.16 abgebildete Graph ist interaktiv, d.h. sobald der Mauszeiger über den Namen eines Users stehen bleibt, werden alle gemeinsamen Freunde hervorgehoben und in einer anderen Farbe visualisiert.

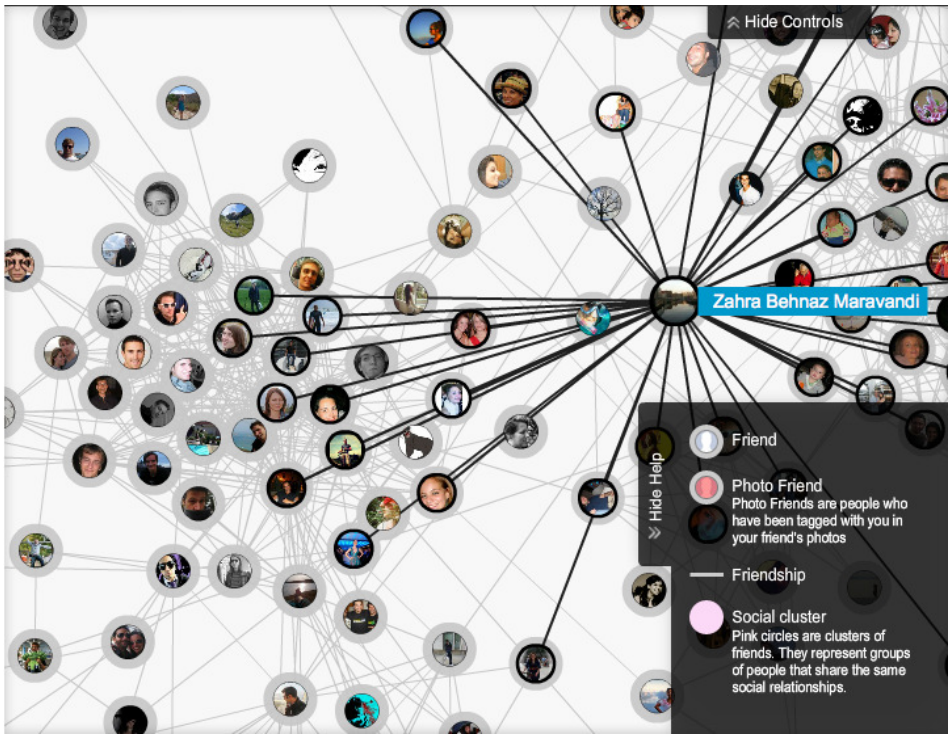


Abbildung 2.16: Social Graph eines Facebook Users

[11] `https://apps.facebook.com/socgraph/` [Zugriff am 19.01.2011]

KAPITEL 3

Flickr: The Commons

3.1 Flickr: The Commons

Viele Kulturinstitutionen werden heute mit einer Reihe von neuen Herausforderungen konfrontiert. Ein immer wichtig werdender Aspekt ist, historische Artefakte für so viele Menschen wie möglich zugänglich zu machen. Gleichzeitig steht man aber aufgrund begrenzter räumlicher Ressourcen vor einem Konflikt, der sich nur durch die Verwendung des World Wide Webs lösen lässt.

Flickr startete am 16.Jänner 2008 das Pilotprojekt „The Commons“ in Zusammenarbeit mit der *Library of Congress*[1] (LOC). Dieses Projekt entstand konkret aus der Fragestellung, in wieweit Social-Software zur Wissensbereicherung beitragen kann und verfolgte folgende Ziele:

- Erhöhung der Bekanntheit von historischen Archiven der LOC, nicht nur für Menschen die die LOC Webseite besuchen, sondern auch für Menschen die sich gerne für Kunst und Fotografie interessieren.
- Ein besseres Verständnis schaffen, wie sowohl die Institution als auch die User von Social Tagging und User Generated Input profitieren können.
- Erfahrung durch die Teilnahme an Web-Communities sammeln und erkennen welche Artefakte in den Beständen der LOC für User interessant sind.

Am Beginn wurde ein Teil der Bildersammlung der *Library of Congress* auf *Flickr* veröffentlich. Die eigene Webseite kam aufgrund technischer Hürden nicht in Frage, aus diesem Grund entschied man sich für Flickr als zentralen Partner. Da die kulturellen Institutionen aber selten die Urheber der Artefakte sind, die sie bereitstellen, musste hier am Beginn der Kooperation eine eigene Rechtserklärung geschaffen werden.

[1] `http://www.loc.gov/index.html`

Zwar können Institutionen Anderen die Fotos zur Nutzung ohne Einschränkung überlassen, aber nur dann wenn der Urheber des Bildes damit einverstanden ist. Bei historischem Bildmaterial sind meistens Informationen wie Erstellerdatum und Name des Urhebers nicht vorhanden. Weiters variiert der Schutz und die Dauer dieser Gesetzgebung von Land zu Land. Letztendlich einigten sich alle teilnehmenden Einrichtungen ihre Archive auf Basis der Rechtserklärung „Keine Urheberrechtsbeschränkung bekannt“[2] zu veröffentlichen und diese auch mit dieser Lizenz verpflichtend zu kennzeichnen (Springer et al., 2008).

Neben einigen Farbbildern der ersten Wirtschaftskrise und des zweiten Weltkrieges veröffentlichte die LOC auch noch einige Schwarzweißbilder eines Pressebüros aus dem zwanzigsten Jahrhundert. Diese beiden Alben erzielten eine sehr große Resonanz, wobei alleine in den ersten 24 Stunden die Flickr-Seite der LOC eine Million Aufrufe verzeichnete. Seit dem Start der Initiative erhöhten sich die Besucherzahlen auf der LOC Webseite um 20%. Weitere Statistiken, vom 23 Oktober 2008, die den Erfolg dieses Projektes bestätigten (Springer et al., 2008):

- LOC-Bilder wurden 10 Millionen mal betrachtet
- 80% der Fotos der LOC wurden als „Favorite“ markiert und erschienen somit auf dem Fotostream anderer User
- 15,000 Benutzer fügten die LOC als Kontakt hinzu und stellten somit einen Link zwischen ihrem Fotostream und dem Fotostream der LOC her
- 8000 Kommentare wurden auf 2873 Fotos abgegeben (von 2562 verschiedenen Benutzern)
- Insgesamt wurden 67,176 Tags hinzugefügt

Das Pilotprojekt war erfolgreich gestartet, und es war nur mehr eine Frage der Zeit bis auch andere Institutionen auf *Commons* vertreten waren. In der Abbildung 3.1 sind 16 ausgewählte Institutionen dargestellt, welche einen hohen Bekanntheitsgrad aufweisen. In dieser Abbildung befinden sich beispielsweise das *Powerhouse Museum (Australien)*[3], die *New York Public Libary (USA)*[4], das *National Maritime Museum (Großbritannien)*[5] und das *Swedish National Heritage Board*[6].

Die gesamte Liste kann im Kapitel 3.2 auf der Tabelle 3.1 eingesehen werden.

[2] Rechtserklärung: `http://www.flickr.com/commons/usage/`
[3] `http://www.flickr.com/photos/powerhouse_museum/` [Zugriff am 13.11.2011]
[4] `http://www.flickr.com/photos/nypl/` [Zugriff am 13.11.2011]
[5] `http://www.flickr.com/photos/nationalmaritimemuseum/` [Zugriff am 13.11.2011]
[6] `http://www.flickr.com/photos/swedish_heritage_board/` [Zugriff am 13.11.2011]

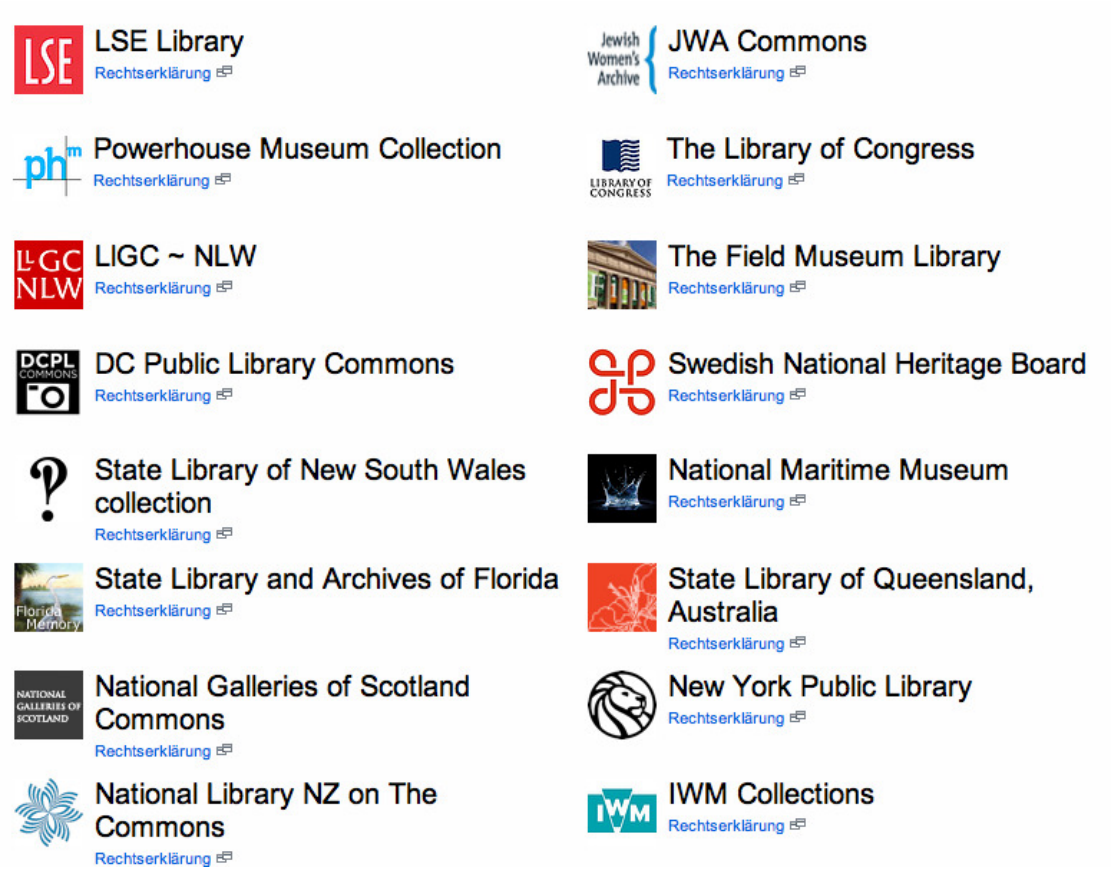


Abbildung 3.1: Ausschnitt von den teilnehmende Institutionen (Flickr, 2011*c*)

3.2 Datenstruktur

In diesem Abschnitt werden die Daten, welche dann im praktischen Teil näher untersucht werden, vorgestellt. Die Datenbasis wurde mithilfe der Flickr-API erstellt und stellt einen Snapshot der Daten zum Zeitpunkt 31.03.2011 dar. Der Aufbau der Datenstruktur ist in der Abbildung 3.2 dargestellt. Es gibt eine Anzahl an Institutionen (ent_institution) die verschiedene Bilder (ent_photos) über ihre Flickr-Profile hochladen. Diese Bilder werden von Usern (ent_user) mit Tags (ent_tags) versehen. Einige User haben im Userprofil auch ihren Aufenthaltsort gepflegt über welchen dann die Taggerlocation (ent_location) ermittelt wird.

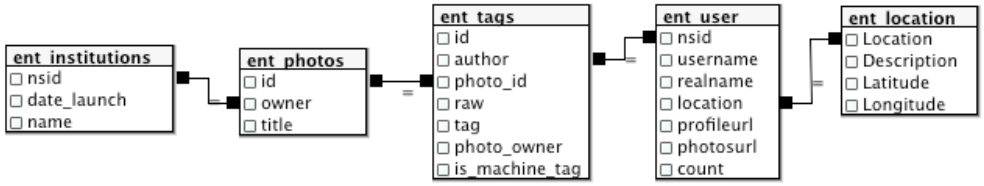


Abbildung 3.2: Datenstruktur *Flickr: TheCommons*

Die Tabelle 3.1 gibt eine Übersicht aller Institutionen, Tags und Kommentare an, welche sich in der Datenbasis befinden. In den Spalten *Tagger bzw. Kommentatoren* befindet sich jeweils die Anzahl an verschiedenen Usern, welche Tags bzw. Kommentare hinterlassen haben.

Betrachtet man diese Tabelle im Detail, so ist ersichtlich, dass die Institutionen nur Anhand ihrer Tags und Kommentare schwer vergleichbar sind. Hier muss auch die Berücksichtigung der Anzahl an verschiedenen Usern erfolgen. Beispielsweise wurden für die *Biblioteca de Arte* 787 Tags vergeben, jedoch nur von einem einzigen User. Dahingegen wurden für die *Australian War Memorial colloection* 804 Tags von 40 verschiedenen Taggern vergeben. Dieses Problem der Vergleichbarkeit wurde im Kapitel 4.2 mit einem eigenen *Beliebtheitsindex* gelöst.

Eine aggregierte Form der nachfolgenden Tabelle ist im Kapitel 4.1 (Tabelle 4.1) abgebildet.

Institution	Tags	Tagger	Kommentare	Kommentatoren
National Maritime Museum	471	6	78	51
National Library of Scotland	3026	5	39	29
Galt Museum Archives on The Commons	820	17	146	101
Powerhouse Museum Collection	409	3	92	49
Smithsonian Institution	486	14	259	107
Muse McCord Museum	1042	35	375	237
Keene and Cheshire County (NH) Historical Photos	781	1	41	34
Bibliothque de Toulouse	101	1	25	23
Biblioteca de Arte	787	1	15	7

National Media Museum	1064	85	1230	883
State Library of New South Wales collection	859	24	383	234
Nationaal Archief	385	12	138	91
Australian War Memorial collection	804	40	900	660
The Library of Virginia	300	1	36	25
Cornell University Library	953	5	19	14
National Galleries of Scotland Commons	1362	84	576	371
The National Archives UK	449	9	154	114
State Library and Archives of Florida	849	8	88	75
Imperial War Museum Collections	53	4	0	0
State Library of Queensland	489	3	11	10
National Library NZ on The Commons	180	10	111	74
New York Public Library	979	13	121	91
Australian National Maritime Museum on The Commons	525	6	149	55
nha.library	577	24	176	91
Swedish National Heritage Board	357	15	269	119
Oregon State University Archives	193	1	35	23
LSE Library	421	7	35	24
The Field Museum Library	1194	3	10	9
Getty Research Institute	578	2	34	26
The U.S. National Archives	461	5	62	27
DC Public Library Commons	370	25	159	96
JWA Commons	1400	10	103	58
Center for Jewish History	686	5	104	39
LlGC NLW	1217	7	64	46
Bergen Public Library	381	10	49	35
Fylkesarkivet i Sogn og Fjordane	3301	4	82	41
UA Archives | Upper Arlington History	1194	3	51	26
SMU Central University Libraries	505	8	120	101
NASA on The Commons	719	8	241	185
Texas State Archives	400	1	17	10
Jewish Historical Society of the Upper Midwest	975	6	13	11
George Eastman House	1072	16	179	115
UW Digital Collections	872	3	36	22
Brooklyn Museum	134	5	33	25
The Library of Congress	426	11	126	86

Lj - smyndasafn Reykjav'kur / Reykjav'k Museum of Photography	177	3	30	18

Tabelle 3.1: Datenbasis mit Institutionen, Anzahl an Tags, Taggern, Kommentaren und Kommentatoren

Da die bereits vorhandenen Tools nicht alle Fragestellungen abdecken konnten, wurden die zwei Analysewerkzeuge FlickrMaps und ClusterTags selbst entwickelt. Beide Werkzeuge machen sich die FlickrAPI zunutze. Bei FlickrMaps handelt es sich um eine Echtzeitanwendung, welche Flickr und GoogleMaps verknüpft. Durch diese Verknüpfung können die *Location*-Daten von Flickr auf einer Landkarte visualisiert werden. Dahingegen arbeitet ClusterTags mit der Datenbasis und clustert naheliegende Ortsinformationen der User zusammen und visualisiert diese schließlich auf einer Google Landkarte. Im nächsten Abschnitt werden diese zwei Anwendungen im Detail erläutert.

