Unlocking Knowledge through Corporate Tags

Céline Van Damme

Proefschrift neergelegd tot het behalen van de graad

Doctor in de Toegepaste Economische Wetenschappen: Handelsingenieur

Promotor: Prof. Dr. E. Vandijck
Samenvatting

Tegenwoordig dienen bedrijven zich te differentiëren van elkaar door middel van kennis. De digitalisering van de wereld heeft echter geleid tot een overload aan informatie waardoor het voor een bedrijf moeilijker wordt op de hoogte te blijven van alle kennis die verscholen zit in haar collectief geheugen of organizational memory. In deze doctoraatsthesis werd onderzocht hoe tags gecreëerd door werknemers in een onderneming, genaamd corporate tags, kunnen worden geanalyseerd om kennis in de organizational memory terug te vinden.

Tags zijn zelfgekozen trefwoorden die gebruikt worden om bronnen (documenten, websites, foto’s, videos) te beschrijven. Ze worden gecreëerd om bronnen achteraf gemakkelijker terug te vinden. De aggregatie van tags vormt een folksonomy zoals gedefinieerd door Thomas Vander Wal. De.licio.us, Flickr en YouTube zijn een aantal gekende voorbeelden van websites die een tagging mechanisme geïmplementeerd hebben.

Onderzoekers binnen het domein van het World Wide Web (WWW) beschouwen tags als een belangrijk fundament voor het bouwen van ontologieën. Ontologieën zijn één van de technologieën die ervoor zorgen dat informatie op het WWW gemakkelijker zal teruggevonden worden en het informatie overload probleem op het WWW zal gereduceerd worden.

Aangezien ondernemingen eveneens kampen met een overload aan informatie werd onderzocht hoe corporate tags een meerwaarde kunnen bieden aan ondernemingen, namelijk hoe ze kennis die verscholen zit in de organizational memory kunnen in kaart brengen en hoe dit de bedrijfsstrategie kan ondersteunen. Hiervoor zijn 4 onderzoeksvragen opgesteld. Om een antwoord te kunnen formuleren op deze onderzoeksvragen werd gebruik gemaakt van een case-study en dataset van een grote Europese distributieonderneming. Deze onderneming laat al meer dan 20 jaar alle berichten gecreeëerd en verstuurd door werknemers taggen.

In eerste instantie werd onderzocht hoe tags van hoge kwaliteit kunnen gevonden worden in een folksonomy omdat het principe van tagging een controle op de gekozen tags uitsluit: iedereen is vrij zijn eigen tags te kiezen. Daarna is de kwaliteit van de corporate tags in de dataset van nabij bestudeerd. Hieruit bleek dat de kwaliteit van de corporate tags, na het opkuisen van de tags, goed was, maar zeker nog kon verbeterd worden door het aanreiken van suggesties.

Daarnaast werd aangetoond dat door middel van een analyse van corporate tags experten binnen de onderneming kunnen gelocaliseerd worden, en bijgevolg de kennis verscholen in de hoofden van de werknemers kan worden vrijgemaakt. Tag dashboards, opgesteld op basis van de bedrijfsstrategie, kunnen gebruikt worden voor het in kaart brengen van kennis die verborgen zit in bedrijfsdocumenten. Daarenboven werd aangetoond hoe de tag dashboards kunnen dienen als een performance measurement tool binnen het IT-governance model.

Op basis van de resultaten van dit onderzoek zijn een aantal basisvereisten en aanbevelingen geformuleerd voor ondernemingen die kennis uit de organizational memory willen halen door middel van corporate tags.
# Contents

Acknowledgement ............................................. 9

I Positioning of the research ................................ 11

1 Introduction ............................................... 13
   1.1 Structure ............................................. 13
   1.2 Overview of literature ............................... 13
   1.3 Research questions ................................... 18
   1.4 Case-study research ................................... 19
   1.5 Contributions ......................................... 20
   1.6 How the thesis is organized ........................... 21
   1.7 Publications ........................................... 23

II Background ................................................. 25

2 Organizational knowledge ................................. 27
   2.1 Introduction .......................................... 27
   2.2 The difference between data, information and knowledge .... 28
   2.3 Tacit and explicit knowledge .......................... 30
   2.4 Organizational memory ................................ 31
      2.4.1 Transactional and individual memories .......... 33
   2.5 Information overload in companies and on the world wide web .... 35
   2.6 Ontologies ............................................. 38
      2.6.1 Goals to use ontologies ......................... 39
      2.6.2 Languages for creating ontologies .............. 40
      2.6.3 Methodologies for creating ontologies .......... 40
      2.6.4 Problems related to ontologies .................. 41
   2.7 Turning folksonomies into ontologies ................ 42
   2.8 Conclusion ............................................ 44

3 Folksonomies and tagging ................................ 45
   3.1 Introduction .......................................... 45
   3.2 Classification ......................................... 45
      3.2.1 Object-based classification ....................... 46
      3.2.2 Subject-based classification ...................... 47
   3.3 Categorization: tagging and folksonomies ............... 51
      3.3.1 Folksonomy versus taxonomy ..................... 54
      3.3.2 Broad versus narrow folksonomy ................. 57
### III Research

#### 4 Quality of tags in broad and narrow folksonomies

- **4.1 Introduction** .................................................. 77
- **4.2 Quality of information and tags** .......................... 79
- **4.3 Parameters for tag quality** .................................. 80
  - **4.3.1 Implictness-explicitness** ................................. 80
  - **4.3.2 Type of folksonomy** ....................................... 81
- **4.4 High-quality tags** ............................................. 82
- **4.5 Related work** ................................................. 83
  - **4.5.1 How our research differs from existing research** .... 86
- **4.6 Algorithms to find high-quality tags in broad folksonomies** ........................................... 86
  - **4.6.1 Tag quality algorithm 1: high-frequency tags** ....... 87
  - **4.6.2 Tag quality algorithm 2: tag agreement** ............... 88
  - **4.6.3 Tag quality algorithm 3: TF-IRF** .......................... 89
  - **4.6.4 Choose an appropriate value for \( n \)** ................. 91
- **4.6.5 Dataset** ..................................................... 91
- **4.6.6 Results** ..................................................... 93
- **4.6.7 Preliminary evaluation** ..................................... 94
- **4.6.8 Limitations of the research** ............................... 98
- **4.6.9 Conclusion** .................................................. 98

#### 5 Description of the case study and dataset

- **5.1 Introduction** ................................................... 101
- **5.2 Tagging in the company** ..................................... 102
- **5.3 Data elements in the dataset** ................................ 103
- **5.4 Tools** .......................................................... 106
- **5.5 Problems dataset** ............................................. 108
  - **5.5.1 Loading text files into the database** .................... 108
  - **5.5.2 Size of the dataset** ....................................... 108
  - **5.5.3 Problems with structure of the dataset** ............... 109
  - **5.5.4 Manual and semi-automatic messages** ................. 109
- **5.6 Descriptive statistics** ........................................ 111
  - **5.6.1 Messages** .................................................. 111
5.6.2 Tags ................................................................. 112
5.7 Conclusion .......................................................... 113

6 CorTagCleaning approach to improve the explicit quality of corporate tags 115
6.1 Introduction ......................................................... 115
   6.1.1 Tag rules in the company ................................. 116
   6.1.2 Tag dataset problems ................................. 117
6.2 Related work ...................................................... 119
   6.2.1 Stemming and lemmatization ............................ 119
   6.2.2 Levenshtein edit distance .............................. 120
6.3 CorTagCleaning approach ....................................... 122
   6.3.1 Step 1: preparatory steps ............................... 123
   6.3.2 Step 2: removal of stop words ....................... 125
   6.3.3 Step 3: adjustment of the c/k words ............... 125
   6.3.4 Step 4: Levenshtein edit distance ............... 126
   6.3.5 Step 5: transform tags to their lemma .......... 126
   6.3.6 Step 6: add tags we removed in the first step .. 128
   6.3.7 Step 7: remove duplicate tags .................. 128
   6.3.8 Overview of results ..................................... 128
6.4 Limitations CorTagCleaning approach ........................ 130
6.5 Conclusion ......................................................... 131

7 Implicit quality of corporate tags 133
7.1 Introduction ......................................................... 133
7.2 Quality of corporate tags ...................................... 133
7.3 Related work ...................................................... 135
   7.3.1 Automatically generated keywords ................ 136
   7.3.2 Tagging compared to automatically generated keywords ................................................. 136
7.4 Methodology ...................................................... 137
   7.4.1 Approach to select high-quality tags automatically in a narrow folksonomy .................... 137
   7.4.2 Methodology to measure the quality of the corporate tags manually ............................. 138
7.5 Results ............................................................. 140
   7.5.1 Definitions and notation ................................ 141
   7.5.2 Question 1: how good is the implicit quality of corporate tags? .................................. 143
   7.5.3 Question 2: can a mining of a textual resource help to find high-quality tags? .................. 150
   7.5.4 Question 3: can the suggested tags (type A) help an employee to select more high-quality tags? .................................................. 150
   7.5.5 Question 4: how do employees experience the difficulty of selecting high-quality tags? .... 152
7.6 Limitations .......................................................... 153
7.7 Conclusion .......................................................... 154