3.2.1 FlickrMaps

Diese Webanwendung wurde mithilfe JSP, Javascript und HTML realisiert und visualisiert im großen und ganzen die geografischen Daten der Tagger und des Objektes auf einer Landkarte.

Mithilfe von FlickrMaps sollen folgende Hypthesen geprüft werden:

1. Geografische Analyse der Schlagwörter und Kommentare

2. Darstellung der geografischen Assoziationsbeziehung zwischen Tagger und dem Objekt

3. Darstellung der Abhängigkeit zwischen Aufnahmeort (falls angegeben), Standort der Institution und dem Tagger

In Abbildung 3.3 wird der Ablauf der Anwendung näher erläutert. Im ersten Schritt baut die Applikation eine Verbindung zu Flickr auf. Hierbei werden *SecretKey*, *DeveloperKey* und ein *Token* als Authentifizierung an Flickr übertragen. Nach der erfolgreichen Authentifizierung, liefert Flickr ein *Zugriffsinterface* zurück. Im zweiten Schritt werden dann die Fotoinformationen, Kommentare und Tags zu einem Bild heruntergeladen. Nach der Validierung dieser Daten durch die Applikation, wird eine HTML Seite erzeugt. Diese Webseite wird immer dynamisch generiert und enthält zur Visualisierung der Daten eine Google Landkarte. Alle zu visualisierenden Daten sind als Javascript in dieser dynamisch generierten Webseite eingebunden und werden dann auf einer *Google Map* dargestellt (Abbildung 3.4). In Algorithm 1 ist ein *Pseudocode* zur Generierung der Marker angegeben.

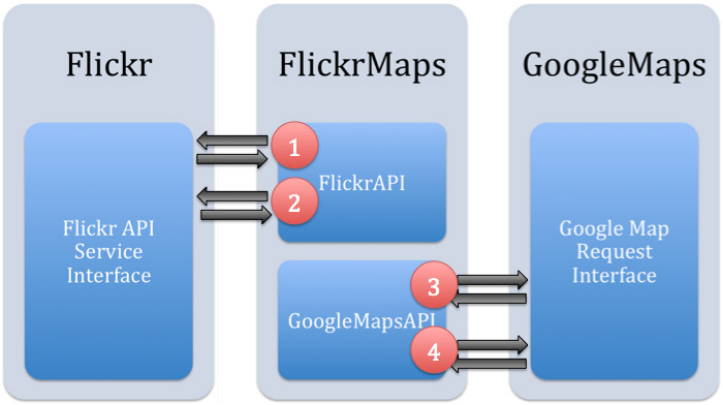


Abbildung 3.3: Architektur der Webanwendung *FlickrMaps*.

Algorithm 1 CreateMarker

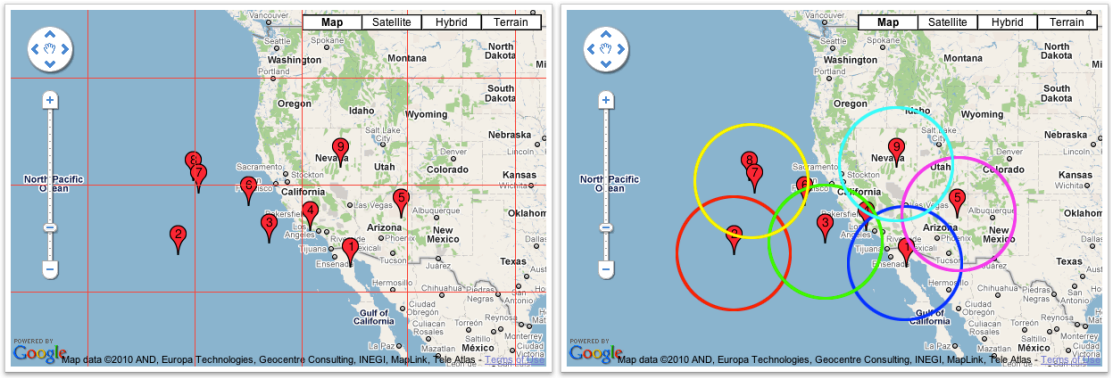
```
1:  procedure CREATEMARKER(type, dataList)
2:      for all t ∈ dataList do                    ▷ Iterate all Items (Tag,Comments) t
3:          t ← VALIDATEDATA(t, tagList)           ▷ t = validated Tag,Comment with Location
4:          icon ← ASSIGNICON(type)                ▷ Different Icons for Tag or Comment
5:          if t is not NULL then
6:              new MARKER ← createMapMarker(t, icon)   ▷ Create new LocationMarker
7:              call DISPLAYMARKER(MARKER)              ▷ Display Marker on Map
8:          end if
9:      end for
10: end procedure
```

3.2.2 ClusterTags

Diese Webapplikation arbeitet mit einer vorher erstellten Datenbank, welche einen Snapshot der „The Commons“-Daten zu einem bestimmten Zeitpunkt darstellt. Dieses Tool baut auf dem GoogleAPI Framework *ClusterMarker* auf. Mit diesem Framework ist es möglich eine große Menge an Ortsmarkierungen auf einer Google Map zu erstellen und diese nach bestimmten Kriterien zusammenzuführen. Bei Clustermarker sind zwei Arten von Clustering möglich:

Beim *Grid-based Clustering* (Abbildung 3.5a) wird die Karte zunächst mal in einzelne Quadrate unterteilt. Die Größe dieser Quadrate hängt von der gewählten Zoomstufe ab und wird bei jeder Vergrößerung oder Verkleinerung der Karte neu berechnet. Danach werden alle Ortsmarkierungen in diesem Quadrat zu einer Markierung gruppiert. Im Gegegensatz dazu werden beim *Distance-based Clustering* die Ortsmarkierungen nach dem geringsten Abstand zwischen der Markierung und dem Clustermittelpunkt (Centroid) gruppiert. Der Pseudocode für das Clustering ist in Abbildung 2 angegeben. Für den Originalquellcode siehe Appendix C.

Abbildung 3.4: Screenshot of FlickrMaps Webapplication



(a) Grid-based **(b)** Distance-based

Abbildung 3.5: Clustering varianten

Algorithm 2 Calculate Cluster

```
1:  procedure CALCULATECLUSTER(arrayMarker)
2:      for all m1 ∈ arrayMarker do                ▷ Iterate all positioned markers t
3:          if m1 is NULL then
4:              CONTINUE
5:          end if
6:          for all m2 ∈ arrayMarker do            ▷ Iterate all positioned markers m
7:              if m2 is NULL then
8:                  CONTINUE
9:              end if
10:             p1 ← M1.point                      ▷ retrieve coordinates m1
11:             p2 ← M2.point                      ▷ retrieve coordinates m2
12:             x := p1.x − p2.x                   ▷ calculate x-difference
13:             y := p1.y − p2.y                   ▷ calculate y-difference
14:             if x^2 + y^2 < compareDistance(p1, p2) then
15:                 CLUSTER.ADD(m2)                ▷ add marker to cluster
16:             end if
17:         end for
18:     end for
19: end procedure
```

Das Clustering der Daten kann eigentlich mit dem *k-Means++* Algorithmus verglichen werden. Dieser Algorithmus stellt eine verbesserte Variante des *k-Means* dar und beinhaltet laut (Wikipedia, 2011*d*) folgende Schritte:

1. Wähle als ersten Cluster-Schwerpunkt zufällig ein Objekt aus
2. Berechne für Objekt den Abstand $D(x)$ zum nächsten Cluster-Schwerpunkt
3. Wähle zufällig als nächsten Cluster-Schwerpunkt ein Objekt aus. Die Wahrscheinlichkeit mit der ein Objekt ausgewählt wird, ist proportional zu $D^2(x)$, d.h. je weiter das Objekt von den bereits gewählten Cluster-Schwerpunkten entfernt ist desto wahrscheinlicher ist es, dass es ausgewählt wird.
4. Wiederhole Schritt 2 und 3 bis k Cluster-Schwerpunkte bestimmt sind
5. Führe nun den üblichen k-Means Algorithmus aus

KAPITEL 4

Empirischer Teil

4.1 Praktische Umsetzung

In den nachstehenden Kapiteln werden die Ergebnisse der Datenanalyse beschrieben. Die Analyse der Daten erfolgte durch verschiedenste Tools, neben den eigenen Entwicklungen *FlickrMaps* und *ClusterTags* (Siehe Kapitel 3.1) kamen auch *Weka*[1] und *PostgreSQL-Database*[2] zum Einsatz.

Um eine besser Übersicht zu erhalten, zeigt die Tabelle 4.1 eine aggregierte Form der Datenbasis (für mehr Details siehe Anhang B). In den Daten sind 46 Institute enthalten, welche 4396 Bilder auf *Flickr* zur Verfügung gestellt haben. Diese Bilder wurden mit 34784 Tags von 569 Benutzern annotiert. 4517 User haben 6936 Kommentare hinterlassen.

Datenelemente	**Einträge**
Institutionen	46
Fotos	4396
Tags	34784
Kommentare	6936
Tagger	569
Kommentatoren	4517

Tabelle 4.1: Aggregierte Datenbasis

4.2 Institutionen

In diesem Abschnitt sollen die auf *Commons* vertretenen Institutionen näher untersucht werden. Die meisten Institutionen kommen aus den USA, Großbritannien und Australien (Siehe Tabel-

[1] `http://www.cs.waikato.ac.nz/ml/weka/` [Zugriff am 11.12.2011]
[2] `http://www.postgresql.org/` [Zugriff am 11.12.2011]

le 4.2). Insgesamt ist die angloamerikanische Region mit 38 von 46 Einrichtungen sehr stark vertreten.

Region	Land	Institutionen
Australien	Australien	4
	Neuseeland	1
Europa	Frankreich	1
	Island	1
	Niederlande	1
	Norwegen	2
	Portugal	1
	Schottland	1
	Schweden	1
	UK	6
Nord Amerika	Kanada	2
	USA	25

Tabelle 4.2: Herkunft der Institutionen

In der Datenbasis befinden sich im Mittelwert 100 Fotos pro Institution, welche mit einer sehr unterschiedlichen Menge an Tags und Kommentaren versehen wurden. Um die Institutionen mit dem höchsten *Benutzerinteresse* zu finden, wurde ein eigener *Beliebtheitsindex* P verwendet. Dieser Index normiert die Anzahl der Tags, Kommentare, Tagger und Kommentatoren auf Basis der jeweiligen Maximalwerte. Dadurch ist der Index nicht rein von der Anzahl der Tags und Kommentare abhängig, sondern berücksichtigt auch die Menge an verschiedenen Benutzern. Auch wird durch die Standardisierung ausgeschlossen, dass einzelne *Poweruser* die Beliebtheit einer Institution hochtreiben.

Definition 5: *Beliebtheitsindex* $P := \frac{t}{\max T} + \frac{c}{\max C} + \frac{ta}{\max T_A} + \frac{k}{\max K}$

- t = Anzahl Tags, T = Menge Anzahl aller Tags, c = Anzahl Kommentare, C = Menge Anzahl alle Kommentare, t_A = Anzahl Tagger, T_A = Menge Anzahl aller Tagger, k = Anzahl Kommentatoren, K = Menge Anzahl alle Kommentatoren.

- wobei $t \in T, c \in C, t_A \in T_A, k \in K$

- $0 \leq P \leq 4$

Die Abbildung 4.1 gibt eine grafische Gegenüberstellung über die Anzahl an Tags, Kommentaren, Taggern und Kommentatoren von zehn Institutionen an, welche einen sehr hohen Beliebtheitsindex aufweisen. Der Index selbst wird mit der selbst definierten Formel (Definition 5) berechnet. Die Tabelle 4.3 zeigt die zehn Institutionen mit dem höchsten Beliebtheitsindex.

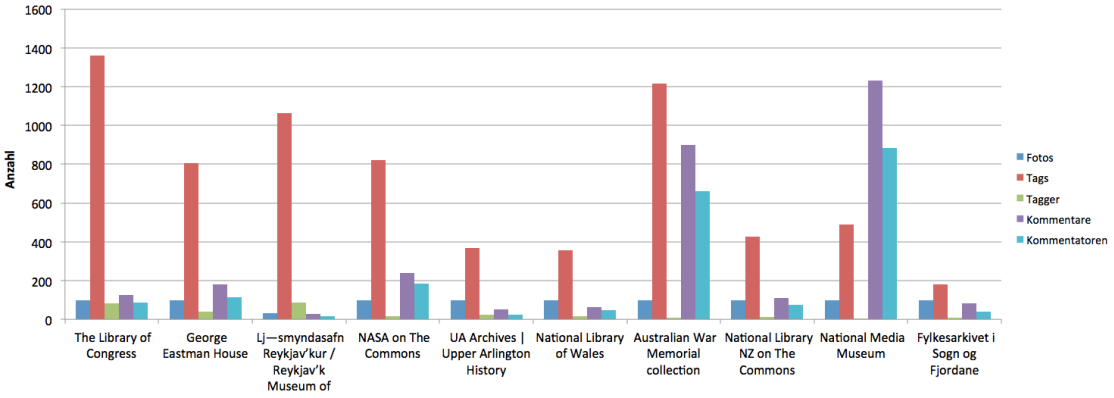


Abbildung 4.1: Grafische Darstellung der Tags, Kommentare, Tagger und Kommentatoren von zehn Institutionen

Institution	Tags	Tagger	Kommentare	Kommentatoren	Index
National Media Museum	489	3	1230	883	2,1
Australian War Memorial collection	1217	7	900	660	1,9
The Library of Congress	1362	84	126	86	1,6
Museum of Reykjav'k	1064	85	30	18	1,3
Galt Museum	3026	5	146	101	1,2
Oregon State University Archives	3301	4	35	23	1,1
National Galleries of Scotland Commons	409	3	576	371	1,0
George Eastman House	804	40	179	115	0,9
Muse McCord Museum	1194	3	375	237	0,9
NASA on The Commons	820	17	241	185	0,8

Tabelle 4.3: Institutionen sortiert nach dem Beliebtheitsindex

Das *National Media Museum*[3] und die *Australian War Memorial collection*[4] weisen einen Index von 2,183 und 1,93 auf (Siehe Tabelle 4.3). Dieser Wert lässt sich auf eine beträchtliche Anzahl an Kommentaren zurückführen. Dahingegen weist die *The Library of Congress*[5] eine geringere Menge an Kommentaren auf, aber hier wird der Index durch die hohe Anzahl an Schlagwörtern und verschiedenen Taggern dominiert. Diese Beispiele zeigen die Robustheit und Unabhängigkeit des Index gegenüber einer Kennzahl.

[3] `http://www.nationalmediamuseum.org.uk/` [Zugriff am 11.12.2011]
[4] `http://www.awm.gov.au/database/` [Zugriff am 11.12.2011]
[5] `http://www.loc.gov/index.html` [Zugriff am 11.12.2011]

4.3 Bilder

Im nächsten Abschnitt sollen einige interessante Bilder der *Commons*-Kollektion präsentiert werden. Als Basis für das Auswahlverfahren wurde die Anzahl an Tags und Kommentaren, sowie der im vorherigen Kapitel beschriebene Beliebtheitsindex herangezogen.

Für alle Bilder aus der Datenbasis wurde der Beliebtheitsindex berechnet. Die Tabelle 4.4 zeigt 10 Bilder mit dem höchsten Index. Das Bild mit der *PhotoID 3007981750* (Rang 1) weist einen Index von 2,72 auf. Dieser Wert lässt auf die hohe Beteiligung unterschiedlicher User an den Kommentaren und Tags zurückführen. Für dieses Bild wurden 12 Tags vergeben, 11 davon stammen von verschiedenen Usern. Die Kommentare wurden auch von einer großen Anzahl an verschiedenen Flickr Benutzern vergeben.

Den zweitbesten Index weist das Bild mit der *PhotoID 3588772325* (Rang. 2) auf, obwohl die Anzahl an Kommentatoren höher als im ersten Bild ist. Für den schlechteren Index sind in diesem Fall die Tags (6 Tags von 3 Taggern) verantwortlich, da sowohl die Tags als auch die Kommentare mit ihrer Anzahl an unterschiedlichen Autoren in die Berechnung einfließen.

Rang	PhotoID	Tags	Tagger	Kommentare	Kommentatoren	Index
1	3007981750	12	11	119	118	2,72
2	3588772325	6	3	155	151	2,52
3	3102880444	16	12	90	86	2,49
4	3084876560	7	7	98	95	1,99
5	3527155504	4	4	123	118	1,95
6	3527157206	7	5	104	104	1,92
7	3526345203	5	5	94	93	1,70
8	3110325921	12	10	44	43	1,65
9	3527160566	7	7	67	66	1,53
10	2922445156	11	10	37	34	1,53

Tabelle 4.4: Bilder sortiert nach dem Beliebtheitsindex

In den folgenden Kapiteln werden einige Bilder vorgestellt. Diese Bilder wurden nach folgenden Kriterien ausgewählt:

- Bild mit den meisten Tags
- Bild mit den meisten Kommentaren
- Bild mit dem höchsten Beliebtheitsindex

4.3.1 Bild mit den meisten Tags

Die Abbildung 4.2 zeigt, dass Bild von *John Turner* (Links) und seiner Familie. Dieses Bild wurde mit 47 Tags versehen. John war ein Offizier in *Green Howards*[6] in England während des

[6]Infanterie-Regiment der britischen Armee

ersten Weltkrieges.

Abbildung 4.2: Paul Stang, *Familie Turner*, Fylkesarkivet i Sogn og Fjordane

4.3.2 Bilder mit den meisten Kommentaren

Die Abbildung 4.4 zeigt eine Mutter mit ihrem Kind. Dieses Bild wurde 155 mal kommentiert. Das Bild wurde 1912 von *Henry Essenhigh Corke* mit dem *Autochromverfahren* aufgenommen (Wikipedia, 2011*a*).

Abbildung 4.3: Henry Essenhigh Corke, *Mother and Child*, National Media Museum

Dieses Bild hat ebenso wie das vorherige 155 Kommentare und zeigt einen amerikanischen Soldaten in Australien.

Abbildung 4.4: John Earl, *An American soldier with a joey 1942*, Australian War Memorial Collection

4.3.3 Bild mit dem besten Index

Das Bild in Abbildung 4.5, weist einen Index von 2,72 auf und zeigt eine Truppe die als Unterstützung, nacheilt.

Abbildung 4.5: Frank Hurley, *Supports going up after battle to relieve the front trenches*, National Media Museum

4.4 BenutzerInnen

In diesem Abschnitt wird ein genauer Blick auf die Benutzer (Tagger) von *The Commons* geworfen. Insgesamt gab es 257 verschiedene User aus 129 unterschiedlichen Regionen[7]. Die *Ortsangaben* wurden aus den Benutzerprofilen extrahiert und werden in der Tabelle 4.5 dargestellt.

Location	**User**
Keine	13097
USA	6412
Schottland	2679
AUS	1263
Wales	1192
Kanada	973
UK	448
Schweden	311
Norwegen	276
Neuseeland	154
Irland	129
Niederlande	17
Griechenland	14
Italien	12
Russland	12
Deutschland	7
Frankreich	7
Portugal	4
Estland	3
Sri Lanka	3
Israel	2
Mexiko	2
Serbien	1
Summe	27019

Tabelle 4.5: Herkunft der Tagger (Länder)

50% der User (13097) hatten keine *Location* in ihrem Profil eingetragen. Die meisten User stammen aus den *USA* (6412), Schottland (2679) und Australien (1263). Eine aggregierte Version der Länder wird in der Abbildung 4.6 dargestellt. Aus dieser Abbildung ist ersichtlich, dass $\frac{13903}{13922}$ =99,9% der User aus *Nord Amerika*, *EU* oder *Australien* stammen. Die restlichen Regionen *Süd Amerika*, *Naher Osten* und *Asien* sind mit einer sehr geringen Benutzeranzahl vertreten.

Auch wurden die Tags pro Benutzer genauer betrachtet. Die Tabelle 4.6 zeigt, dass die ersten acht User für 45% aller Tags verantwortlich sind.

[7]Regionen: Bundesländer, Länder

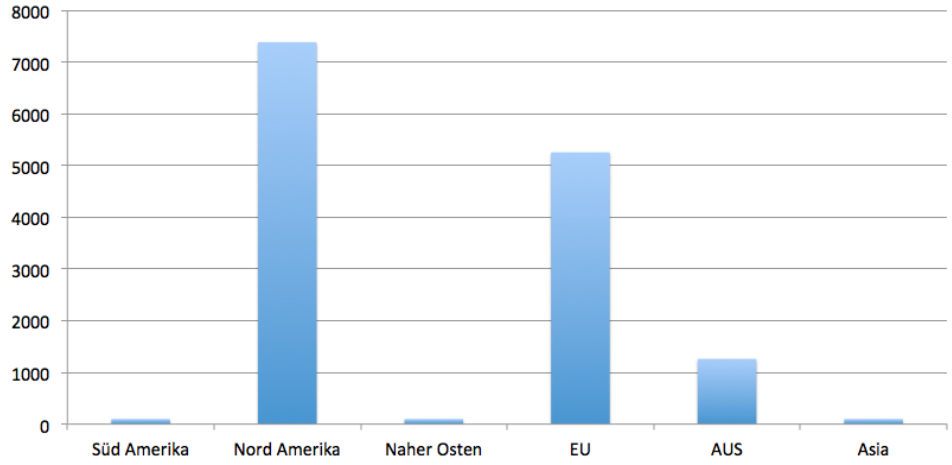


Abbildung 4.6: Herkunft der Tagger (Region)

Rang	User	Tags
1	37547255@N08	3218
2	14456531@N07	2624
3	11418107@N02	1678
4	37199428@N06	1192
5	37784107@N08	1186
6	35310696@N04	1184
7	36281769@N04	922
8	32951986@N05	871
	Summe	12875

Tabelle 4.6: Tags pro User

Je mehr User in die Analyse einfließen, desto mehr passt sich die Verteilung der Tags pro User an eine Potenzverteilung an (Abbildung 4.7).

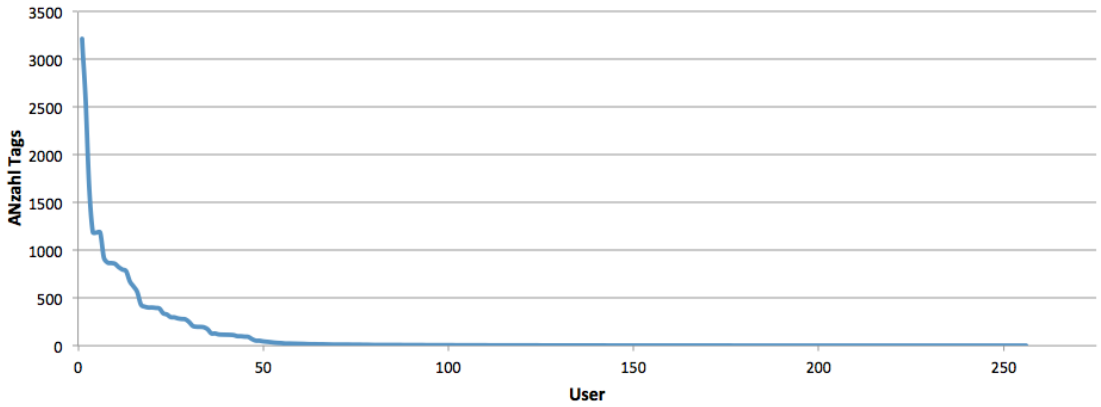


Abbildung 4.7: Verteilung der Tags aller Users. 20% der User (*Poweruser*) sind für 80% der Tags verantwortlich.

Unabhängig von der Ebene der Abstraktion, entsteht in den meisten Fällen eine Potenzverteilung, egal ob man einen oder mehrere User betrachtet. In Abbildung 4.8 wird die Verteilung der Tags eines Benutzers betrachtet. Diese Verteilung zeigt, dass nur 20% der Tags 80% der Gesamthäufigkeit der Schlagwörter ausmachen d.h. ein User verwendet sehr häufig dasselbe Vokabular um verschiedene Objekte zu beschreiben.

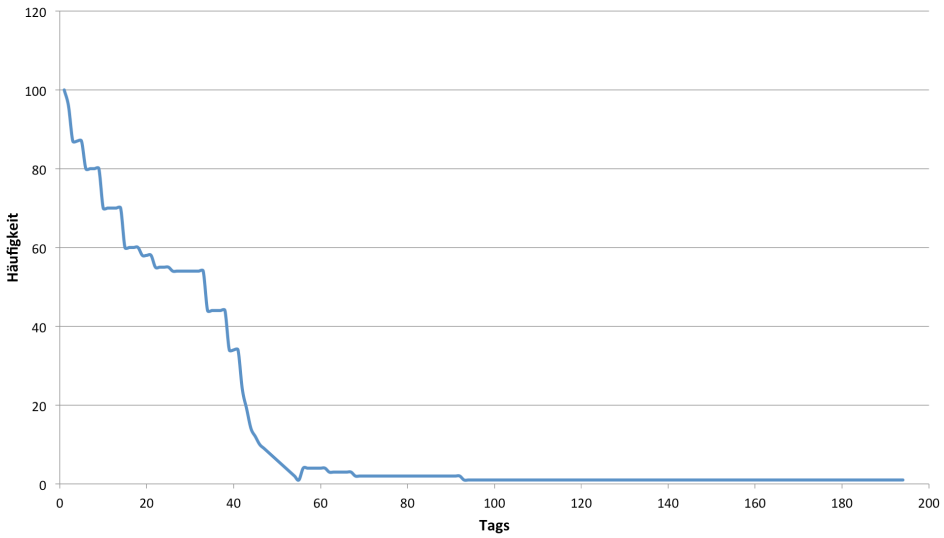


Abbildung 4.8: Verteilung der Tags eines Users

4.5 Schlagwörter

In der Datenmenge befinden sich insgesamt ca. 34,000 Tags. Sehr häufige Tags beschreiben eher generelle Aspekte und seltene Tags eher spezielle Eigenschaften eines Objekts. Beispielsweise ist der Tag *Portrait* (siehe Tabelle B) ein sehr genereller Begriff, da er ein Gemälde, eine Fotografie oder eine andere künstlerische Darstellung einer oder mehrerer Personen bezeichnet (Wikipedia, 2011*g*). Auch die weiteren Tags wie *History*, *woman*, *The Great War* sind Oberbegriffe. Im Gegensatz dazu beschreiben seltene Schlagwörter wie z. B. *Mar del Norte* und *Robeson, Paul (1898-1976)* ein ganz spezifisches Detail. Beispielsweise *Mar del Norte* ist die spanische Bezeichnung für die Nordsee und gibt im Gegensatz zum Obergriff *Meer* viel mehr Auskunft. Generelle Begriffe wie *Portrait* werden durch Schlagwörter wie *Robeson, Paul (1898-1976)* erweitert und auf eine ganz neue Detailebene gebracht.

Rang	Tag	Anzahl
1	portrait	267
2	women	195
3	history	189
4	woman	177
5	jewish	173
6	worldwari	160
7	usa	155
8	thegreatwar	144
9	worldwarone	134
10	man	133

(a) Top 10 Schlagwörter

Rang	Tag	Anzahl
8001	Major General Sir William Throsby Bridges	1
8002	Manned Spacecraft Center	1
8003	Mar del Norte	1
8004	Million dollar mermaid	1
8005	Miss Brazil	1
8006	Morris, William (1889-1979)	1
8007	Manned Spacecraft Center	1
8008	Robeson, Paul (1898-1976)	1
8009	Manned Spacecraft Center	1
8010	Rue de la Casbah	1

(b) Flop 10 Schlagwörter

Tabelle 4.7: Sehr häufige Schlagwörter im Vergleich zu selten vorkommenden Tags

Auch die Anzahl an verwendeten Zeichen der Schlagwörter wurde analysiert. Die nachstehende Tabelle gibt eine Übersicht der sechs häufigsten Wortlängen.

Nr	Wortlänge	Anzahl
1	8 Zeichen	2249
2	5 Zeichen	2185
3	6 Zeichen	1916
4	7 Zeichen	1901
5	9 Zeichen	1729
6	10 Zeichen	1680

Tabelle 4.8: Analyse der verwendeten Tag-Wortlänge.

Die durchschnittliche Zeichenlänge wurde anhand des gewichteten arithmetischen Mittels berechnet: $\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$ = 12,9168. Vergleicht man diesen Mittelwert von 13 Zeichen pro Tag, mit der Tabelle Ba so ist ersichtlich, dass alle Tags der Top 10 Tabelle eine Wortlänge unter 13 Zeichen aufweisen. Das gleiche gilt auch für Tabelle Bb der nicht so häufigen Tags, hier haben alle Tags mehr als 13 Zeichen.

Die Theorie, dass die Wortlänge mit der Häufigkeit des Wortes in Verbindung steht, basiert auf George Kingsley Zipf[8]. Der Linguist war überzeugt davon, dass wir den Aufwand beim Sprechen und Schreiben so gering wie möglich halten möchten und deswegen sind häufig gebrauchte Wörter kürzer. Zipf spricht in diesem Zusammenhang vom Prinzip der geringsten Anstrengung (Wikipedia, 2011*i*; Köhler, 2005).

In der Wissenschaft bezeichnet man dieses Gesetz als *falsches Zipfsches Gesetz*. Es besagt, dass Äußerungen in einer Sprache immer aus einem Kompromiss zwischen zwei entgegengesetzten Tendenzen im Sprecher entstehen (Wikipedia, 2011*i*):

- einerseits aus dem Wunsch, eine Information möglichst verständlich zu vermitteln, was zu Wiederholung (Redundanz) und Ausführlichkeit führt, und

- andererseits aus Sparsamkeit, dem Bedürfnis, möglichst wenig physische und geistige Energie bei der Sprachproduktion aufzuwenden.

Neben diesem Zusammenhang folgt auch die gesamte Verteilung der Schlagwörter einer Power-Law-Kurve (siehe 4.9). Die ersten 20% der Tags (8000 von 34000) machen 80% der Häufigkeit aller Tags aus. Diese 80% werden auch als sogenannten *Long Tail* bezeichnet.

Durch diese Abbildung wird auch ersichtlich, dass das verwendete Vokabular sich zwar voneinander deutlich unterscheidet, aber der größte Anteil der vergebenen Tags sich auf wenige, häufige Tags verteilt.