8 Unlocking tacit knowledge through tags: expert finding 157
8.1 Introduction .......................................................... 157
8.2 Motivation for expert finding .................................. 159
8.3 Expert recommender systems .................................. 160
8.4 Related work .......................................................... 160
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4.1 The use of tags and experts</td>
<td>160</td>
</tr>
<tr>
<td>8.4.2 Experts and social network analysis based on tags</td>
<td>163</td>
</tr>
<tr>
<td>8.5 Methodology</td>
<td>165</td>
</tr>
<tr>
<td>8.5.1 Description of the techniques</td>
<td>165</td>
</tr>
<tr>
<td>8.5.2 Dataset</td>
<td>168</td>
</tr>
<tr>
<td>8.5.3 Interviews</td>
<td>169</td>
</tr>
<tr>
<td>8.6 Interview results analysis</td>
<td>170</td>
</tr>
<tr>
<td>8.6.1 Question 1: which technique reflects the expertise of an employee best?</td>
<td>170</td>
</tr>
<tr>
<td>8.6.2 Question 2: does tagging of corporate resources well reflect employees’ knowledge?</td>
<td>171</td>
</tr>
<tr>
<td>8.6.3 Question 3: do employees who use the same tags also have the same knowledge field?</td>
<td>172</td>
</tr>
<tr>
<td>8.7 Limitations</td>
<td>175</td>
</tr>
<tr>
<td>8.8 Conclusion</td>
<td>175</td>
</tr>
<tr>
<td>9 Enabling business value by unlocking explicit knowledge through tags</td>
<td>177</td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>177</td>
</tr>
<tr>
<td>9.2 Information overload and the impact on the organization</td>
<td>179</td>
</tr>
<tr>
<td>9.3 Evolution of approaches to information management and decision making</td>
<td>181</td>
</tr>
<tr>
<td>9.4 Dashboards and scorecards</td>
<td>182</td>
</tr>
<tr>
<td>9.5 Tag dashboard approach</td>
<td>184</td>
</tr>
<tr>
<td>9.5.1 Related work</td>
<td>185</td>
</tr>
<tr>
<td>9.5.2 Tag Dashboards: description and motivation</td>
<td>186</td>
</tr>
<tr>
<td>9.5.3 Example and interpretation of a tag dashboard</td>
<td>189</td>
</tr>
<tr>
<td>9.5.4 Implementation of tag dashboards as a performance measurement tool in the company</td>
<td>190</td>
</tr>
<tr>
<td>9.5.5 Step-by-step implementation plan</td>
<td>191</td>
</tr>
<tr>
<td>9.5.6 Methodology</td>
<td>197</td>
</tr>
<tr>
<td>9.6 Conclusion</td>
<td>203</td>
</tr>
<tr>
<td>IV Results</td>
<td>207</td>
</tr>
<tr>
<td>10 Results</td>
<td>209</td>
</tr>
<tr>
<td>10.1 Introduction</td>
<td>209</td>
</tr>
<tr>
<td>10.2 Prerequisite 1: quality of tags</td>
<td>210</td>
</tr>
<tr>
<td>10.2.1 Prerequisite 1a: explicit quality of tags</td>
<td>211</td>
</tr>
<tr>
<td>10.2.2 Prerequisite 1b: implicit quality of tags</td>
<td>217</td>
</tr>
<tr>
<td>10.3 Prerequisite 2: integration of tagging into the working processes</td>
<td>219</td>
</tr>
<tr>
<td>10.4 Prerequisite 3: involving all the employees</td>
<td>219</td>
</tr>
<tr>
<td>10.5 Recommendations</td>
<td>221</td>
</tr>
<tr>
<td>10.5.1 Recommendations to unlock tacit knowledge from tags: expert finding</td>
<td>221</td>
</tr>
<tr>
<td>10.5.2 Recommendations to unlock explicit knowledge from tags: tag dashboards</td>
<td>222</td>
</tr>
<tr>
<td>10.6 Conclusion</td>
<td>222</td>
</tr>
</tbody>
</table>
V Conclusion

11 Conclusion and future research

11.1 Introduction

11.2 Question 1: how can we select high-quality tags in a broad and narrow folksonomy?

11.3 Question 2: how good is the implicit quality of corporate tags?

11.4 Question 3: how can an analysis of employees’ tags help a company to unlock the tacit knowledge that resides in the organizational memory?

11.5 Question 4: how can employees’ tags help a company to unlock the explicit knowledge that resides in the organizational memory and how does this influence the business strategy?

11.6 Prerequisites and recommendations

11.7 Future research

11.7.1 Tag dashboards

11.7.2 Tag purpose

11.7.3 Corporate ontology through tags improves the information retrieval problem

11.8 Final conclusion

A Appendix chapters 6 and 7

A.1 Introduction

A.2 Methodology to verify tag rules in Chapter 6

A.3 Preparatory steps for interviews in Chapter 7

A.3.1 Sample of messages

A.3.2 Participants

A.3.3 List of suggested tags (type B)

A.3.4 Test interview

A.3.5 Post processing: tags and word cleansing

A.3.6 List of synonyms used to clean-up tags

A.3.7 List of c/k words used to clean-up tags

A.3.8 List of questions used in Chapter 6

B Appendix chapter 8

B.1 Preparatory steps to apply tag analysis techniques

B.2 Preparation social networks

B.3 Interview questions

B.3.1 Part I: evaluation of the tag analysis techniques

B.3.2 Part II: social network

C Appendix chapter 9

C.1 Creation tag dashboards

C.2 Preparatory steps for interviews

C.2.1 Description of interviews

D Appendix SPSS

Glossary
Acknowledgement

Johann Wolfgang von Goethe was right: “Magic is believing in yourself, if you can do that, you can make anything happen.”. After almost five years of hard work, I am very proud to deliver this dissertation to you. As every PhD student I had my ups and downs, but I never gave up even when some things did not go well.

During these past five years many people have played a very important role in my life and work. This dissertation would not have been possible unless with the guiding advice and support of all of them. So, thank you with all my heart to each one of you!

First of all I would like to thank my supervisor Prof. Dr. Eddy Vandijck to help, encourage, trust, and believe in me all these years. Of course, I cannot forget to thank Prof. Dr. Frank Plastria who gave me a chance to work as an assistant researcher/teacher at the MOSI-research group. Also, I would like to thank all the other members of my Phd commission, Prof. Dr. Eddy Torfs and Prof. Dr. Olga De Troyer, as well as the other jury members Prof. Dr. M. Despontin, Prof. Dr. J. Vanthienen, Prof. Dr. P. De Leenheer, and Dr. T. Coenen for all their valuable feedback, support and their interest in my research.

Next, I am very grateful to Ria, Marc, Liesbeth, and An and all the other employees of the Company for giving me such a valuable dataset and allowing me to do all these interviews. Without their cooperation I could not have gotten such relevant data for my research. I would also like to thank them for being always approachable.

I would like to make a special reference to my love Bert, who was always there to help and advise me. I could always rely on him when something did not go well or when I was stuck with a problem. But more importantly, he helped me gain more self-confidence and convinced me I could learn new things. The last year has been very stressful for both of us as we both had our own project going on, but no matter what he always stood by me.

I would like to express gratitude to those who took the time out to encourage me and
put me on track at the beginning of my research: Dr. Dirk Kenis and Dr. Tanguy Coenen. In 2005, there was not yet a lot of research available on folksonomies and tagging and I was not sure whether corporate tagging would be a valuable research topic for my PhD. Dr. Dirk Kenis and Dr. Tanguy Coenen supported my choice of topic and I am very pleased with the choice I made. Next to this, I am also very grateful to Dr. Dirk Kenis for putting me in contact with the Company and helping me to get the valuable dataset I could use for my research.

I owe my deepest gratitude to Prof. Dr. Martin Hepp who gave me very valuable feedback on my research and provided me with some interesting research directions for my PhD.

I would also want to thank Mark Van Lokeren and Prof. Dr. Marc Despontin to help me with the preparation of the interviews discussed in Chapter 7. Also, I am very grateful to Mark Van Lokeren and Dennis Van den Broeck for helping me with the statistical analysis in the dissertation and for validating the analysis.

I am also heartily thankful to many people, it would have been next to impossible to write this thesis without the help and guidance of my colleagues Eiblin Matthys, Andrei Fierens and all the other MOSI-colleagues, the input from the thesis students Sarah Berghmans, Christophe Hautier and Vincent Sterken, the fruitful discussions I had with many other people such as Prof. Dr. Pieter De Leenheer, Stijn Christiaens, Christophe Debruyne, Aldo Demoor, and Jaak Leenknegt. Of course, I cannot forget to mention Katleen Cornelis, our MOSI secretary, who always had a listening ear and a word of encouragement.