[8] US-amerikanischer Linguist

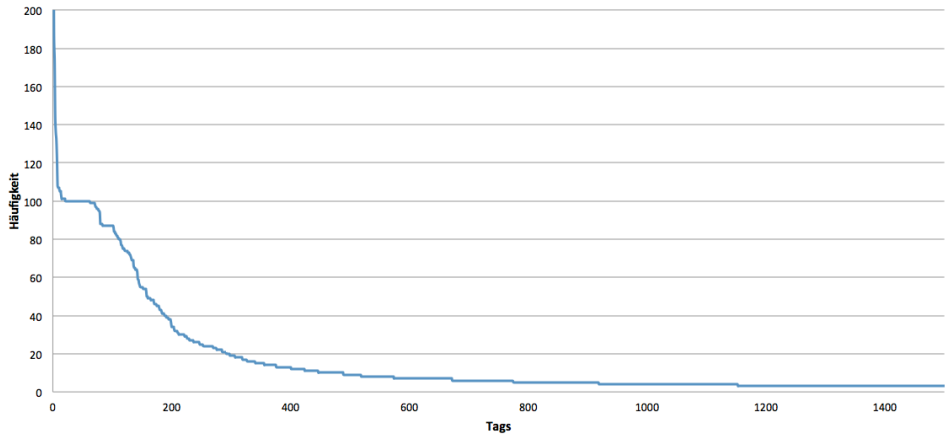


Abbildung 4.9: Verteilung der aller Tags

Um eine bessere Übersicht über Schlagwörter und über die Objekte die sie beschreiben zu erhalten, wurden die Bilder in fünf Klassen, abhängig von der Anzahl ihrer Schlagwörter unterteilt (Siehe Tabelle 4.9). Diese Tabelle zeigt, wie viele Bilder in einer Klasse vorkommen. Anhand der Spalte *Useranteil* wird außerdem ersichtlich, wie viel Prozent der Tags von den Usern vergeben wurden. Der restliche Anteil der Tags wurde somit von den Institutionen selbst vergeben.

Beispielsweise wurden die meisten Bilder mit 1-9 Tags versehen, wobei hier 65% der Tags von den Benutzern stammen und 35% von den Institutionen selbst. In Klasse 2 wurden bereits 70% der Schlagwörter von den Usern annotiert und somit nur mehr 30% von den Institutionen. Mit der Zunahme der Tags wird der *User*-Anteil immer höher, bis er schließlich bei Klasse 5 90% beträgt.

Nr	**Tags**	**Fotos**	**Useranteil**
Klasse 1	1 - 9	2879	65%
Klasse 2	10 - 19	644	70%
Klasse 3	20 - 29	121	80%
Klasse 4	30 - 39	87	85%
Klasse 5	40 - 50	15	90%

Tabelle 4.9: Unterteilung der Tags in Klassen

In Abbildung 4.10 und 4.11 sind die Tagclouds der Klasse 1 und 5 visualisiert. Durch diese zwei Abbildungen ist ersichtlich, dass die Anzahl an verschiedenen Sprachen von Klasse zu Klasse immer höher wird. In Klasse 1 befinden sich ausschließlich englische Tags, wohingegen in Klasse 5 sehr viele skandinavische Tags enthalten sind.

Im Anhang A wird für alle fünf Klassen jeweils ein Bild präsentiert.

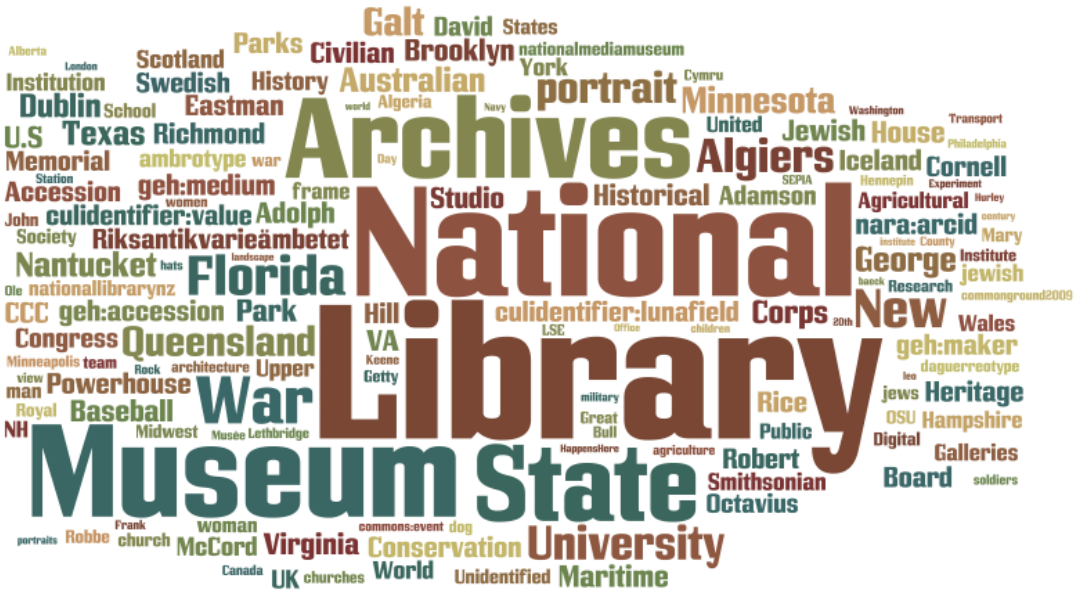


Abbildung 4.10: Tagcloud der Klasse 1

Abbildung 4.11: Tagcloud der Klasse 5

4.6 Kommentare

Auch in den Kommentaren steckt sehr viel Information. In Abbildung 4.12 wird die Anzahl der Zeichen der insgesamt 7014 Kommentare dargestellt. 63% der Kommentare weisen eine Zeichenlänge von maximal 100 auf. Der längste Kommentar umfasst 32691 Zeichen und beinhaltet neben reinen Textinformationen auch sehr viele HTML-Tags. Der kürzeste Kommentar weist eine Zeichenlänge von zwei auf. Dabei handelt es sich um einen *Smiley :)*.

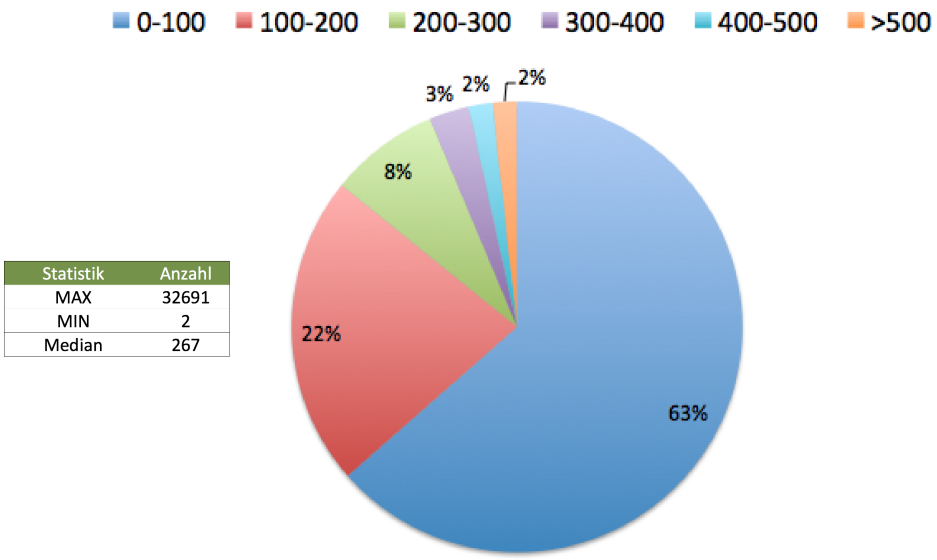


Abbildung 4.12: Zeichenlänge der Kommentare

Im Gegensatz zu Tags beschreiben Kommentare keine Charakteristika der Ressourcen. Sie drücken viel mehr die mit dem Objekt verknüpfte Emotion aus. Extrahiert man die Schlagwörter aus den Kommentaren, so erhält man die in der Tabelle 4.10 dargestellten Begriffe. Das häufigste Kommentar in der Datensammlung war „Hi, I'm an admin for a group called..“, daraus resultierend sind auch die Keywords *called*, *group* und *admin* am häufigsten bei der Auswertung vertreten. Mit diesem Kommentar will man das Bild in einer anderen Gruppe zusätzlich aufnehmen. Danach folgen *love*, *beautiful*, *great* und *amazing*, welche alle samt Emotionen zu den Bildern ausdrücken. Neben reinem Text, enthalten 42% der Kommentare auch HTML Information (siehe Abbildung 4.13). Die *HTML* Kommentare können wiederum in *Link-Tags* und *Image-Tags* unterteilt werden. Auch lassen sich bei 12% der Kommentare direkt Emotionen durch Erkennung von *Emoticon* feststellen. *Emoticons* sind Zeichenfolgen (aus normalen Satzzeichen) die Gefühlszustände nachbilden. Wobei hier auch wiederum zu 70% ein *Emoticon* sofort auf Eigenschaftswort(z. B. *Love* :>, *beautiful* ;D etc.) folgt.

Weiters wurde auch die Relation zwischen den Tags und den Kommentaren untersucht. Nur 5% der Schlagwörter waren auch in den Kommentaren wieder zu finden. Diese Relation gilt auch für die Verfasser. User welche Schlagwörter vergeben hatten, waren in sehr seltenen Fällen

auch als Kommentatoren tätig. Hier wurde eine Übereinstimmung von 10% berechnet.

Nr	Keyword	Anzahl
1	called	69
2	group	63
3	quot	62
4	admin	61
5	love	51
6	beautiful	43
7	Sydney	24
8	photos	23
9	great	23
10	wonderful	22
11	amazing	22

Tabelle 4.10: Häufigste Keywords aus den Kommentaren

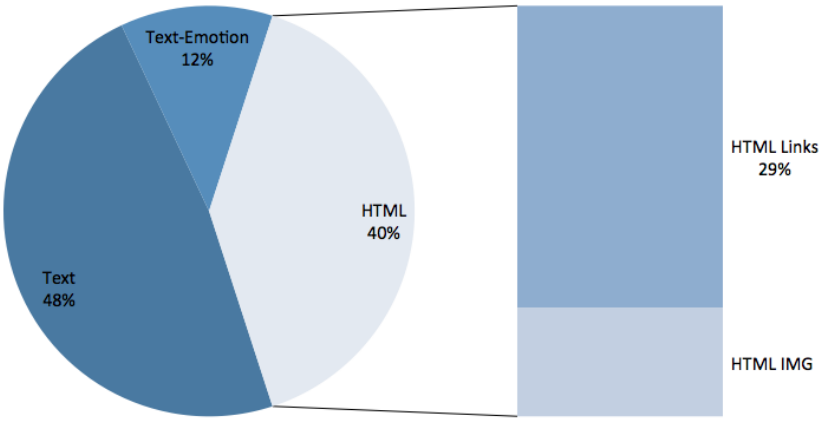


Abbildung 4.13: Aufteilung der Kommentare

Die untersuchten Kommentare weisen sehr häufig auf falsche oder fehlende Informationen (Falscher Bildtitel, Tags) hin. Um diese Art von Kommentaren zu finden, wurden nach den Wörtern *wrong* bzw. *false* gesucht. Die Auswertung zeigte, dass fast jedes vierte Kommentar diese beiden Wörter enthielt (Siehe Abbildung 4.14).

Auch muss an dieser Stelle hinzugefügt werden, dass das alleinige Auffinden dieser Wörter in einem Kommentar kein verlässlicher Indikator für fehlerhafte Informationen sein muss. Um hier ein genaueres Ergebnis zu liefern, muss auch der Inhalt der Kommentare genauer analysiert werden. Dies könnte beispielsweise mit einer *Stimmungserkennung* durch *Sentiment*-Analyse erfolgen (Choi et al., 2009). Jedoch würde diese Analyse den Rahmen dieser Diplomarbeit sprengen und wurde somit nicht durchgeführt.

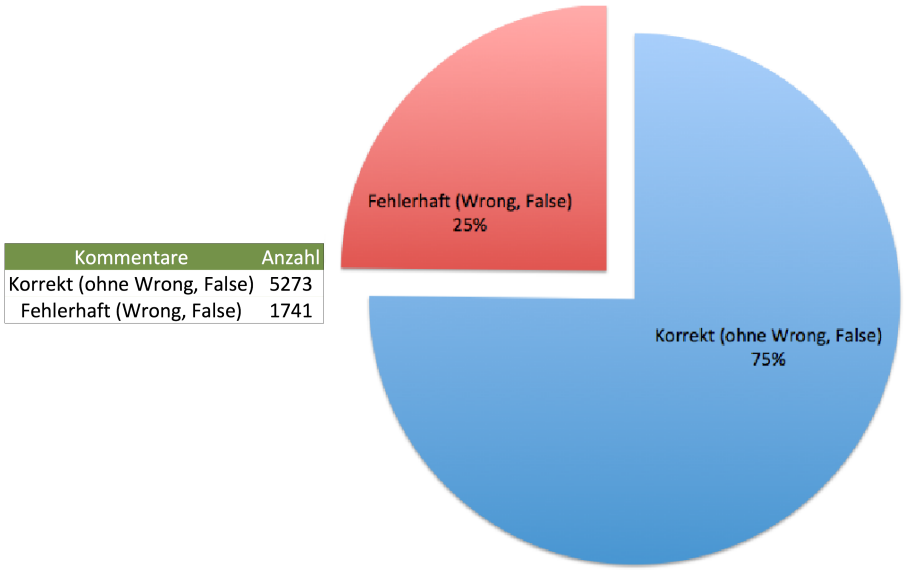


Abbildung 4.14: Verhältnis fehlerhafter Kommentare

4.7 Geograpischer Zusammenhang

In diesem Abschnitt wird die geografische Relation zwischen Tagger und den Kultureinrichtungen untersucht. Im ersten Teil wird der geografische Einfluss der auf *The Commons* vertretenen Institutionen errechnet. Wobei hier zwischen lokalem und globalem Einfluss unterschieden wird.

Um einen globalen bzw. lokalen Einfluss zu messen, wurde für diese Arbeit folgende Definition festgelegt:

- Ein globaler Einfluss herrscht dann, wenn zumindest $\frac{1}{3}$ der Tags von Usern, welche nicht aus dem gleichen Land wie die Institution stammen, vergeben wurden. Ansonsten ist der Einfluss lokal.

Abschließend wird auch die geografische Verteilung der in Kapitel 4.3 vorgestellten Bilder visualisiert. Zu diesem Zweck kommt die Echtzeit-Applikation *FlickrMaps* zum Einsatz.

Alle nachfolgenden Berechnungen wurden auf Basis der Stichprobe (Datenbasis) errechnet und beziehen sich auf das Herkunftsland der Institution und der Tagger. Wobei nur Tags zu denen eine geografische Angabe vorhanden war, in die Analyse einflossen.

Mithilfe der Applikation *ClusterTags* wurden für alle in der Datenbasis enthaltenen Institutionen der geografische Einfluss ermittelt. Folgende Schritte werden benötigt um den Einfluss der Einrichtungen zu ermitteln:

1. Standorte aller Institutionen recherchieren
2. Filtern der Tags (nur Tags mit *User-Location* Eintrag)
3. Validierung der *Location* (Überprüfung auf Korrektheit der *Locations*)
4. Ermittlung der geografischen Daten der Schlagwörter (Herkunft der User/Tagger)
5. Auswertung und Visualisierung der Daten
6. Berechnung des Einflusses auf Basis der Herkunftsländer der Institution und Tagger

Die Schritte 1 und 5 wurden manuell errechnet (*Internet und Excel*). Alle anderen Schritte (2-4) werden durch die Anwendung *ClusterTags* automatisch durchgeführt.

4.7.1 Globaler Einfluss

Die Tabelle 4.11 zeigt die Institutionen, welche einen globalen Einfluss aufweisen, wobei die Spalte *Global/Lokal* das Verhältnis zwischen globalen und lokalen Schlagwörtern darstellt. Insgesamt konnte bei 4 von 46 Institutionen ein globaler Einfluss nachgewiesen werden.

Institution	Global/Lokal
National Media Museum (UK)	238 / 17
Muse McCord (Kanada)	110 / 214
Swedish National Heritage Board (Schweden)	79 / 253
Reykjav'k Museum (Island)	27 / 0

Tabelle 4.11: Institutionen mit globalen Einfluss

Beim *National Media Museum*[9] (Bradford, Großbritannien) wurden sogar 238 Schlagwörter von nicht britischen Taggern vergeben und nur 17 Tags von lokalen Usern. Die Visualisierung durch *ClusterTags* zeigt die geografische Verteilung der Tagger (Siehe 4.17).

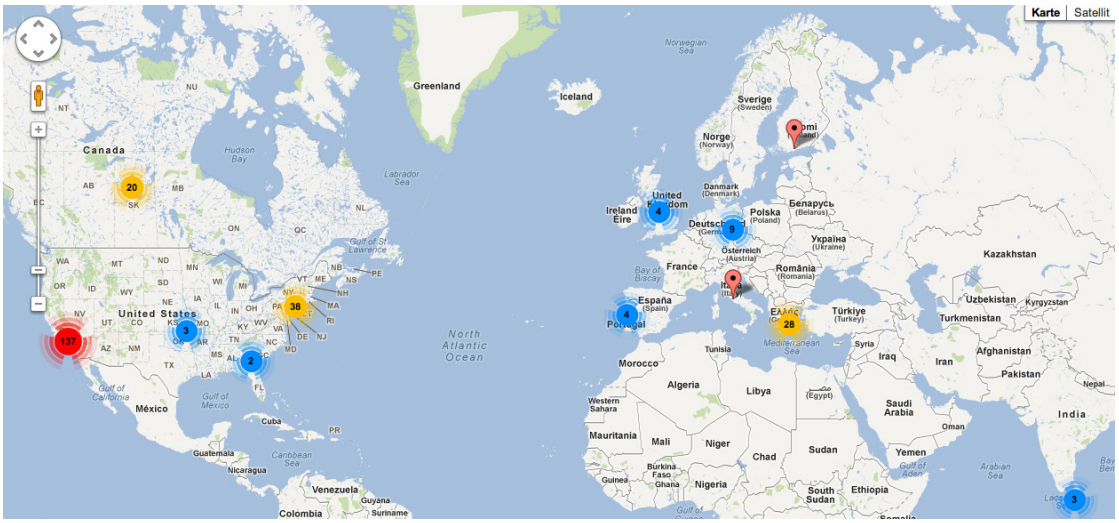


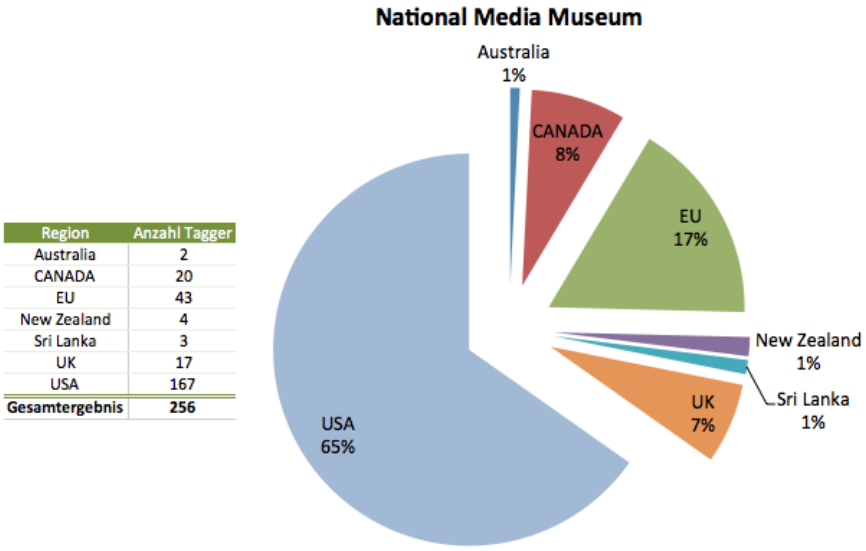
Abbildung 4.15: *ClusterTag*-Visualisierung der Herkunft der Tagger (*National Media Museum*)

Gruppiert man diese verschiedenen Ortsangaben der Tagger zu Regionen bzw. Ländern, so erhält man die Abbildung 4.16. In dieser Abbildung ist ersichtlich, dass die Mehrheit der User des *National Media Museum* aus den USA stammen (65%). Die Europäische Union dahingegen ist nur mit 17% (ohne UK) der User vertreten. Darüber hinaus weist diese *Institution* auch ein relativ geringeres nationales Interesse von 7% auf.

Das *Reykjav'k Museum (Island)* weist einen globalen Einfluss von 100% auf. Aus der Tabelle 4.11 ist ersichtlich, dass für dieses Museum überhaupt keine lokalen Tags vergeben wurden. Im Anhang (Tabelle B.2) sind die Einzelnachweise aller lokalen und globalen Tags für alle Einrichtungen aufgelistet. Aus dieser Tabelle kann abgelesen werden, dass die 27 Tags des *Reykjav'k Museum (Island)* alle samt von einem User aus *San Francisco, USA* vergeben wurden. In diesem Fall handelt es sich um einen *Poweruser* aus den USA, der ein großes Interesse an den Archiven dieser Kultureinrichtung aufweist.

[9] `http://www.nationalmediamuseum.org.uk/` [Zugriff am 20.11.2011]

Da aber nur 27 globale Tags in der Datenbasis enthalten sind und alle von einem User vergeben wurden, hat dieses Ergebnis wenig Aussagekraft. Um hier eine bessere Aussage zu treffen, muss die Datenbasis vergrößert werden und der geografische Einfluss erneut berechnet werden.



Region	Anzahl Tagger
Australia	2
CANADA	20
EU	43
New Zealand	4
Sri Lanka	3
UK	17
USA	167
Gesamtergebnis	**256**

Abbildung 4.16: Verteilung der Tags von *National Media Museum* nach Regionen.

Auch für das *Muse McCord*[10] (Quebec, Kanada) wurde ein globaler Einfluss errechnet. In der Tabelle 4.12 bzw. der Abbildung 4.17 sind die Ortsangaben der Tagger ersichtlich. Aus diesen Darstellungen kann abgelesen werden, dass die meisten User aus dem *angloamerikanischen* Bereich stammen. Betrachtet man diesen Bereich als eine Region und bezieht den geografischen Einfluss darauf, dann wäre der Einfluss als 100% lokal zu sehen (Siehe Tabelle 4.13).

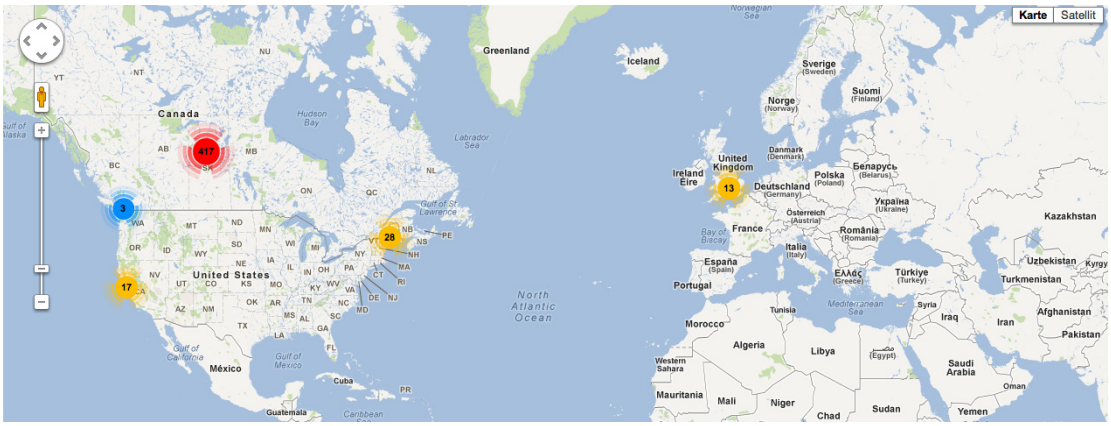


Abbildung 4.17: *ClusterTag*-Visualisierung der Herkunft der Tagger (*Muse McCord*)

[10] `http://www.mccord-museum.qc.ca/en/` [Zugriff am 20.11.2011]

Location	Tags
Australia	55
Toronto, Canada	208
London, UK	12
Maine, USA	23
Montreal, Canada	1
New York City, United States	2
Oxford, UK	1
San Francisco, USA	17
Vancouver, Canada	2
Winnipeg, Canada	3
Gesamtergebnis	324

Tabelle 4.12: Herkunft der Tagger (*Muse McCord*)

Region	Land	Anzahl
Angloamerikanische Region	Australien	55
	Kanada	214
	UK	13
	USA	42

Tabelle 4.13: Darstellung der Herkunft der Tagger als Region (Muse McCord)

4.7.2 Lokaler Einfluss

Viele dieser Institutionen weisen eine sehr große Anzahl an lokalen Tags auf, welche nicht über die Landesgrenzen hinaus vergeben wurden. In der Tabelle 4.14 werden die Institutionen mit lokalem geografischen Einfluss dargestellt. Die *New York Public Library* weist ein Global/Lokal Verhältnis von $\frac{13}{51}$= 25% auf und verfehlt somit die Grenze von $\frac{1}{3}$ globaler Tags nur sehr knapp. Für die restlichen Institutionen wurde der lokale Einfluss relativ deutlich nachgewiesen.

Institution	Global/Lokal
The Library of Congress (USA)	0 / 15
Powerhouse Museum Collection (Australien)	0 / 343
Smithsonian Institution (USA)	0 / 138
George Eastman House (USA)	2 / 409
National Maritime Museum (UK)	0 / 203
State Library of New South Wales collection (AUS)	6 / 1042

Australian War Memorial collection (Australien)	2 / 428
Imperial War Museum Collections (UK)	0 / 20
National Library NZ on The Commons (Neuseeland)	6 / 164
New York Public Library (USA)	13 / 51
National Galleries of Scotland Commons (UK)	107 / 691
nha.library (USA)	5 / 278
DC Public Library Commons (USA)	3 / 138
National Library of Wales (UK)	2 / 1192
LSE Library (UK)	10 / 53
Bergen Public Library (Norwegen)	15 / 552
Getty Research Institute (USA)	0 / 578
UA Archives | Upper Arlington History (USA)	0 / 1190
Galt Museum (Kanada)	16 / 621
JWA Commons (USA)	1 / 10
Texas State Archives (USA)	0 / 800
National Library of Scotland (Schottland)	4 / 2624
UW Digital Collections (USA)	0 / 866
The National Archives UK (USA)	0 / 15
Jewish Historical Society of the Upper Midwest (USA)	0 / 865

Tabelle 4.14: Institutionen mit lokalem Einfluss

Beispielsweise hat die *National Library of New Zeeland*[11] eine relativ große Anzahl an lokalen Taggern. Insgesamt wurden 170 Schlagwörter von 10 verschiedenen Usern vergeben. Davon stammen 164 aus Neuseeland, dies ergibt einen prozentualen Anteil von $\frac{164}{170} = 96,47\%$ lokalen Tags. Weiters gibt es 6 Tags die von einem User aus den USA vergeben wurden. Das Interesse an der Kulturinstitution ist lokal, da es sehr wenige Tags gibt die über die Landesgrenzen hinaus vergeben wurden. In Abbildung 4.18 sowie Tabelle 4.15 ist die Herkunft der Tagger sowohl geografisch als auch tabellarisch abgebildet.

[11] `http://www.natlib.govt.nz/` [Zugriff am 20.11.2011]

Zeilenbeschriftungen	Ergebnis
Christchurch, New Zealand	32
Washington, DC, USA	6
Wellington, New Zealand	132
Gesamtergebnis	170

Tabelle 4.15: Tabellarische Darstellung der Herkunft der Tagger (*National Library of New Zeeland*)

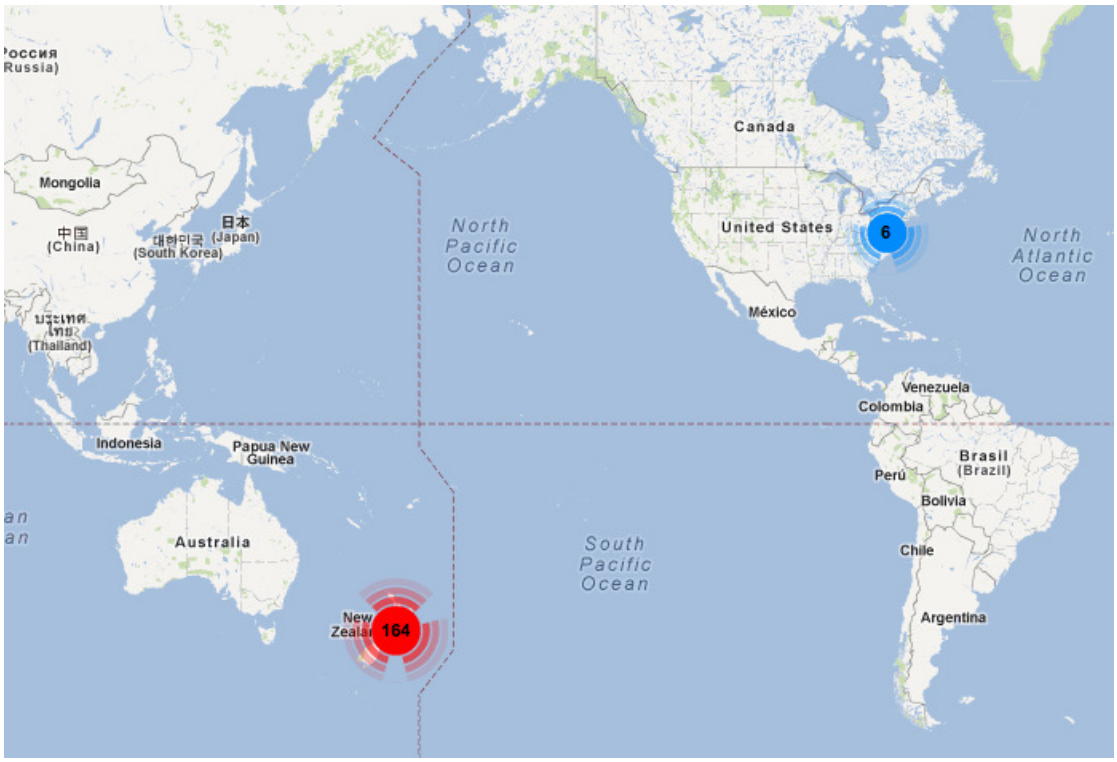


Abbildung 4.18: *ClusterTag*-Visualisierung der Herkunft der Tagger (*National Library of New Zeeland*)

4.7.3 Unbekannter Einfluss

Bei einigen Kultureinrichtungen konnte dieser Einfluss aufgrund fehlender *Location*-Daten, nicht ermittelt werden (Siehe Tabelle 4.16). Aufgrund dieser fehlenden Information kann keine Aussage über den geografischen Einfluss dieser Einrichtungen gemacht werden. In diesem Fall sind in der Datenbasis zu wenig Tags mit Angabe einer Tagger-Location vorhanden.