Lastly, I would like to thank my mother and father for all their unconditional love and support all these years.
Part I

Positioning of the research
Chapter 1

Introduction

1.1 Structure

In this chapter, we start by giving an overview of literature related to the research we discuss in this dissertation. To detail our research questions, we introduce a number of concepts which are discussed in the other chapters. Since our research is based on case-based research, we discuss the misunderstandings that often exist for this type of research and evaluate their relevance to our work in Section 1.4. Then, we elaborate on our contributions and provide an overview of all the chapters covered in this dissertation in Section 1.6. We end this chapter by giving an overview of all our publications related to our research.

1.2 Overview of literature

In the last few decades knowledge has been considered a competitive advantage for companies (Nonaka [1994]). The economic resources by which a company differentiates from its competitors has moved away from the traditional resources, such as labor and capital (Drucker [1994]). Knowledge is generated in the mind of human beings through the pro-
cess of absorbing and processing information (Alavi and Leidner [2001]).

The corporate information is acquired, stored, and retrieved in the organizational memory. An organization does not really have a memory, because it is not able to remember things and events. Walsh and Ungson [1991] defined it as “stored information from an organization’s history that can be brought to bear on present decisions”. They claim that experiences from the past that are stored in a number of storage bins can help to solve problems of the present. The organizational memory can actually be considered as a collection of explicit and tacit knowledge that resides in the organizational memory (Alavi and Leidner [2001]). Nonaka [1994] describes tacit knowledge as knowledge which is stored in people. Explicit knowledge is knowledge which is available in any kind of tangible format (Nonaka [1994]).

Companies, however, are dealing with an overload of information. The information and communication technology (ICT) is regarded as one of the main causes of information overload. ICT has simplified the process of creating information in an electronic or digital format, for example electronic mail (Bawden et al. [1999]). Human beings are said to be able to process a certain amount of information, but when a critical amount of information is reached, additional information processing becomes a burden because it has an impact on the decisions someone has to take (Schroder et al. [1967], Eppler and Mengis [2004], Ceglar and Roddick [2006]).

Eppler and Mengis [2004] classified the effects or symptoms of information overload in an organization into four categories: limited information retrieval possibilities, problems in organizing all the information, impact on performance of the individual, and impact on the decision-making process.

The problem of information overload is also present, and even on a much larger scale, on the World Wide Web. The World Wide Web contains several billion web pages which need to be quickly and easily accessible by its users (Lyman et al. [2003]). Nowadays, web users experience more and more problems to retrieve the information that fulfills
their needs through search engines. To solve this problem, researchers are working on an extension of the current World Wide Web, a Web where all the information will be interpretable by machines: the Semantic Web (Berners-Lee et al. [2001]). To make the Semantic Web a reality, technology is needed. The World Wide Web Consortium (W3C), the organization that creates standards and technologies for the World Wide Web, has suggested building a Semantic Web Stack. The Web Stack consists of several technologies, such as ontologies (Berners-Lee et al. [2001]). Briefly, ontologies try to describe all the concepts, instances and relations of a certain domain, mostly in a formal or machine understandable language (Gruber [2008]).

In the domain of information retrieval, ontologies are suggested to be used for query expansion and for indexation of documents and queries. Query expansion extends queries with semantically related concepts, for example synonyms, to improve the search results.
Search results can also be improved by using the concepts of the ontology to index the documents and queries (Gamper et al. [1999]). Since ontologies are expensive to build, researchers are looking for new means to make them less expensive (Specia and Motta [2007], Van Damme et al. [2008c]).

Recently, academic researchers active in Semantic Web research started to explore the possibilities of tagging to aid in building ontologies (Specia and Motta [2007], Braun et al. [2007], Van Damme et al. [2008c]).

Users can enter any words that enter their minds when they label resources. This makes them active participators in creating new tags. Aggregating this user-created metadata leads to a flat, bottom-up taxonomy\(^1\), also known as a *folksonomy*. It was Thomas Vanderwal who coined the term in 2004\(^2\). He observed tagging activities on sites such as del.icio.us\(^3\) and Flickr\(^4\). On these sites, users describe their bookmarks or pictures with tags. People, or the folk, are creating their own taxonomy or *folksonomy*, a contraction of the words *folk* and *taxonomy* (VanderWal [2007]).

Since companies are dealing with a similar problem of information overload (Edmunds and Morris [2000]), we wondered whether tagging can also generate an added value to them. Can the information overload problem be tackled through unlocking the knowledge in the organizational memory? More specifically, *we were interested in researching how employees’ tags can help a company to unlock the tacit and explicit knowledge that resides in the organizational memory and how this can influence the business strategy.*

We are not aware of any research discussing the added value of tagging towards unlocking the explicit and tacit knowledge located in the organizational memory. Although some papers discuss how tags can be used to find experts in the company, none of them presents tag analysis techniques that have been tested in a corporate setting. Most research

---

\(^1\)Definition of *taxonomy*: Chapter 3 or *Glossary*.
\(^2\)More detailed information is provided in chapter 3.
\(^3\)http://del.icio.us
\(^4\)http://www.flickr.com
on the use of corporate tags is focused on how bookmarking systems such as del.icio.us can be applied in a corporate setting. For example, the authors in Casey et al. [2007] present a social bookmarking tool, called Dogear. In Dogear, employees can bookmark and annotate the web pages they retrieved on the corporate intranet as well as on the world wide web. However, we believe that tagging in a corporate environment has far more value: retrieving knowledge through an analysis of employees’ tags.

Since users can enter any words that enter their minds when they tag resources, *we have to question the quality of tags*. In a corporate environment, we might have to inspect the quality even more since employees might not directly receive some kind of personal benefit, when they create tags, as is the case on the World Wide Web. To investigate whether the analysis of employees’ tags can help a company to unlock the explicit and tacit knowledge that resides in the organizational memory, we need to have a closer look at the quality of tags and, more specifically, the quality of corporate tags. Before we can measure the quality of corporate tags, we need to define the *quality of tags* and find techniques that will allow us to retrieve *high-quality tags*.

To discuss the quality of tags, we believe a distinction should be made along two dimensions: *implicitness-explicitness* and *type of folksonomy*. We define *implicit* tag quality as a measurement of tag quality that is not directly observable by analysing a tag on its own. More specifically, the implicit quality of a tag is related to the other tags as well as the annotated resource.

In contrast to implicit tag quality, we can measure the *explicit* tag quality by studying a tag itself. To measure the explicit tag quality, we need to verify whether the tag is consistent or in line with a uniform tag format. Of course, before the *implicit* quality of tags can be measured, we need to be sure to have consistent tags or tags of good explicit quality. Therefore, we will propose a methodology of how to have consistent or cleaned up tags in the case of a corporate tag dataset with mainly Dutch tags.

Next to *implicitness-explicitness*, we believe the type of folksonomy should also be taken into account. VanderWal [2005] distinguished two types of folksonomies: *broad*...
and narrow folksonomies. Whereas many users annotate the same resource in a broad folksonomy, only a few users label a resource in the case of a narrow folksonomy, mostly one person. As we will detail later on in this dissertation, the type of folksonomy is an important dimension to take into account when we have to measure implicit quality of tags.

Until now, only a small number of publications discuss the quality of tags used on the World Wide Web. None of these publications mention that a difference should be made between broad and narrow folksonomies when evaluating the quality of tags. Most of them implicitly assume a broad folksonomy, whereas a narrow folksonomy also exists, as we discussed above. Besides, none of the publications have yet to investigate the quality of corporate tags.

To investigate our research questions, we will use a del.icio.us dataset as well as the tag data set from a large European distribution company with headquarters in Belgium. The company employs more than 15,000 people across Europe. Even though the tagging system at this company is somewhat different from current web-based tagging practices, the 20-years worth of tagged messages represents a valuable case. Such cases are rare, as not many organizations have adopted tagging in a way which allows the analysis of a large body of tags.

To obtain the dataset a non disclosure agreement between the university and the company was signed. For this reason, we cannot mention the name of the company and we will refer to it as the Company throughout this dissertation.

1.3 Research questions

In this dissertation, we will discuss the research questions we formulated below

---

6 In Chapter 4, we will explain why we did not mainly use the dataset of the Company.
• Question 1: How can we select high-quality tags in a broad and narrow folksonomy?

• Question 2: How good is the implicit quality of corporate tags?

• Question 3: How can employees’ tags help a company to unlock the tacit knowledge that resides in the organizational memory?

• Question 4: How can employees’ tags help a company to unlock the explicit knowledge that resides in the organizational memory and how does this influence the business strategy?

1.4 Case-study research

The research strategy we use in this dissertation can be categorised as case-study research.

“Case study research consists of a detailed investigation, often with data collected over a period of time, of phenomena, within their context. The aim is to provide an analysis of the context and processes which illuminate the theoretical issues being studied.” (Cassell and Symon [2004])

Cassell and Symon [2004] and Yin argue that how and why research questions are most appropriately answered through case-study research. Therefore, we mainly used how research questions in our dissertation.

Case-based research is often criticized in that it cannot be considered as a valuable research strategy. Flyvbjerg [2006] did an extensive literature research on case-based research. Based on the literature research, he concluded that five misunderstandings exist about this kind of research. In his work, however, he argues why each of these statements are incorrect and reformulates each of them. For instance, in literature, case studies are often criticized because they cannot create scientific value because it is not possible to generalize from one case. Flyvbjerg [2006] gives an overview of a number of valuable
scientific contributions that were realized based on the analysis of only one case.

Since cases where companies let their employees annotate messages or documents in the company are rare, difficult to find, and not always accessible, we focused on one case study for the research. Cassell and Symon [2004] explain that it is allowed to do a single case study when it is hard to find another case. However, “the challenge is to disentangle what is unique to that organization from what is common to other organizations”. Therefore, we formulate in Chapter 10 a number of recommendations based on the answers to our research questions.

1.5 Contributions

This dissertation contains a detailed body of work on how corporate tags can be used to unlock knowledge in a company. We summarize the key contributions as follows:

- **Guidelines regarding the quality of corporate tags**: since tagging does not have a built-in control mechanism, we explained how this can have an impact on the quality of the tags. The results of measuring the explicit as well as implicit tag quality of the *Company* even showed that the quality of tags can indeed be an issue. We provide guidelines on how the explicit as well as implicit quality of tags should be verified and improved in a corporate environment.

- **An approach to unlocking tacit knowledge that is in companies**: we explain how experts in the company can be found through an analysis of tags. On the one hand we tested three tag analysis approaches, and visualized the social and ego networks on the other hand. We applied the approach to a sample of the dataset of the *Company* and had the results evaluated by 10 employees.

- **An approach to unlocking explicit knowledge that is in companies**: to unlock explicit knowledge, we suggest a tag dashboard approach derived from the tradi-
tional dashboard approach of Kaplan and David [1992]. To verify the approach we had interviews with five managers within the Company.

- **Recommendations for a company that considers using tags to unlock knowledge in the company:** we provide a list of issues that a company should be aware of, as well as a number of recommendations that a company should take into account when they plan to use a tag as a tool to reveal knowledge within the company.

- **Business strategy:** we will show how unlocking corporate knowledge can effect the business strategy and can help a company to determine whether or not the business objectives are met. In addition, we will show how tag dashboards can be completely integrated in the IT-governance model.

## 1.6 How the thesis is organized

The thesis is organized around eleven chapters as shown in Figure 1.2. After this introductory chapter, *Chapter 1*, we start with the dissertation’s *background*.

The *background* provides an overview on the literature related to the research we will discuss in this dissertation. It consists of two chapters. *Chapter 2* explains some concepts in the domain of knowledge management, such as the differences between data, information and knowledge, and the concept of organizational memory. Next, we discuss the information overload problem on the World Wide Web and how this problem is currently handled on the World Wide Web.

On overview of tagging and folksonomies is detailed in *Chapter 3*. We start the chapter by explaining the difference between classification and categorization and argue why tagging should be considered as a categorization technique. Next, we give an overview of the main topics often discussed in publications on tagging and corporate tagging.

The third part of the dissertation, the research part, contains all the chapters that
formulate an answer to each of the research questions. In Chapter 4, we discuss the problem of tag quality in general. Next, we define high-quality tags and explain how they could be retrieved in a broad and in a narrow folksonomy. In the last section, we discuss the results of testing three tag quality algorithms on a del.icio.us dataset.

A complete overview and description of the case study and dataset we used to investigate the other research questions is detailed in Chapter 5. In addition, we also list the problems we encountered with the dataset and provide some descriptive statistics. The CorTagCleaning approach to clean up the tags of the dataset, in order to obtain consistent tags or tags with a good explicit quality, is explained in Chapter 6. In Chapter 7, we discuss the results we obtained when measuring the implicit quality of tags of the Company.

To find out how we can unlock the tacit knowledge based on an analysis of employees’ tags, we suggested three techniques to analyse tags in order to find experts in the Company. In Chapter 8, we discuss the results from these techniques and detail the
feedback we received from employees on these techniques.

In Chapter 9, we explain how tag dashboards based on an analysis of employees’ tags might help a company to unlock the explicit knowledge that resides in the organizational memory. We detail the results we obtained when discussing the tag dashboards with a number of managers within the Company.

The answers to each of the research questions allow us to detach the results from the case study. In Chapter 10 we formulate general answers to each of these research questions as well as a number of recommendations for other companies that want to use tagging as a business tool.

We end the dissertation with a conclusion and a number of interesting directions for future research in Chapter 11.

Details from the research that are not described in the other chapters are discussed in the Appendices.

In the Glossary, we provide an overview of definitions of frequently used words.

1.7 Publications

Below we provide an overview of all our publications related to the research we present in this dissertation.


C. Van Damme, C. Approaches to Analyse Corporate Tags for Business Intelligence Purposes. In *Proceedings of the first International Workshop on Ontology-supported Business Intelligence (OBI2008) in conjunction with ISWC2008*, Karlsruhe, Germany,


Part II

Background
Chapter 2

Organizational knowledge

2.1 Introduction

To detail the research questions in this dissertation, we need to provide an overview of the literature in the domain of knowledge management. In this chapter we provide an overview of a number of concepts. We start by explaining the difference between data, information and knowledge in Section 2.2, explain the differences between tacit and explicit knowledge in Section 2.3 and discuss the concept of organizational memory in Section 2.4. Next, we highlight the problem of information overload in organizations as well as on the World Wide Web. In Section 2.6, we explain how ontologies might solve the problem of information on the World Wide Web and discuss the problems that ontologies currently face.
2.2 The difference between data, information and knowledge

It is important to make a distinction between data, information, and knowledge, because the words data, information, and knowledge are often used interchangeably. However, a clear distinction between the terms exists (Devlin [2000]).

Let us start by explaining the difference between data and information. Devlin [2000] defines data as any kind of output we get from computers or other sources. For example, the bits (0 or 1) held by the transistors in a microchip of a computer can be considered as data. A bit holding the value 1 is generated when electric current is passing through, and a 0 when it is not passing through. All these bits together or data do not tell us anything unless we add a meaning or interpretation to it. Human beings added a meaning to each of these bits in order to represent every number, word and graphic as a combination of bits (White and Downs [2008]). When data are given a meaning, information is created (Anand et al. [1998] and Lillrank [2003]).

Another example of data are the observations in meteorology and astronomy. For instance, the number 80 is considered as data but does not tell us anything unless we add a meaning or interpretation to it: the number of moons in the solar system or the average wind speed last month.

Next, the terms information and knowledge are also often used as synonyms, but a clear distinction between them exists, as stated by Nonaka [1994], and Alavi and Leidner [2001].

Today, many definitions of knowledge exist. The epistemological definition of the word knowledge is derived from ancient Greek. Ancient Greeks defined something as knowledge when three conditions were met: that it was “justified, true, and believed”. This means that something must be true before someone can have knowledge of it. This definition is extended by others such as Nonaka, who defines knowledge from the know-
CHAPTER 2. ORGANIZATIONAL KNOWLEDGE

In his work, Nonaka [1994] explains that more focus should be put on the personal belief and justification of the individual. Therefore, he defines knowledge as “created and organized by the very flow of information, anchored on the commitment and beliefs of its holders” (Nonaka [1994]). He argues that information is one of the main fundamentals for creating knowledge.

We notice several approaches regarding the relations of data, information, and knowledge. On the one hand, there is the well known hierarchical relation which states that knowledge can be considered as an extension of information and information as an extension of data (Devlin [2000]). On the other hand, there is the reversed hierarchy as Tuomi [1999] proposes in his work. He believes there cannot be data before there is information and knowledge. Knowledge is needed before someone can create information and consequently data. We do not agree with latter approach. In the case of meteorological observations, a lot of data is available that needs to be analysed in order to become information. In the long run the information can be turned into knowledge.

Next to the top-down or bottom-up hierarchical approaches, there is also the approach of Alavi and Leidner [2001], who state that information is turned into knowledge once it is processed in the minds of human beings. It is transformed into information when it is put into words again.

We agree with the latter approach as well as with the hierarchical approach. Knowledge is broader than information and information is consequently broader than data. However, we believe knowledge can easily be transformed into information and vice versa, as argued by (Alavi and Leidner [2001]). But, we also agree with Nonaka [1991] that knowledge can be explicit as well as tacit as we will explain in the next section. In the rest of this dissertation, we will use the term knowledge when we talk about information that is interpreted by humans and information when it is not yet interpreted by humans.
2.3 Tacit and explicit knowledge

In his work, the knowledge creating company Nonaka [1991] argues that a company should create knowledge and transform this knowledge into innovations, for example by creating new products and technologies. By giving a number of examples of how innovations in Japanese companies have occurred, he shows the importance of the principle of the knowledge creating company.

A basic condition for the creation of knowledge starts with knowledge from the individual. The individual knowledge should be transformed into organizational knowledge and as such it becomes a valuable asset for the whole company. Although large amounts of knowledge are available in a tangible format, not all the knowledge is available in such a format. Some knowledge is stored within people (Nonaka [1994]).