Institution	Global/Lokal
The Field Museum Library (USA)	0 / 6
Cornell University Library (USA)	1 / 2
Fylkesarkivet i Sogn og Fjordane (Norwegen)	2 / 0
NASA on The Commons (USA)	3 / 0
Nationaal Archief (Niederlande)	4 / 6
Center for Jewish History (USA)	0 / 1
Australian National Maritime Museum (AUS)	0 / 2
The U.S. National Archives (USA)	0 / 1

Tabelle 4.16: Einfluss der Institutionen nicht bestimmbar

4.7.4 Geografische Abhängigkeit

Um die geografische Abhängigkeit zwischen Usern und einer Institution zu ermitteln, wurde die Applikation *FlickrMaps* verwendet. In diesem Abschnitt soll geklärt werden, ob ein direkter Zusammenhang zwischen Aufnahmeort des Bildes und dem Tagger besteht. Dieser Zusammenhang kann beispielsweise dadurch entstehen, dass das Foto in unmittelbarer Umgebung des Taggers aufgenommen wurde bzw. von einem naheliegenden Museum veröffentlicht wurde.

Mithilfe der Anwendung *FlickrMaps* soll dieser Zusammenhang geklärt werden. Zunächst wurden 50 Bilder zufällig aus der Datenbasis ausgewählt, wobei für die Hälfte der Bilder ein Aufnahmeort vorhanden war. Bei der anderen Hälfte waren keine Ortsangaben über den Aufnahmeort bekannt.

Die Abbildung 4.19 zeigt die Auswertung eines Bildes durch *FlickrMaps*. Neben der grafische Auswertung erstellt die Anwendung auch eine *tabellarische* Statistik (siehe Abbildung 4.20). Diese Statistik wurde von allen 50 Bildern erstellt und manuell ausgewertet.

Abbildung 4.19: Geografische Darstellung eines Bildes mit *FlickrMaps*

CurrentData

Powerhouse Museum Collection
PhotoID: 2581845285
Location Sydney, Australia

Tags with Location

TagID	Tag	Author	Location
24762863-2581845285-791	nature	Meloearth.com	Northern California, USA
24762863-2581845285-15901	nsw	peckhamryecrow	London, UK
24762863-2581845285-2014647	katoombafalls	peckhamryecrow	London, UK
24762863-2581845285-1550559	geographile	marymactavish	San Lorenzo, California, USA
24762863-2581845285-25079597	dc:identifier=httpwwwpowerhousemuseumcomcollectiondatabaseirn30259	Powerhouse Museum Collection	Sydney, Australia
24762863-2581845285-114386	powerhousemuseum	Powerhouse Museum Collection	Sydney, Australia
24762863-2581845285-2354	waterfall	peckhamryecrow	London, UK
24762863-2581845285-9507	bluemountains	MargaretsFamily	Sydney, Australia
24762863-2581845285-8512	australia	peckhamryecrow	London, UK
24762863-2581845285-18719	newsouthwales	peckhamryecrow	London, UK
24762863-2581845285-8499312	xmlns:dc=httppurlorgdcelements11	Powerhouse Museum Collection	Sydney, Australia

Abbildung 4.20: Tabellarische Auswertung eines Bildes mit *FlickrMaps*

Die Ergebnisse dieser Auswertung sind in der Abbildung 4.21 dargestellt. Dieses Diagramm stellt die globalen und lokalen Tags der Bilder mit Angaben zum Aufnahmeort dar. Im Vergleich ist in der Abbildung 4.22 die Auswertung der Bilder ohne Aufnahmeort dargestellt. In der Tabelle 4.17 und 4.18 befinden sich die jeweiligen statistischen Auswertungen zu den Diagrammen. Bei allen 25 Bildern mit Ortsangaben zum Aufnahmeort wurde festgestellt, dass der lokale Taggeranteil bei jedem Bild höher war als der globale Taggeranteil. Das heißt, es gab mehr Tagger aus der direkten Umgeben als bei Bildern ohne geografische Angaben. Das Diagramm 4.22 bestätigt das Ergebnis aus der ersten Analyse. In diesem Diagramm sind alle Bilder ohne geografische Angaben dargestellt. Hier wechseln sich lokale und globale Tagger ab. Im Summe gibt es bei Bildern mit Locationangaben mehr Tagger. Darüber hinaus bewirkt die Ortsangabe ein unmittelbares Interesse der User aus der lokalen Umgebung.

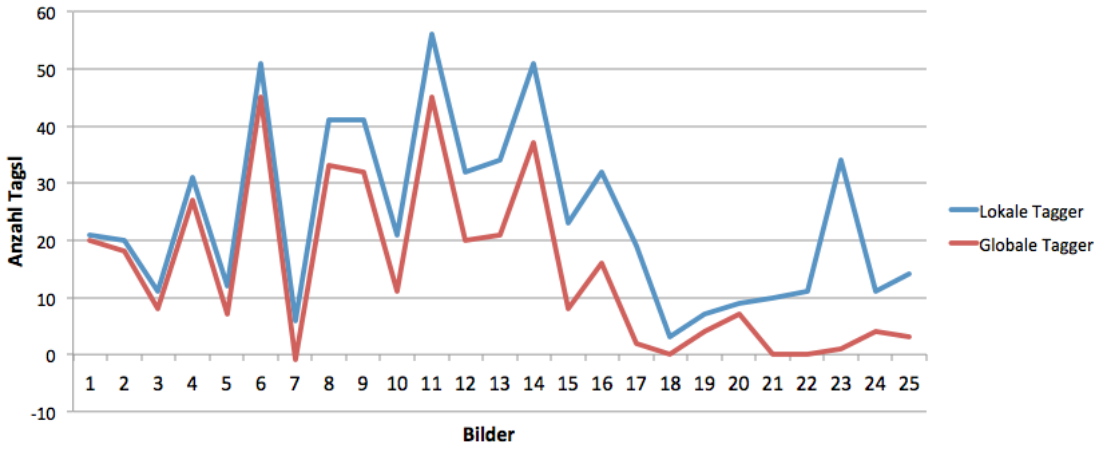


Abbildung 4.21: Auswertung der Bilder mit Aufnahmeort

Statistik	Lokale Tags	Globale Tags
Mittelwert	24,04	14,76
Varianz	238,45	209,52
Standardabweichung	15,44	14,47
Maximum	56	45
Minimum	3	0

Tabelle 4.17: Statistische Auswertung der 25 Bilder mit Aufnahmeort

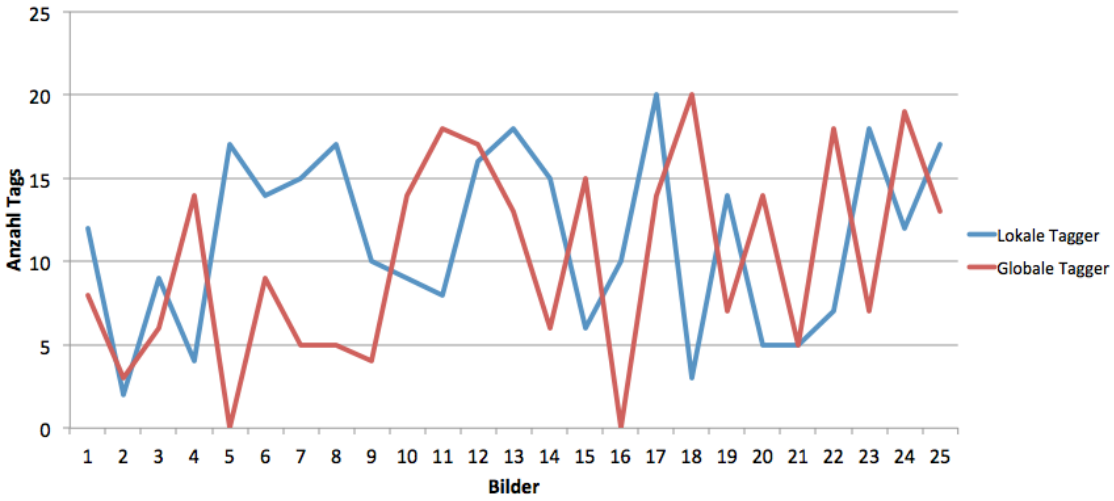


Abbildung 4.22: Auswertung der Bilder ohne Aufnahmeort

Statistik	Lokale Tags	Globale Tags
Mittelwert	11,32	10,16
Varianz	27,49	35,17
Standardabweichung	5,35	6,05
Maximum	20	20
Minimum	2	0

Tabelle 4.18: Statistische Auswertung der 25 Bilder ohne Aufnahmeort

Mithilfe des Zweistichproben-t-Tests wird geprüft ob sich die Grundgesamtheiten der Stichproben signifikant voneinander unterscheiden. Dabei werden die Mittelwerte μ_1 und μ_2 miteinander verglichen.

Folgende Hypothesen sollen dabei geprüft werde:

1. $H_0 : \mu_X - \mu_Y = 0$ (Nullhypothese)
2. $H_1 : \mu_X - \mu_Y \neq 0$ (Alternativhypothese)

Aus den Tabellen 4.17 (Stichprobe X) und 4.18 (Stichprobe Y) werden die Mittelwerte $\bar{X}$, $\bar{Y}$ sowie die zugehörigen Standardabweichungen S_X und S_Y über einen Durchschnittswert berechnet:

- $\bar{X} = \frac{24{,}04+14{,}76}{2}$ = 19,4
- $\bar{Y} = \frac{11{,}32+10{,}16}{2}$ = 10,74
- $S_X = \frac{15{,}44+14{,}47}{2}$ = 14,57
- $S_Y = \frac{5{,}35+6{,}05}{2}$ = 3,64

Diese Parameter dienen dann zur Berechnung der Teststatistik:

1. Teststatistik $t = \frac{\bar{X}-\bar{Y}-\omega_0}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \approx t_\nu$=2,88
2. Freiheitsgraden $\nu = \frac{\left(\frac{s_x^2}{n}+\frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1}+\frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}$ =9,85≈10

In der Tabelle für die t-Verteilung wird nun der Wert für $1-\alpha$ = 95% abgelesen: $t_{1-\alpha/2;\nu}$=2,262

Aufgrund des Ablehnungsbereiches $\{t|t > t_{1-\alpha/2;\nu}\}$ und $\{t|t < -t_{1-\alpha/2;\nu}\}$ liefert der t-Test folgendes Ergebnis:

- $t > t_{1-\alpha/2;\nu}$ –> 2,88 > 2,262 (WAHR)
- $t < -t_{1-\alpha/2;\nu}$ –> 2,88 < -2,262 (FALSCH)

Hier ist ersichtlich, dass der beobachtete t-Wert in den kritischen Bereich fällt, sodass die Nullhypothese verworfen wird und somit die Stichproben sich signifikant voneinander unterscheiden.

In diesem Kapitel konnte gezeigt werden, dass Ortsangaben in Summe die Anzahl der Schlagwörter zu einem Bild erhöhen. Vorallem bewirkt diese Angabe ein erhöhtes lokales Interesse der User. Vergleicht man die beiden Abbildungen 4.17 und 4.18 so ist ersichtlich, dass bei Abbildungen 4.17 der lokale Taggeranteil stets höher ist als der globale Taggeranteil. In Abbildung 4.18 dahingegen wechseln sich lokale und globale Tags ab.

KAPITEL 5

Zusammenfassung und Ausblick

5.1 Zusammenfassung

Zu Beginn gibt diese Arbeit einen Überblick über die aktuellen Erkenntnisse im Bereich Social Tagging und Folksonomien. Im ersten Kapitel wird gezeigt, dass gerade Social Tagging ein sehr effektiver Weg zur Datenorganisation ist. Trotz vieler Nachteile, weisen die gemeinschaftlich erzeugten Metadaten eine sehr hohe Qualität für die Klassifikation von Objekten auf. Vor allem bei sehr großen Datenmengen ist die Kategorisierung mittels Social Tagging, im Gegensatz zu traditionellen Kategorisierungsverfahren, sehr schnell und effektiv. Das Vergeben von Schlagwörtern benötigt keinerlei Vorkenntnisse und nimmt nur wenig Zeit und Mühe in Anspruch. Die Kategorisierung mittels Social Tagging basiert auf einen Konsens verschiedener Benutzer mit unterschiedlicher Expertise, wodurch andere Sichtweisen gefördert werden. Neben diesen Vorteilen der Datenorganisation, fördert Social Tagging auch die Zusammenarbeit von *Communities*.

Im zweiten Kapitel wird das Projekt *Flickr: The Commons* näher erläutert. Dieses Projekt dient dazu einer breiten Masse die öffentlichen Fotoarchive der Welt zugänglich zu machen und diese durch deren Wissen zu bereichern. In diesem Kapitel wird auch die Datenbasis sowie die zwei eigens entwickelten Tools *FlickrMaps* und *ClusterTags* vorgestellt und dient somit als Vorbereitung auf den empirischen Teil dieser Arbeit. Die Datenbasis stellt eine Stichprobe der Daten dar und beinhaltet die Institutionen, die Bilder sowie die zugehörigen Schlagwörter und Kommentare.

Im letzten Kapitel (praktischer Teil) wird die Datenbasis mithilfe verschiedenster Tools analysiert und visualisiert. Am Beginn werden die Institutionen untersucht und dabei die Populärsten errechnet. Die Beliebtheit bezieht sich hier auf das *User*-Interesse auf Basis der Anzahl an Tags und Kommentaren. Die eigentliche Berechnung erfolgt dann mit einem selbst definierten Beliebtheitsindex P. Dieser normiert die Anzahl der Tags, Kommentare, Tagger und Kommentatoren auf Basis der jeweiligen Maximalwerte und ist somit robust gegenüber *Powerusern*.

Nach den Institutionen wurden alle Bilder aus der Datenbasis untersucht. In die engere Auswahl gelangten Bilder, mit der höchsten Anzahl an Tags, Kommentaren sowie einem hohen Beliebtheitsindex P.

Anschließend wurden die User (Tagger) genauer betrachtet. Neben einer Häufigkeitsanalyse wurde auch die Herkunft der Tagger genauer untersucht. Die meisten User stammten aus den USA, Schottland und Australien. In diesem Abschnitt konnten auch die *Poweruser* nachgewiesen werden, welche für 80% aller Tags verantwortlich sind.

Danach wurden 34,000 Schlagwörter untersucht. Diese Analyse zeigte, dass sehr häufig Tags eher generelle Aspekte beschreiben und seltene Tags sich auf spezielle Eigenschaften eines Objektes beziehen. Anschließend wurde die Zeichenlänge und die Häufigkeit der Wörter analysiert und in Verbindung mit dem *Zipfschen Gesetz* erklärt.

Aus den Kommentaren wurden sehr viele Information gewonnen. Kommentare beschreiben keine Charakteristika einer Ressource sondern drücken viel mehr die mit dem Objekt verknüpfte Emotion aus. Bei der Analyse der Kommentare wurden neben der Häufigkeit und Zeichenlänge auch der Inhalt untersucht. Neben reinem Text wurden in 40% der Kommentare auch HTML Tags gefunden. Die untersuchten Kommentare weisen auch sehr häufig auf falsche bzw. fehlende Informationen hin. Hier konnte nachgewiesen werden, dass nahezu jeder vierte Kommentar die Wörter *false* oder *wrong* enthielt.

Zu guter Letzt wurde der geografische Einfluss der Kulturinstitutionen und der geografische Zusammenhang zwischen Aufnahmeort und Tagger ermittelt. Für 4 Kulturinstitutionen konnte hier ein globaler Einfluss anhand der Datenbasis errechnet werden. Ein globaler Einfluss herrscht dann wenn zumindest $\frac{1}{3}$ der Tags von Usern, welche nicht aus dem gleichen Land wie die Institution stammen, vergeben wurden. Für alle anderen Einrichtungen konnte entweder kein Einfluss ermittelt werden oder nur ein lokaler Einfluss.