Nonaka explains that Polanyi was the first to make a philosophical classification for knowledge. Later, Nonaka extended the definition to a more practical use of the terms. He describes tacit knowledge as knowledge which resides in individuals whereas explicit knowledge is knowledge which is available in any kind of tangible format (Nonaka [1994]).

Both types of knowledge can be transformed from one form into the other. In his work, the knowledge creating company, Nonaka [1991] describes four patterns that can be distinguished when creating knowledge:

1. Socialization: Tacit to Tacit: when someone explains something to someone else during a conversation, tacit knowledge is exchanged.

2. Combination: Explicit to Explicit: when writing a summary of a book or paper.

3. Articulation: Tacit to Explicit: taking notes during a presentation and then including it in a report transforms tacit knowledge into explicit knowledge.

4. Internalization: Explicit to Tacit: reading a book transforms the explicit know-
Figure 2.1: The four knowledge creating patterns of Nonaka (own visualization)

Some researchers argue that explicit knowledge is far more important than tacit, whereas others claim the opposite. We agree with Alavi and Leidner [2001], who state that none of these statements are right: neither type of knowledge can be seen separated from the other. They both need each other and there needs to be a shared knowledge between both parties (Alavi and Leidner [2001]).

### 2.4 Organizational memory

The corporate information is acquired, stored, and retrieved in the organizational memory. An organization, however, does not really have a memory, because it is not able to remember things and events. Many definitions of an organizational memory exist. In this
work, we refer to the definition of Walsh and Ungson [1991]. Walsh and Ungson [1991] define organizational memory as "stored information from an organization's history that can be brought to bear on present decisions". It implies that experiences from the past can help to solve the problems of the present. However, the authors warn that this cannot be done blindly, as although at first sight there might be an analogy with the past, there are not always similarities between the present and past. Therefore, some caution is necessary.

In their work, Walsh and Ungson [1991] explain that organizational memory can be stored in five different bins inside the organization and one outside. The external archives contain information on the organization that is not kept inside. It can be collected by an external source, for instance a company, that keeps specific information on certain companies or employees that have left the company a couple of years previously. The internal bins contain all the information that is collected inside the organization: individuals, culture, transformations, structures, and ecology.

1. Individuals: store all the information they encounter and observe while performing their jobs.

2. Culture: holds the experiences from the past that can be used in the future.

3. Transformations: are the transition from an input to an output. An input can be anything, a raw material or a new recruit. The output can be a finished product or a loyal employee. Each time a transformation takes place, acquiring and processing information becomes necessary.

4. Structures: are used in the sense of organizational structure: what role or profession does a certain employee play in the company? Walsh and Ungson [1991] make a comparison with sociology where people expect a certain behaviour from people who carry out a certain role or profession. Therefore roles can be considered as a repository to store organizational information.
5. Ecology: concerns the organization’s workplace or how the employees are placed in the office: where people’s desks are placed and how they are organized.

The organizational memory can actually be considered as a collection of (1) written documents and other types of information stored in information systems or printed format, the so called explicit knowledge, and (2) all the tacit knowledge (Alavi and Leidner [2001]).

Stein and Zwass [1995] claim that organizational memory needs to be supported by information technology. When there is so much information present in a company, efficient ways to manage and retrieve all this information and as a consequence knowledge are required. Therefore, information technology is considered as a means for this.

In the last few years, we noticed that social software tools, such as wikis\(^1\) and blogs\(^2\), that were initially used on the World Wide Web\(^3\) have now found their way into companies to help employees archive knowledge from the organizational memory (Munson [2008]). These social software tools are easy-to-use content tools with low technology barriers. In such a way, they increase the amount of explicit knowledge by archiving the implicit knowledge. In Munson [2008], wikis are suggested as a technological means to help employees access knowledge on past projects. In the same way, other social software tools could easily be applied for “articulation” of the organizational memory.

### 2.4.1 Transactional and individual memories

The information of individuals, one of the internal bins of the organizational memory, is actually stored in the individual memory systems. These systems allow communication among the different individuals (Wegner [1986]). The collection of all the individual

---

1. A wiki is a tool which allows you to create web pages and links between these pages very easily with a WYSIWYG editor (http://en.wikipedia.org/wiki/wiki, retrieved 15 August 2009).
2. A blog or weblog is a kind of online personal diary (http://en.wikipedia.org/wiki/Blog, retrieved 15 August 2009).
3. The most well-known example is probably Wikipedia.
memory systems makes up the transactional memory. In this way, we can say that the transactional memory is part of the organizational memory and consists of different individual memory systems.

Wegner [1986] explains that the processes that take place in a person’s memory occur in three different steps: encoding, storage, and retrieval. When humans store information, we automatically label it and connect it to other sets of information. This means we do not store the items of information one by one as is the case with external memory. During the encoding, human beings sort items into categories and organize the words into conceptual groups. For example, a human will automatically associate the words ball and round with each other and therefore information will be retrieved very easily, whereas external memory requires an extended classification mechanism\(^4\) to retrieve the same kind of information. This is the reason why human beings are good at retrieving information within a very small period of time. Meanwhile, an external memory requires a good classification mechanism\(^5\) that takes all these semantic relations into account (Wegner [1986]).

The individual memory can be categorized into two groups: episodic and semantic memories. The term semantic memory was first mentioned by Quillian in his dissertation (Tulving [1972]). Tulving extended this definition and introduced episodic memory as a contrast to semantic memory. Episodic memory stores and retrieves information that is related to personal experiences whereas semantic memory is necessary for the use of language. For instance, recalling that NaCl is the chemical formula for salt is part of the semantic memory whereas remembering a specific event such as one’s wedding is part of the episodic memory (Tulving [1972]).

\(^4\) An overview of all the classification techniques will be provided in Chapter 3.

\(^5\) More detailed information is provided in Chapter 3.
CHAPTER 2. ORGANIZATIONAL KNOWLEDGE

2.5 Information overload in companies and on the world wide web\(^6\)

The organizational memory can actually be considered as a collection of all the explicit as well as tacit knowledge that resides in the company (Alavi and Leidner [2001]).

The *information overload*, however, makes it very difficult for companies to manage the information and thus knowledge stored in the organizational memory.

According to Bawden et al. [1999] and Edmunds and Morris [2000] no generally accepted definition of the term *information overload* exists. However, in Eppler and Mengis [2004], an overview of frequently used definitions of information overload is given, but each of these definitions contains different components or dimensions, as the literature analysis done by Eppler and Mengis [2004] shows. Bawden et al. [1999] suggest to describe it as “*information overload occurs when information received becomes rather a hindrance than a help when the information is potentially useful*”.

Information and communication technology (ICT) is regarded as one of the main causes of information overload, because ICT has simplified the process of creating information in an electronic or digital format, for example electronic mail (Bawden et al. [1999]). The problem of information overload is not something which started in the 1990s with the rise of the Internet; the problem already existed in the 1950s (Bawden et al. [1999], Edmunds and Morris [2000]) due to the increase in the number of scientific publications and the start of mechanized documentation (Bawden et al. [1999]). At that time, it was already stated that information overload influences people’s effectiveness, for instance the effectiveness of the researcher in finding the most appropriate resources for his work (Bawden et al. [1999]).

The problem of information overload also exists on the World Wide Web, but on a

---

\(^6\)This section is largely based on a section of: C. Van Damme, Folksonomies and enterprise folksonomies. Master thesis, Vrije Universiteit Brussel, 2006.
much larger scale. Every day a huge amount of information is published on the World Wide Web. In 2000, the School of Information Management and Information Systems (SIMS) of the University of Berkeley estimated that there were 20 to 50 terabytes of information available on the surface or visible Web. Three years later this school did the same test and concluded that the volume of information had already tripled to 167 terabytes. That is almost 17 times the amount of information which is stored in the nineteen million books and other print collections of the Library of Congress in Washington (Lyman et al. [2003]).

The surface Web consists of a collection of static Web pages, Web pages which are always available, of course when the Web server is not down. Therefore they can be easily indexed by traditional or commonly used search engines such as Google and Yahoo. Most search engine use spiders or Web crawlers which index every Web page they encounter on their list of URLs. After they indexed a Web page, they store the results of the indexation in a database. Each time a Web page includes a link to another Web page, the URL address is added to the list of URLs. This is the reason why it sometimes takes some time before an Internet page is indexed by the spider and listed in the results of the search engine (Morville and Rosenfeld [2006]).

This is not the case with the deep or invisible Web. Web pages on the invisible or dark Web are dynamically created and are hence not constantly available for indexation by search engines. A dark Web page is created at the moment that the user asks for particular information that is stored in a database, for example a phone number. After the information is retrieved from the database, the Web page is created. The Web page is removed when the user closes the Web page (Bergman [2001]). According to BrightPlanet, a company which has a search technology for the deep Web, the volume of this dark or invisible Web is estimated to be 500 times larger than that of the surface Web (Bergman [2001]).