Neben dem Einfluss der Kultureinrichtungen, wurde auch der geografische Zusammenhang zwischen Aufnahmeort eines Bildes und dem User analysiert. Mittels einer kleinen Stichprobe wurde ein direkter Zusammenhang zwischen Aufnahmeort und dem Tagger nachgewiesen. Ein Bild hat um 50% mehr lokale Tags, wenn bei Flickr die geografischen Daten zum Aufnahmeort vorhanden sind.

Diese Ergebnisse beziehen sich alle auf eine Datenbasis, welche nur einem geringen Teil der Gesamtdaten entspricht. Wegen dem Umfang der Analyse haben wir uns hier auf eine Auswertung von zufälligen Stichproben verlassen, die aber deutliche Tendenzen erkennen lassen.

Viele dieser Ergebnisse, können in der Praxis angewendet werden. Institutionen können, durch eine solche Analyse ihre Besuchergruppe ausfindig machen und einschränken. Danach können durch gezielte Werbung potenzielle Besucher sowohl für die Online- als auch Offlinepräsenz der Einrichtungen angesprochen werden. Beispielsweise können mit sehr einfachen Mitteln z. B. Herkunftsangaben zu den Bildern die Anzahl der Tagger auf *The Commons* gesteigert werden. Ein großer Ansturm auf *Flickr* wirkt sich mit einer hohen Wahrscheinlichkeit auch sehr positiv auf die Besucherzahlen der eigenen Homepage auf. Dies hat wiederum zur Folge, dass die realen Besuche in den Einrichtungen selbst auch gesteigert werden.

5.2 Ausblick

Die Menge der heterogenen Datensätze wird in Zukunft eher größer als kleiner werden, somit stellt Social Tagging ein äußerst aktuelles Thema in der Wissensorganisation dar. In naher Zukunft wird Social Tagging nicht traditionelle Klassifikationsverfahren ersetzen, sondern wird als Komplement zu den bestehenden Verfahren eingesetzt werden.

Diese Arbeit liefert eine sehr gute Basis über Folksonomien und Social Tagging. Der praktische Teil zeigt, mit welchen Methoden und Tools das Tagging-Verhalten der Benutzer erforscht werden kann. Um noch genauer und aussagekräftiger Ergebnisse zu liefern, sollte in erster Hinsicht die Datenbasis auf 30-40% der Echtdaten angepasst werden. Auch würde es Sinn machen, mehrere zeitlich voneinander unabhängige Snapshots der Daten zu erstellen. Dadurch ist es möglich eine Zeitspanne genauer zu durchleuchten und somit Veränderungen der Daten zu erkennen. Eventuell können die Veränderungen, auf Ereignisse der Gegenwart zurückgeführt werden.

Die Kalkulation des Beliebtheitsindex sollte bereits auf Datenbankebene erfolgen und schon am Beginn für alle Bilder und Institutionen berechnet werden. Mithilfe dieser Vorkalkulation können Veränderungen der Daten relativ schnell erkannt werden, da nur mehr die Indizes miteinander verglichen werden müssen. Darüber hinaus muss dieser Index noch mit einem Glättungskoeffizient versehen werden, welcher auch die Gesamtanzahl der Bilder pro Institution berücksichtigt.

Um die Schlagwörter noch genauer zu unterscheiden, wäre es auch sinnvoll das Vokabular der Schlagwörter zu untersuchen. Beispielsweise könnte mithilfe von *WordNet*[1] eine Trennung des Vokabulars in Nomen, Verben und Adjektive vorgenommen werden. Nach diesem Schritt könnte noch eine zusätzliche Verfeinerung der Nomen in Orts- bzw. Zeitangaben erfolgen.

Der geografische Einfluss in dieser Arbeit bezieht sich auf die Herkunftsländer der Tagger bzw. Einrichtungen. Um noch genauere Ergebnisse in diesem Abschnitt zu erzielen, sollte der Vergleichsparameter verkleinert werden. Beispielsweise könnten neben den Ländern, auch Bundesländer bzw. Städte als Basis für den geografischen Einfluss verwendet werden.

Die Erforschung der verwendeten Sprache auf *The Commons* sollte auch einige interessante Zusammenhänge liefern. Da sehr viele Institutionen aus dem angloamerikanischen Bereich stammen, ist Englisch die meist verwendete Sprache. Jedoch kann es sein, dass bei Bildern wo sich der Aufnahmeort von dem der Institution unterscheidet, die Sprache der Kommentare und Tags durch den Aufnahmeort und nicht durch die Institution geprägt wird. Diese Analyse könnte Beispielsweise durch eine *TranslatorAPI* z. B. Google Translator durchgeführt werden, diese ist jedoch kostenpflichtig weswegen wir in dieser Arbeit Abstand davon genommen haben. In naher Zukunft könnte es gut sein, dass es hier einen kostenfreien Dienst geben wird, welcher dann für eine tiefer gehende Sprachanalyse verwendet werden könnte.

Die Forschung an Folksonomien und Social Tagging steckt noch in den Kinderschuhen, jedoch lassen die bereits publizierten wissenschaftlichen Arbeiten darauf schließen, dass gemeinschaftliches Indexieren auch für die Zukunft ein sehr interessantes Thema bleibt.

[1] `http://wordnet.princeton.edu/` [Zugriff am 11.12.2011]

ANHANG A

Bilder

In Kapitel 4.5 wurden Schlagwörter in fünf Klassen abhängig von der Anzahl ihrer Tags unterteilt. Die folgenden Bilder gehören jeweils zu einer dieser fünf Klassen. Diese Auswahl der Bilder erfolgt zufällig beim Durchschauen der Fotoarchive.

Abbildung A.1: Bild aus Klasse 1 von National Media Museum - *Baby show* (Flickr, 2011*d*)

Abbildung A.2: Bild aus Klasse 2 von State Library of New South Wales collection - *Theatre Royal chorus, Tamarama Beach, ca. 1938 / by Sam Hood* (Flickr, 2011*g*)

Abbildung A.3: Bild aus Klasse 3 von Australian War Memorial - *A possum and a movie camera 1943* (Flickr, 2011*a*)

Abbildung A.4: Bild aus Klasse 4 von National Media Museum - *Wild Eye, the Souvenir King* (Flickr, 2011*e*)

Abbildung A.5: Bild aus Klasse 5 von National Media Museum - *The morning of Aug 8th 1918. German prisoners just taken, returning in charge of a single Australian past their own burning dugouts* - (Flickr, 2011*f*)

ANHANG B

Tabellen

Im folgenden Abschnitt werden einige ergänzende Tabelle bereitgestellt.

Die nachstehende Tabelle zeigt die 35 häufigsten Tags der *The Commons*-Kollektion. Die Tabelle dient als Ergänzung zu der in Kapitel 4.5 dargestellten Tabelle .

Nr	**Tags**	**Anzahl**
1	portrait	265
2	History	187
3	woman	174
4	The Great War	144
5	World War I	136
6	man	132
7	great war	122
8	ww1	108
9	war	107
10	Women	107
11	Florida	106
12	Canada	105
13	First World War	105
14	United States	103
15	wwi	101
16	Library of Congress	101
17	New York Public Library	101
18	Minnesota	101
19	world war one	101
20	Wales	101
21	George_Eastman_House	100

22	CCC	100
23	U.S. National Archives	100
24	Black-and-white prints (photographs)	100
25	Photojournalism	100
26	Powerhouse Museum	100
27	Photographic prints	100
28	War photography	100
29	Llyfrgell Genedlaethol Cymru	100
30	Propaganda	100
31	Upper Arlington	100
32	National Galleries of Scotland	100
33	Civilian Conservation Corps	100
34	Adolph B. Rice Studio	100
35	Texas State Parks	100

Tabelle B.1: 35 häufige Tags

Die nachstehende Tabelle beinhaltet alle Einzelnachweise zur Ermittelung des geografischen Einflusses (Kapitel 4.7) aller untersuchten Institutionen. Mithilfe der Spalte *Location* wird die Herkunft der Tags angegeben. Die Spalte *Einfluss* gibt an, abhängig von dem Standort der Institution, ob diese Tags lokal oder global vergeben wurden.

Institution	**Location**	**Einfluss**	**Tags**
The Library of Congress (USA)	Washingen USA	Lokal	7
	New York, NY, USA	Lokal	7
Powerhouse Museum (Australien)	Sydney	Lokal	343
Smithsonian Institution (USA)	Highland Park, NJ, United States	Lokal	8
	San Francisco, USA	Lokal	8
	Tuscaloosa, AL, USA	Lokal	15
	United States	Lokal	1
	Washington, DC, USA	Lokal	111
George Eastman House (USA)	London, United Kingdon	Global	2
	North Pole, AK, US	Lokal	2

	Rochester, NY, United States	Lokal	2
	Springdale,Arkansas , USA	Lokal	400
	Woodland Park, CO, USA	Lokal	3
National Media Museum (UK)	Athens, Greece	Global	28
	Bakersfield, USA	Global	28
	Brighton, United Kingdom	Lokal	44
	Canada	Global	2
	Dharga Town, Sri Lanka	Global	20
	Entroncamento, Portugal	Global	3
	Espoo, Finland	Global	1
	GA, USA	Global	1
	Greater Los Angeles, United States	Global	2
	LDN, UK	Lokal	23
	Leipzig, Deutschland	Global	13
	Manchester, UK	Lokal	1
	Maungaturoto, New Zealand	Global	1
	New York, US	Global	4
	Oakland, CA, United States	Global	7
	Porto, Portugal	Global	15
	Portsmouth, United Kingdom	Lokal	3
	Roma, Italy	Global	1
	San Francisco, USA	Global	1
	south shore, cape, boston, usa	Global	55
	Sydney, Australia	Global	2
	The Netherlands	Global	1
	usa	Global	8
	Williamsburg, Virginia, United States	Global	3
	Wyong, Australia	Global	16
National Maritime Museum (UK)	Birmingham,UK	Lokal	197
	London, UK	Lokal	197
	UK	Lokal	2

State Library (AUS)	Sydney	Lokal	818
	BCN, Iberio	Global	818
	Canberra, Australia	Lokal	16
	ciudad de mexico	Global	4
	Cronulla, NSW, AU, Australia	Lokal	2
	Melbourne, Australia	Lokal	4
	Sydney, Australia	Lokal	198
Muse McCord Museum (Kanada)	London,UK	Global	110
	Canada	Lokal	110
	City of Rexdale (part of Toronto), Canada	Lokal	414
	London, UK	Global	2
	Maine, USA	Global	12
	Montreal, Canada	Lokal	23
	New York City, United States	Global	1
	Oxford, UK	Global	2
	San Francisco, USA	Global	1
	Vancouver, Canada	Lokal	17
	Winnipeg, Canada	Lokal	3
Nationaal Archief (Niederlande)	Amsterdam, Netherlands	Lokal	6
	San Francisco, USA	Global	6
Australian War Memorial (Australien)	New York, USA	Global	2
	Brisbane, Australia	Lokal	2
	Cambridge, MA	Global	2
	Canada	Global	2
	Canberra, Australia	Global	10
	Denver Colorado suburbs, USA	Global	344
	Deutschland	Global	6
	Europe	Global	4
	France	Global	3
	San Francisco, USA	Global	1
	San Mateo, USA	Global	41
	Sturgeon Bay, WI, USA	Global	1
	Wellington, New Zealand	Global	1
	Yekaterinburg, Russia	Global	1
Imperial War Museum (UK)	London, UK	Lokal	20

National Library NZ (Neuseeland)	Wellington, NZ	Lokal	32
	Washington, DC, USA	Global	32
	Wellington, New Zealand	Lokal	6
New York Public Library (USA)	Norway	Global	3
	San Francisco, USA	Lokal	3
	Toronto, Canada	Global	51
	UK	Global	8
National Galleries of Scotland	Edinburgh, Scottland	Lokal	104
	Anaheim, California, USA	Global	104
	Australia	Global	8
	Belfort, France	Global	118
	Bloomington, Indiana, USA	Global	2
	Bray, Ireland	Global	2
	Canada	Global	190
	Currently London, UK	Global	148
	Edinburgh, Scotland, UK	Global	6
	Falun, Sweden	Global	106
	GA, USA	Global	1
	Glasgow, UK	Global	10
	Gold Bar, WA, USA	Global	2
	Italy	Global	1
	Kent, Ohio, USA	Global	1
	Koblenz, Germany	Global	3
	Moncton, NB, Canada	Global	2
	New York City, United States	Global	1
	Northern town, England, UK	Global	2
	Rosyth, Scotland	Lokal	1
	San Francisco, CA, USA	Global	3
	San Francisco, USA	Global	7
	Seattle, United States of America	Global	24
	St Helens, uk	Global	11
	Sydney, Australia	Global	6
	Torino, Italy	Global	9
	Toronto, Canada	Global	1

nha.library (USA)	Washington, USA	Lokal	8
	Maine, USA	Lokal	8
	Nantucket, USA	Lokal	4
	New York City, United States	Lokal	248
	New Zealand	Global	4
	Oakland, CA	Global	1
	Rishon LeZion, Israel	Global	2
	San Francisco, USA	Lokal	2
	Springfield, MO, USA	Lokal	8
	Winchendon, Massachusetts, United States	Lokal	1
Swedish National Heritage Board	Paris, France	Global	68
	New York City, United States	Global	68
	NL	Global	8
	Ottawa, ON, Canada	Global	2
	Sweden	Lokal	1
	Visby, Sweden	Lokal	208
DC Public Library Commons (USA)	Washington DC	Lokal	103
	New Orleans	Lokal	103
	San Francisco, USA	Lokal	5
	Stockholm, Sweden	Global	7
	Sydney, Australia	Global	1
	Washington DC, USA	Lokal	2
	Washington, D.C., USA	Lokal	5
	Washington, DC, US	Lokal	3
	Washington, DC, USA	Lokal	1
The Field Museum Library (USA)	USA	Lokal	6
National Library of Wales (UK)	NY, USA	Global	2
	Wales/Cymru	Lokal	2
LSE Library (UK)	London,UK	Lokal	53
	Paris, France	Global	5
	The Hague, Netherlands	Global	5
	Vancouver, Canada	Global	1
Bergen Public Library (Norwegen)	Oslo, Norway	Lokal	552
	Boston, United States	Global	12
	San Francisco, USA	Global	3
Getty Research Institute (USA)	CA, USA	Lokal	562
	San Francisco, USA	Lokal	16

Upper Arlington History (USA)	USA New York	Lokal Lokal	1186 4
Galt Museum (Kanada)	Toronto, USA New York City, United States Visby, Sweden	Lokal Global Global	621 6 10
JWA Commons (USA)	New York, USA New York City, United States Sydney, Australia	Lokal Lokal Global	1 1 9
Texas State Archives (USA)	Texas, USA	Lokal	800
National Library of Scotland	Paris, UK Scotland	Global Lokal	4 4
UW Digital Collections (USA)	USA	Lokal	866
The U.S. National Archives (USA)	USA	Lokal	1
The National Archives UK (USA)	USA	Lokal	15
Jewish Historical Society (USA)	USA Minneapolis, MN, USA San Francisco, USA	Lokal Lokal Lokal	4 4 858
Center for Jewish History (USA)	NY, USA	Lokal	1
AUS Maritime Museum (USA)	USA	Lokal	2
Cornell University Library (USA)	USA Tallinn, Estonia	Lokal Global	2 2
Fylkesarkivet (Norwegen)	Stockholm, Sweden	Global	2
Reykjav'k Museum (Island)	San Francisco, USA	Global	27
NASA on The Commons (USA)	Vienna, Austria	Global	3

Tabelle B.2: Geografischer Einfluss aller Institutionen

ANHANG C

Quellcode

Der nachfolgende Code stellt eine Implementierung des in Kapitel 3.2.2 dargestellten *Pseudocodes* in Javascript dar.

```
private function calculateClusters():Array  {
    var positionedMarkers:Dictionary = new Dictionary();
    var positionedMarker:PositionedMarker;

    for each (positionedMarker in _positionedMarkers) {
      positionedMarkers[positionedMarker.id] = positionedMarker;
     }

     var compareDistance:Number = Math.pow(_clusterRadius
     * Math.pow(2, 21 - _zoom), 2);

    var clusters:Array = [];
    var cluster:Array;
    var p1:Point;
    var p2:Point;
    var x:int;
    var y:int;
    var compareMarker:PositionedMarker;
    for each (positionedMarker in positionedMarkers) {
      if (positionedMarker == null) {
          continue;
      }

      positionedMarkers[positionedMarker.id] = null;
      cluster = [positionedMarker.marker];
```

```
      for each (compareMarker in positionedMarkers) {
         if (compareMarker == null) {
                  continue;
               }
               p1 = positionedMarker.point;
               p2 = compareMarker.point;
               x = p1.x - p2.x;
               y = p1.y - p2.y;
               if (x * x + y * y < compareDistance) {
                  cluster.push(compareMarker.marker);
                  positionedMarkers[compareMarker.id] = null;
               }
            }
            clusters.push(cluster);
         }
         return clusters;
      }
   }
}
```

Literaturverzeichnis

Albrecht, C. (2006), Folksonomy, Master's thesis, TU Wien, Wien. Diplomarbeit am Institut für Gestaltungs- und Wirkungsforschung, TU Wien. (Cited on page 21.)