---

7 A definition is given in the next paragraph.
A lot of the information which is available on the visible and dark Webs is created through social software tools such as blogs and wikis. The low technology entry barriers and low costs are among the reasons why so much information is created online these days. Many wiki and weblog software applications are offered as a free online service. Knowledge of HTML or other kinds of scripting languages is no longer a prerequisite to creating content online.

Of course, the large amount of information on the World Wide Web makes it very hard to seek information online. Directories and search engines are among the tools most frequently chosen to retrieve information from the World Wide Web. Nowadays, it is becoming hard for users to find the information they are looking for. Search engines are not yet capable of interpreting the semantic meaning that exists between the search keywords entered by a user on the World Wide Web. When you try to find the opening hours of a specific shop, for instance, there is only a slight chance that the search engine will actually be able to locate for you the website that displays the opening hours of the particular shop you asked for (Guarino et al. [1999]). Also, search engines are not yet capable of understanding the semantic meaning of the words on a web page when they are making an indexation of the web page. In this way, we can say that surface pages can be considered as dark pages.

To solve this problem, researchers are working on an extension of the current World Wide Web, a Web where all the information will be interpretable by machines: the so-called Semantic Web (Berners-Lee et al. [2001]). To make the Semantic Web a reality, technology is needed. The World Wide Web Consortium (W3C), the organization that creates standards and technologies for the World Wide Web, has suggested building a Semantic Web Stack.\(^9\) The Web Stack consists of several technologies that can be used to create the Semantic Web. Ontologies are one of the technologies of the Semantic Web Stack that make the Web a reality.

\(^9\)We will not discuss all the technologies that are on the semantic web stack. We will only focus on ontologies.
2.6 Ontologies

The word ontology is derived from Greek and means “being or existence”. In the 1980s, the Artificial Intelligence (AI) community was attracted by the concept of ontologies. It considered ontologies as a means to allow communication between knowledge based systems. As ontologies are used to describe existence, the AI community believed they can also be used to represent formal knowledge which is needed to allow communication between knowledge based systems. When one system asks a question at another system, it should be able to answer that question. Therefore, knowledge based systems need to use the same formal knowledge representation (Gruber [1995]).

The most often cited definition of an ontology is the one given by Gruber. He describes it as an “explicit specification of a conceptualization”. As the words specification and conceptualization are often stated in other publications, Gruber rephrases the core elements of his definition in a recently published encyclopedia on database systems (Gruber [2008]). He defines an ontology as a collection of concepts, relationships, and other elements that are critical to describe a domain. To describe a domain, classes, relations, and other kinds of representational vocabulary, as Gruber defines it, are necessary to add meaning as well as a restriction on the use of the vocabulary (Gruber [2008]).

Jarrar and Meersman [2002] describe ontologies as “a branch of knowledge engineering, where agreed semantics of a certain domain is represented formally in a computer resource, which then enables sharing and inter operation between information systems (IS).”.

De Troyer et al. [2003] define ontologies as “concepts in a domain as well as relationships between these concepts and the terminology used”.

Of course, many other definitions regarding ontologies exist, but we do not mention them here. A more detailed overview can be found in Gomez-Perez et al. [2004].

More recently, ontologies were also introduced in domains other than AI, such as
CHAPTER 2. ORGANIZATIONAL KNOWLEDGE

database theory (Gruber [2008]), computer linguistics, and knowledge engineering (Gomez-Perez et al. [2004]). More specifically, they are used in domains such as knowledge management, natural language processing, e-commerce, information retrieval, and the Semantic Web (Gomez-Perez et al. [2004]).

A distinction between two kinds of ontologies should be made: lightweight and heavyweight ontologies. Lightweight ontologies are considered as the basic form of an ontology. They consist of concepts and relationships between concepts. Heavyweight ontologies are an extension of lightweight ontologies. They are more formal and also include restrictions or axioms (Gomez-Perez et al. [2004]).

2.6.1 Goals to use ontologies

Uschold [1996] distinguishes three goals of using ontologies: communication between people and organizations, interoperability between machines, and improving system engineering. The goal of the ontology will determine the level of formality. This means that an ontology needed for communication between people will be rather informal, whereas one used for the communication between machines needs a far more formal approach. A formal ontology is an ontology which is expressed in such a way that it is that it is easy for computers to understand it, an ontology language, as we will explain in a next section, whereas natural language is used to express an ontology in an informal way (Uschold [1996]).

In the domain of information retrieval, ontologies are suggested to be used for query expansion and for indexation of documents and queries. Query expansion extends queries with semantically related concepts, for example synonyms, to improve the search results. When someone is looking for information on a motorcar, the query would give more interesting results when the query is extended with the synonyms automobile and car. Search results can also be improved by using the concepts of the ontology to index the documents and queries (Gamper et al. [1999]).
2.6.2 Languages for creating ontologies

To use ontologies for the interoperability between systems, the ontology should be made formal (Uschold [1996]) by using an ontology language.

There are many languages which can be used to express an ontology, such as the Web Ontology Language (OWL) and Resource Description Framework (RDF). Again, for a complete overview of the ontology languages, we refer to Gomez-Perez et al. [2004] and Hepp et al. [2007]. OWL is considered an ontology standard language by W3C. It is written in Extensible Markup Language (XML) and has more expressiveness than RDF. OWL can be categorized into three groups: OWL Lite, OWL DL, and OWL Full.\(^{10}\)

As it is often very hard to use an ontology language, tools are used to simplify the creation process. Mostly, they have a very easy to use Graphical User Interface (GUI) which allows a translation to the ontology language, such as OWL or RDF. For example, Protégé\(^{11}\) allows you to create an ontology with an easy to use GUI. The elements of the ontology are entered by using different screens for instances, relations, and restrictions. Afterwards, the user can export the ontology to one of the ontology languages: RDF or OWL.

2.6.3 Methodologies for creating ontologies

Today, many methodologies exist to build ontologies. An overview of methodologies is given in Gomez-Perez et al. [2004], Hepp et al. [2007]. Most of the traditional ontology methodologies, however, lack user feedback. A few of them include this in their approach, for example the DOGMA mess approach (Moor et al. [2006]), the Ontology Maturing approach (Braun et al. [2007]), and the six principles of the Community-based Ontology Evolution approach (De Leenheer [2009]).

\(^{10}\)http://www.w3schools.com/RDF/rdf_owl.asp (retrieved in July 2009) 
2.6.4 Problems related to ontologies

Until now, there have not been so many high quality ontologies or non-toy ontologies, as formulated in Hepp [2007]. He argues that there are four problems, technical as well as social issues, that hamper the creation of high-quality ontologies. We paraphrase these problems (Hepp [2007]) in the paragraphs below.

1. It is hard to have an ontology which is not outdated because the creation as well as the process of updating the ontology takes time. Meanwhile, the entities in the domain are constantly evolving. The size of this gap will determine the impact of this disadvantage.

2. There should always be a trade-off between the amount of resources invested to build the ontology versus the benefits the ontology delivers, as is also stated in (Menzies [1999], Simperl and Sure [2008]). As an ontology should describe all the concepts, attributes, relationships between concepts and attributes, and sometimes the restrictions of a domain, it often becomes very time-consuming to build an ontology from scratch.

3. There is a mismatch between the persons who create the ontology and the persons who actually use the ontology. It is important that both parties are using the same vocabulary and thus share the same ontology. De Leenheer [2009] defines this problem as the *impedance mismatch*.

4. To make interoperability possible, the ontologies should be free to use. This means that they cannot be under intellectual property rights, as is often the case, Hepp [2007] argues.

Based on these problems, we can conclude that it is not easy to build high quality ontologies. Therefore, Semantic Web researchers are looking for new ways to simplify the process of building ontologies and to overcome the problems as argued by Hepp [2007].
2.7 Turning folksonomies into ontologies

Recently, academic researchers active in Semantic Web research started to explore the possibilities of tagging for building ontologies (Specia and Motta [2007], Braun et al. [2007], Van Damme et al. [2008c], Peters and Weller [2008], Angeletou et al. [2009]).

Users can enter any words that enter their minds when they label resources. This makes them active participators in creating new tags. Aggregating this user-created metadata leads to a flat, bottom-up taxonomy, also known as a folksonomy. It was Thomas Vanderwal who coined the term in 2004. He observed tagging activities on sites such as del.icio.us and Flickr. On these sites, users describe their bookmarks or pictures with tags. People are creating their own taxonomy or folksonomy, a contraction of the words folk and taxonomy (VanderWal [2007]).

Despite its strengths, tagging has its weakness: no conceptual meaning or any kind of semantic relation is added to tags. As a consequence, tags have no synonyms, different words expressing the same things, or homonyms, words that are pronounced or spelled the same way as each other but describe different concepts. Furthermore, specialized as well as general tags can be used to annotate the same resource (Golder and Huberman [2006], Guy and Tonkin [2006]). Whereas a vet would use the tag Beagle to describe a picture of a Beagle, a novice in the subject of dogs would simply tag the same photo as dog.