Ames, M. and Naaman, M. (2007), Why we tag: motivations for annotation in mobile and online media, *in* 'CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York, NY, USA, pp. 971–980.
URL: *http://portal.acm.org/citation.cfm?id=1240624.1240772* (Cited on pages xi and 12.)

Begelman, G., Keller, P. and Smadja, F. (2006), Automated tag clustering: Improving search and exploration in the tag space, *in* 'Proceedings of the WWW Collaborative Web Tagging Workshop', Edinburgh, Scotland. (Cited on page 13.)

Cattuto, C., Baldassarri, A., Servedio, V. D. P. and Loreto, V. (2007), Vocabulary growth in collaborative tagging systems.
URL: *http://www.citebase.org/abstract?id=oai:arXiv.org:0704.3316* (Cited on page 6.)

Cha, M., Mislove, A. and Gummadi, K. P. (2009), A measurement-driven analysis of information propagation in the flickr social network, *in* 'Proceedings of the 18th international conference on World wide web', WWW '09, ACM, New York, NY, USA, pp. 721–730.
URL: *http://doi.acm.org/10.1145/1526709.1526806* (Cited on page 16.)

Choi, Y., Kim, Y. and Myaeng, S.-H. (2009), Domain-specific sentiment analysis using contextual feature generation, *in* 'Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion', TSA '09, ACM, New York, NY, USA, pp. 37–44.
URL: `http://doi.acm.org/10.1145/1651461.1651469` (Cited on page 57.)

Cialdini, R. and Truus, S. (2005), *Influence: science and practice*, Pegasus.
URL: *http://books.google.at/books?id=426APQAACAAJ* (Cited on page 19.)

Dalamagas, T., Farmakakis, T., Maragkakis, M. and Hatzigeorgiou, A. G. (2010), 'Freepub: Collecting and organizing scientific material using mindmaps', *CoRR* **abs/1012.1623**.
URL: *http://arxiv.org/abs/1012.1623* (Cited on page 28.)

Delicious (2012), 'Delicious', `http://www.delicious.com/`. [Zugriff: 19.01.2012]. (Cited on page 15.)

Ding, Y., Du, Y., Hu, Y., Liu, Z., Wang, L., Ross, K. and Ghose, A. (2011), Broadcast yourself: understanding youtube uploaders, *in* 'Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference', IMC '11, ACM, New York, NY, USA, pp. 361–370.
URL: *http://doi.acm.org/10.1145/2068816.2068850* (Cited on page 17.)

Ebersbach, A., Glaser, M. and Heigl, R. (2008), *Social Web*, Vol. 3065 of *UTB : Medien- und Kommunikationswissenschaft, Soziologie, Pädagogik, Informatik*, UVK Verl.-Ges., Konstanz.
URL: *http://www.gbv.de/dms/ilmenau/toc/555663728.PDF* (Cited on page 22.)

Flickr (2011*a*), 'Australian War Memorial – A possum and a movie camera 1943', `http://www.flickr.com/photos/australian-war-memorial/3527157206/`. [Zugriff: 20.01.2012]. (Cited on pages xii and 74.)

Flickr (2011*b*), 'Flickr – Help / FAQ / Tags'.
URL: *http://www.flickr.com/help/tags/* (Cited on page 10.)

Flickr (2011*c*), 'Flickr: The Commons', `http://www.flickr.com/commons`. [Zugriff: 01.11.2011]. (Cited on pages v, vii, xi, 16, and 35.)

Flickr (2011*d*), 'National Media Museum – Baby show; Reuben Saidman (1906-1967); Digital positive from glass negative', `http://www.flickr.com/photos/nationalmediamuseum/3588770019/in/photostream/`. [Zugriff: 20.01.2012]. (Cited on pages xii and 73.)

Flickr (2011*e*), 'National Media Museum – Wild Eye, the Souvenir King', `http://www.flickr.com/photos/nationalmediamuseum/3007981618/`. [Zugriff: 20.01.2012]. (Cited on pages xii and 75.)

Flickr (2011*f*), 'National Media Museum - The morning of Aug 8th 1918. German prisoners just taken, returning in charge of a single Australian past their own burning dugouts', `http://www.flickr.com/photos/nationalmediamuseum/3007145371/`. [Zugriff: 20.01.2012]. (Cited on pages xii and 75.)

Flickr (2011*g*), 'State Library of New South Wales collection – Theatre Royal chorus, Tamarama Beach, ca. 1938 / by Sam Hood', `http://www.flickr.com/photos/statelibraryofnsw/3072233763/`. [Zugriff: 20.01.2012]. (Cited on pages xii and 74.)

Ganter, B. and Wille, R. (1999), *Formal Concept Analysis: Mathematical Foundations*, Springer, Berlin/Heidelberg. (Cited on page 8.)

GMW (2008), *Gesellschaft für Medien in der Wissenschaft e.V – Good Tags, Bad Tags : Social Tagging in der Wissensorganisation*, Medien in der Wissenschaft, Bd. 47, Waxmann, Münster.
URL: *http://www.worldcat.org/isbn/9783830920397* (Cited on pages 5 and 7.)

Golder, S. and Huberman, B. A. (2005), 'The Structure of Collaborative Tagging Systems'.
URL: *http://arxiv.org/abs/cs.DL/0508082* (Cited on pages 9, 15, 24, and 25.)

Golub, K., Moon, J., Tudhope, D., Jones, C., Matthews, B., PuzoD, B. and Lykke Nielsen, M. (2009), EnTag: enhancing social tagging for discovery, *in* 'Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries', JCDL '09, ACM, New York, NY, USA, pp. 163–172.
URL: *http://doi.acm.org/10.1145/1555400.1555427* (Cited on page 21.)

Guy, M. and Tonkin, E. (2006), 'Folksonomies: Tidying up Tags?', *D-Lib Magazine* **12**(1).
URL: *http://dx.doi.org/10.1045/january2006-guy* (Cited on page 24.)

Hammond, T., Hannay, T., Lund, B. and Scott, J. (2005), 'Social Bookmarking Tools (I): A General Review', *D-Lib Magazine* **11**(4).
URL: *http://www.dlib.org/dlib/april05/hammond/04hammond.html* (Cited on page 15.)

Hotho, A., Benz, D., Eisterlehner, F., Jäschke, R., Krause, B., Schmitz, C. and Stumme, G. (2010), 'Publikationsmanagement mit BibSonomy – ein Social-Bookmarking-System für Wissenschaftler', *HMD Praxis der Wirtschaftsinformatik* **Heft 271**(271), 47–58. (Cited on page 15.)

Kammergruber, Walter Christian, M. L. (2009), Tagging als soziales Bindeglied für Communities, *in* K. Meissner and M. Engelien, eds, 'Virtuelle Organisation und Neue Medien 2009 (GeNeMe 2009)', Dresden, pp. 35–43.
URL: *http://w3-mmt.inf.tu-dresden.de/geneme/* (Cited on page 6.)

Köhler, R. (2005), *Quantitative Linguistik*, Handbücher zur Sprach- und Kommunikationswissenschaft / mitbegr. von Gerold Ungeheuer. Hrsg. von Hugo Steger ... ; 27, de Gruyter. (Cited on page 53.)

Lackes, S. (2009), 'Web 2.0 - Gabler Wirtschaftslexikon', `http://wirtschaftslexikon.gabler.de/Archiv/80667/web-2-0-v7.html`. [Zugriff: 01.01.2012]. (Cited on pages 1 and 6.)

Leimstoll, U. and Stormer, H. (2007), Collaborative recommender systems for online shops, *in* J. A. Hoxmeier and S. Hayne, eds, 'Americas Conference on Information Systems', Association for Information Systems, p. 156.
URL: *http://aisel.aisnet.org/amcis2007/156* (Cited on page 7.)

Lewis, D. (2006), 'What is web 2.0?', *ACM Crossroads* **13**(1), 3.
URL: *http://dblp.uni-trier.de/db/journals/crossroads/crossroads13.html* (Cited on page 6.)

Marlow, C., Naaman, M., Boyd, D. and Davis, M. (2006), Ht06, tagging paper, taxonomy, flickr, academic article, to read, *in* 'Proceedings of the seventeenth conference on Hypertext and hypermedia', HYPERTEXT '06, ACM, New York, NY, USA, pp. 31–40.
URL: *http://doi.acm.org/10.1145/1149941.1149949* (Cited on pages xi, 11, 12, 19, 20, and 21.)

Müller, S., Miller, G. and Fels, S. (2010), Using temporal video annotation as a navigational aid for video browsing, *in* 'Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology', UIST '10, ACM, New York, NY, USA, pp. 445–446.
URL: *http://doi.acm.org/10.1145/1866218.1866263* (Cited on page 18.)

Norton, Q. (2006), ''I want to build something that grows' | Media | The Guardian'.
URL: *http://www.guardian.co.uk/media/2006/jan/26/newmedia.technology1* (Cited on page 6.)

O'Reilly, T. (2006), 'Web 2.0: Compact Definition? - O'Reilly Radar'.
URL: *http://radar.oreilly.com/archives/2005/10/web-20-compact-definition.html* (Cited on page 5.)

Peters, I. and Stock, W. G. (2008), 'Folksonomies in Wissensrepräsentation und Information Retrieval', *Information – Wissenschaft und Praxis* **59**(2), 77–90.
URL: `http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/56/1204547947stock212_h.htm` (Cited on pages 22 and 23.)

Schmitz, C., Hotho, A., Jäschke, R. and Stumme, G. (2006), Mining association rules in folksonomies, *in* 'Data Science and Classification. Proceedings of the 10th IFCS Conf.', Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg, pp. 261–270. (Cited on pages 7 and 8.)

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M. and Riedl, J. (2006), tagging, communities, vocabulary, evolution, *in* 'Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work', CSCW '06, ACM, New York, NY, USA, pp. 181–190.
URL: *http://doi.acm.org/10.1145/1180875.1180904* (Cited on page 5.)

Shirky, C. (2003), 'Power laws, weblogs, and inequality'.
URL: *http://shirky.com/writings/powerlaw_weblog.html* (Cited on page 26.)

Smith, G. (2008), *Tagging : people-powered metadata for the social web*, New Riders, Berkeley, Calif.
URL: *http://scans.hebis.de/HEBCGI/show.pl?19488071_toc.pdf* (Cited on pages 11, 13, and 14.)

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D. and Zinkham, H. (2008), For the common good – the library of congress – flickr pilot project, Technical report, Library of Congress.
URL: *http://www.loc.gov/rr/print/flickr_report_final_summary.pdf* (Cited on page 34.)

Tanaka, J. W. and Taylor, M. (1991), 'Object categories and expertise: Is the basic level in the eye of the beholder?', *Cognitive Psychology* **23**(3), 457–482.
URL: *http://dx.doi.org.arugula.cc.columbia.edu:2048/10.1016/0010-0285(91)90016-H* (Cited on page 24.)

Vander Wal, T. (2005), 'Explaining and showing broad and narrow folksonomies'. (Cited on pages xi, 22, 23, and 27.)

Voss, J. (2006), 'Collaborative thesaurus tagging the Wikipedia way'.
URL: *http://arxiv.org/abs/cs.IR/0604036* (Cited on page 27.)

Wal, T. V. (2007), 'Folksonomy coinage and definition'.
URL: *http://vanderwal.net/folksonomy.html* (Cited on page 21.)

Weghuber, S. (2009), Implementierung und Anwendung semantisch angereicherter Folksonomien in Medienarchiven, Master's thesis, Digitale Medien; FH Oberösterreich – Fakultät für Informatik, Kommunikation und Medien, Hagenberg, Austria. (Cited on page 14.)

Wikipedia (2011*a*), 'Autochrom Verfahren', `http://de.wikipedia.org/wiki/Autochromverfahren`. [Zugriff: 16.12.2011]. (Cited on page 47.)

Wikipedia (2011*b*), 'Folksonomien', `http://de.wikipedia.org/wiki/Folksonomy`. [Zugriff: 11.11.2011]. (Cited on page 15.)

Wikipedia (2011*c*), 'Hashtags', `http://de.wikipedia.org/wiki/Twitter`. [Zugriff: 01.12.2011]. (Cited on page 11.)

Wikipedia (2011*d*), 'K-Means Algorithmus', `http://de.wikipedia.org/wiki/K-Means-Algorithmus`. [Zugriff: 17.12.2011]. (Cited on page 41.)

Wikipedia (2011*e*), 'Mindmaps', `http://de.wikipedia.org/wiki/Mind-Map`. [Zugriff: 5.12.2011]. (Cited on page 28.)

Wikipedia (2011*f*), 'Paretoprinzip', `http://de.wikipedia.org/wiki/Paretoprinzip`. [Zugriff: 11.11.2011]. (Cited on page 26.)

Wikipedia (2011*g*), 'Portrait', `http://de.wikipedia.org/wiki/Portr%C3%A4t_(Kunst)`. [Zugriff: 03.11.2011]. (Cited on page 52.)

Wikipedia (2011*h*), 'Tagcloud', `http://de.wikipedia.org/wiki/Schlagwortwolke`. [Zugriff: 11.10.2011]. (Cited on page 27.)

Wikipedia (2011*i*), 'Zipfsches-Gesetz', `http://de.wikipedia.org/wiki/Falsches_Zipfsches_Gesetz`. [Zugriff: 25.10.2011]. (Cited on pages vi, viii, and 53.)

Zhang, Z.-K. and Liu, C. (2010), 'Hypergraph model of social tagging networks'.
URL: *http://arxiv.org/abs/1003.1931* (Cited on pages xi and 8.)

Zhu, H. and Wu, H. (2009), Sloppy tags and metacrap? quality of user contributed tags in collaborative social tagging systems, *in* R. C. Nickerson and R. Sharda, eds, 'Americas Conference on Information Systems', Association for Information Systems, p. 438.
URL: *http://aisel.aisnet.org/amcis2009/438* (Cited on page 1.)