These weaknesses can be solved by: (1) giving users tools that enable them to add more information to their tags (e.g. cluster tags, as on del.icio.us; (Guy and Tonkin [2006])) and/or (2) trying to generate more information on the tags by employing text mining and statistical techniques and asking for additional feedback from the community (Van Damme et al. [2008c]).

---


13More information will be given in chapter 3.

14http://del.icio.us

15http://www.flickr.com
In Peters and Weller [2008] the authors propose an overview of techniques, such as formatting guidelines and other existing classifications systems\textsuperscript{16}, which can be applied to enrich tags in a folksonomy data set. The authors compare these techniques to gardening activities. They also introduced Tag Care, a tool that can help the user to manage his own tags created on different systems that rely on tagging systems.

Next, some suggestions have already been made about how folksonomies can be enriched or how they can be turned into an ontology.

In Schmitz [2006] tags of the photo-sharing site Flickr were used in an experiment to induce a taxonomy, the simplest form of an ontology (Uschold and Jasper [1999]). The approach of Schmitz [2006] is based on statistical natural language-processing techniques where a hierarchical relation was deducted.

The authors of Specia and Motta [2007], Van Damme et al. [2008c] and Peters and Weller [2008] suggest including different techniques as well as the wealth of existing online Web resources such as Wordnet, Wikipedia, Google, online dictionaries and existing ontologies.

Specia and Motta [2007] present an approach to enrich tags with semantics to make it possible to integrate folksonomies and the semantic Web. The authors use online lexical resources (e.g. Wordnet, Wikipedia, Google) and ontologies to map tags into concepts, properties or instances and determine the relations between mapped tags. However, the resources are tapped in one way (e.g. Wikipedia is used as a spelling checker for tags) and the community is not involved to confirm the semantics obtained from existing ontologies and resources. Consequently, tags that reflect new concepts, relations or instances or new relations between tags are neglected.

On the contrary, the opposite is suggested in Van Damme et al. [2008c]: ontologies are derived from folksonomies. Online lexical resources are suggested to be exploited in several ways. For instance, Wikipedia is suggested as a spelling checker as well as a

\textsuperscript{16}Definition Classification systems: Chapter 3.
tool for finding concepts and homonyms. Furthermore, the authors suggest involving the community.

2.8 Conclusion

The problem of information overload is also present, and even on a much larger scale than in companies, on the World Wide Web. We discussed how ontologies can be a means to solve this problem, but they are hard to build. Recently, tagging has been considered to be a solution to the current ontology engineering problems.

Since companies are dealing with a similar problem of information overload, we wondered whether tagging can also create an added value to companies. More specifically, we investigated how an analysis of employees’ tags can help a company to unlock and to access the tacit and explicit knowledge that resides in the organizational memory. In addition, we researched how accessing corporate knowledge through an analysis of tags can influence the business strategy and can help to determine whether or not the business objectives are met.

To discuss the results of this research, we first provide some literature on tagging and corporate tagging in the next chapter.
Chapter 3

Folksonomies and tagging

3.1 Introduction

Tagging can be considered a form of categorization. Because confusion exists between classification and categorization, we start this chapter by explaining the difference between classification and categorization. We provide an overview of the different types of classification in Section 3.2. In Section 3.3, we detail the concept of tagging, give an example, summarize the differences between folksonomy and taxonomy and provide an overview of related work on tagging. Related work on corporate tagging is explained in Section 3.4.

3.2 Classification

A clear difference between classification and categorization exists. Whereas classification tries to assign each object to one specific and unique class, categorization is less rigid and assigns objects to categories that have similarities (Jacob [2004]).
Garshol [2004] makes a distinction between two kinds of classification: object-based and subject-based classification. There exists a clear difference between resources that are described based on their objects and the subjects they discuss. Metadata can be considered to be an object-based classification whereas controlled vocabularies, taxonomies, thesauri and ontologies are examples of subject-based classifications (Garshol [2004]). A discussion of these two kinds of classifications is given in the paragraphs below.

3.2.1 Object-based classification

Although metadata is now very popular in computer science, they originated in the domain of library science. Metadata are often described as “data about data”, but in the domain of information architecture, metadata are defined as “information about objects” (Garshol [2004]). The word meta is derived from the Greek language and can be translated into with, that which accompanies data and is not regarded as a vital part of the data as Morville [2005] states it.

Library catalogues are a well-known example of metadata. They are used by librarians to store information about library resources (books and other library items). These catalogues contain metadata about the library resources’ author, title, date of creation, subject coverage, and location number (Morville [2005]).

In computer science, more specifically on the World Wide Web, metadata are used for the indexation of web pages. Metadata can be created manually or automatically. For example, search engine spiders, HTML and XML editors and generators can create metadata automatically. They generate more general data such as date of creation, but in most cases these kinds of metadata are not sufficient. Therefore, more intellectual data are necessary and thus non automatically generated metadata or data generated by human processors are required (Greenberg et al. [2001]).

Metadata can be embedded with XML and HTML in the web page by the creator of the web page. Since they are the creators of the resource, they have a good overview of
the content of the web page. Nowadays, there is an increasing demand for “standardized descriptive metadata” to improve semantic interoperability and thus information retrieval on the World Wide Web. Dublin Core Metadata are an example of such standardization and consist of 15 elements as described in Hillman [2005]. They can be included for example in the Resource Description Framework (RDF) or Extensible Markup Language (XML). RDF and XML are both part of the semantic web stack.¹ W3C considers RDF to be a standard to describe the semantics of data in a machine-understandable language and is expressed in XML (Hillman [2005]).

3.2.2 Subject-based classification

Resources that discuss the same topics or subjects are grouped together in case of a subject-based classification. Different techniques can be used to group these resources: controlled vocabulary, taxonomy, thesaurus, faceted classification, and ontology (Garshol [2004]).² Except for ontologies which are already detailed in a previous chapter, we discuss each of the other techniques in the paragraphs below.

Controlled vocabulary

This contains a list of predefined terms and is in library science sometimes called an indexing language. Garshol [2004] remarks that a clear difference exists between the words term and concept. Whereas a term is used to name a concept, a concept can have several names depending on which subject it refers to. Garshol [2004] explains that the purpose of a controlled vocabulary is to prevent its users from using terms that are too broad or too narrow or are misspelled. He defines it thus: “the simplest form of a controlled vocabulary is simply a list of terms and nothing more.” Morville and Rosenfeld [2006], however, propose a definition that is a little bit more strict: “controlled vocabularies are

¹Definition semantic web stack: Chapter 2, Section 2.6 or Glossary.
²Garshol [2004] mentions also a fifth group, other techniques, which contains all the other types of classification that do not belong to one of the other groups, for example a synonym ring. We will not discuss this group as it contains a collection of a lot of small groups.
predetermined vocabularies of preferred terms that describe a specific domain; typically include variant terms.”

In our opinion, the definition of a controlled vocabulary lies at the intersection of these two definitions and therefore we prefer to describe it as: “a list of predetermined terms that describe a specific domain.”

Taxonomy

The word taxonomy originated from the life sciences. In the 18th century, a taxonomy was used to classify all the plants and animals on earth. Every plant and animal could be placed or classified on a specific node where their hierarchical relationship to other plants and animals was described (Garshol [2004]).

However, in the domain of information science, no uniform definition exists for the term taxonomy. Gilchrist [2003] remarks that the term taxonomy has become a more generic term. He distinguishes five different meanings for the term taxonomy and classifies them according to their application: web directories, taxonomies to support automatic indexing, taxonomies created by automatic categorization, front end filters and corporate taxonomies. The definitions for each of these applications are not completely different from one another: there exists some overlap between them.

Garshol [2004] defines a taxonomy as: “a subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy without doing anything further.”

Hepp and de Bruijn [2007] consider a subsumption hierarchy or a transitive subclass off hierarchy, as a prerequisite for a taxonomy. This implies that an instance of a class should also be an instance of all the parent classes in order to be defined as a taxonomy. For example, a Chrysler is an instance of the class van, but also an instance of the subclass car. A van is a subclass off a car. When there is no subsumption hierarchy, Hepp and de Bruijn [2007] argue it should be called a hierarchical classification instead of a
CHAPTER 3. FOLKSONOMIES AND TAGGING

taxonomy.

The Dewey Decimal Classification (DDC) is an example of a system with a hierarchical classification. This classification uses 10 main classes to classify entire library collections. These main classes are further divided into 100 divisions: each main class has 10 divisions. Each division is subdivided into ten sections. These are all represented by Arabic numerals. It is a library classification system that was created by Melvil Dewey in 1873 and published for the first time in 1876. At this moment, it is the most widely used classification system in the world: more than 135 countries use it to organize their library collections (OCLC [2003]).

In our opinion, we believe a hierarchical classification can also be considered as a taxonomy. However, we agree with Hepp and de Bruijn [2007] that a taxonomy with a subclass off relationship has more semantic value than merely a hierarchical relationship.

Thesaurus

Some consider a thesaurus as an extension of a taxonomy (Morville and Rosenfeld [2006], Garshol [2004]) whereas others such as Hepp and de Bruijn [2007] claim the opposite. Morville and Rosenfeld [2006] define it as: “a controlled vocabulary in which equivalence, hierarchical and associative relationships are identified for the purpose of improved retrieval.” Apart from arranging the subjects into a hierarchy, they also contain other relations and information such as broader and narrower relationships between the subjects, related terms, top-level terms, and explaining the meaning within the thesaurus, as Garshol [2004] explains. These elements are all described within the ISO 27883 standard to build monolingual thesauri. Wordnet (Fellbaum [1998]) can be considered as an example of a thesaurus.

However, Hepp and de Bruijn [2007] argue the opposite. A broader/narrower relation is, in terms of semantics, less specific than a subclass of relation that is used in a

---

3 International Organization for Standardization (ISO).
Figure 3.1: Semantic relations in a thesaurus (Morville and Rosenfeld [2006])

taxonomy. In contrast to a taxonomy, an instance of a class in a thesaurus is not always an instance of all the parent classes. In this way, they regard a thesaurus to be less specific than a taxonomy.

Since a thesaurus also contains variant as well as related terms, we believe it contains more semantic value than a taxonomy defined by Hepp and de Bruijn [2007]

**Faceted classification**

Faceted classification is a classification mechanism introduced in 1930 by the Indian mathematician and librarian S.R. Ranganathan. He came up with this new library classification scheme, because he considered the classification systems such as DDC at that time very poor. These mechanisms pushed a user to classify a resource in a predetermined class. S.R. Ranganathan realized that it would become impossible to classify objects in the

---

4This section is largely based on a section of: C. Van Damme, Folksonomies and enterprise folksonomies. Master thesis, Vrije Universiteit Brussel, 2006.
long run, as more and more knowledge was created. A classification of new knowledge required a combination of several subjects and that was not possible with the existing classification schemes such as DDC and the Library of Congress Classification (GLASSEL, 1998). He applied this classification onto his own classification system: the Colon Classification. Therefore, he created a number of categories which he called PMEST. We paraphrase them below (Hjorland [2008]):

- Personality: the characteristics of a subject,
- Matter: the composing materials of a subject,
- Energy: any action related to the subject,
- Space: geographic location of the subject,
- Time: the subject’s period.

S.R. Ranganathan argued that these facets were universal and consequently could be used to classify everything that exists in the world (Morville and Rosenfeld [2006]). However, Morville and Rosenfeld [2006] suggests to use other facets nowadays to classify information on the web.

This subject-based classification system is still used. For example, Wine.com uses four facets to classify its wines: type, region, price and winery. Users can search for their favourite wine by making a combination of these facets (Morville and Rosenfeld [2006]).

### 3.3 Categorization: tagging and folksonomies

During the last few years, we have noticed an increasing interest from web users as well as semantic web researchers in user-created keywords or tags. Users describe their online resources (e.g. bookmarks, pictures or scholarly publications) with freely chosen key-
words or tags in order to simplify the information seeking process. Of course, tags can be used for many other purposes as we will discuss later on in this chapter.

Tags can be considered as a kind of metadata. Adam Mathes makes a distinction between professional, author and user-created metadata and categorizes tags into the last category (Mathes [2004]). Freely chosen keywords or tags cannot be considered to be a totally new concept. The process of choosing keywords to describe the content of a certain resource already existed for quite a long time in other disciplines. For example, in library science, keywords are used as metadata to describe books or articles. Today, authors of certain scientific papers are still asked by the publisher to add some keywords, although these are not always freely chosen, to describe the content of the paper.

Tags differ from previously known keywords because they are analysed and studied in a different way. As there are four\(^5\) parties or entities involved in tagging a resource - actors, tags, resources and tagging systems (Gruber [2005], Van Damme et al. [2008c]) - additional information is generated simultaneously with the creation of tags. For example, tags entail a new way of navigation on the web, called social navigation, because tags as well as annotated resources are mostly publicly available (Morville [2005]) as we will discuss in Section 3.3.3. When someone annotates a resource with a tag that is in our field of interest, it might be interesting to have a look at all the resources that have been tagged by that person in order to retrieve similar information more easily.

Nowadays, more and more websites incorporate a tagging mechanism. For instance, the social bookmark manager Del.icio.us\(^6\), the image sharing system Flickr\(^7\), and the publication sharing system that supports Bibtex Bibsonomy\(^8\). Users can enter any words that enter their mind and create new tags on the spot. This makes them active participants in creating new tags. Aggregation of this user-created metadata leads to a flat, bottom-up taxonomy, also known as a folksonomy as coined by Thomas Vander Wal (VanderWal

\(^5\)However, some authors only consider three parties: actors, tags and resources.
\(^6\)http://Del.icio.us
\(^7\)http://www.flickr.com
\(^8\)http://www.bibsonomy.org
Vander Wal explains on his blog that it all started with a mailing list of information architects. Gene Smith, an information architect, asked the question whether there already existed a term for the phenomenon where people use tags to describe and share information such as is the case on Del.icio.us. After exchanging some messages back and forth, Vander Wal responded to the question: “So the user-created bottom-up categorical structure development with an emergent thesaurus would become a Folksonomy?” A few days later, Gene Smith introduced the term on his blog.

There has been some discussion about the term folksonomy. Some state that it is incorrect to describe it as a taxonomy, as the art of tagging cannot be considered to be a classification but should be regarded as a categorization (Mathes [2004], Golder and Huberman [2006], Halpin et al. [2007]). Although classification and categorization are often used synonymously, a clear difference exists between them (Jacob [2004]). Categorization organizes entities into categories which have boundaries that are not so clear. In contrast, classification puts entities into distinct classes which are mutual exclusive (Jacob [2004]). Golder and Huberman [2006] prefer not to use the term folksonomy in their paper because of this debate and opt for the term collaborative tagging.

Although we are aware of the fact that a folksonomy cannot be considered *stricto senso* as a flat taxonomy, we will use the term in this dissertation, because it enables you to categorize information, but in a different way from a taxonomy. In case there is no taxonomy at all, we believe a folksonomy can be considered as an interesting alternative. This corresponds to the statement Stewart Butterfield, one of the creators of Flickr, made on his weblog on folksonomy: “It’s like 90% of the value of a proper taxonomy but 10 times simpler.” He argues that the lack of hierarchy, synonym control and semantic precision are the reasons why it works (Butterfield [2004]). Emanuelle Quintarelli claims that it definitely cannot be seen as a substitute for the professional classification schemes.

---

10However, we must remark that we will mostly use both terms: folksonomy and tagging.
of librarians. Folksonomies are powerful and innovative provided that they are used in the right way (Quintarelli [2005]).

3.3.1 Folksonomy versus taxonomy

Although, there is some debate on the fact that a folksonomy cannot really be considered to be a taxonomy because it is a categorization and not a classification technique as is the case with a taxonomy, we provide a brief overview of the differences between a taxonomy and folksonomy in the paragraphs below.

1. Controlled vocabulary: A taxonomy consists of a controlled vocabulary. Users always have to select one of the terms of the taxonomy to classify a resource. A central authority decides what new categories will be introduced in the taxonomy (Morville and Rosenfeld [2006]). Maintaining such a controlled vocabulary is very time consuming and a mechanism is needed to maintain the words regularly. In contrast, folksonomies do not cope with this problem: they are chosen by the user when he annotates and thus categorizes an item. Moreover folksonomies consist of freely chosen keywords, which do not follow formal guidelines or restrictions (Guy and Tonkin [2006]). There is no central authority who is going to forbid the tagger to use a specific tag. Even the recommended tags\(^\text{11}\) are only given as a suggestion and the user is not forced to accept recommendations. The free spirit of tagging lets a user select whatever tags he wants to use, he does not have to justify his tag choice, but as a consequence we may question the quality of them as we will discuss in Chapters 4, 6, and 7. An analysis of the tags on the web shows that people sometimes tag in an idiosyncratic or personal way. For example, they use the tags *mydog, myhouse, whatisinmybag* (Golder and Huberman [2006]). Everyone also uses their own format when choosing compound words for annotation, for example *augmented_reality* and *augmentedreality*. As spaces are not always allowed in tags,

\(^{11}\) An example of recommended tags will be given in Section 3.3.3.
some people will write it in one word, while others will split the word. Spelling errors can partially be solved by calculating the Levenshstein edit distance or by applying stemming or lemmatization techniques as we will explain in Chapter 6.

2. Total cost of ownership\textsuperscript{12}: Letting everyone participate in tag creation is less expensive than building a taxonomy “the beauty of tagging is that it taps into an existing cognitive process without adding much cognitive cost” (Sinha [2005]). As we will discuss in Section 3.3.6, people already associate a number of concepts to a resource they see. In this way, they can use these conceptual observations to choose the tags and do not need to invest in an additional cognitive cost. Letting users select a tag instead of asking them to type a search query implies also a lower cost (Sinclair and Cardew-Hall [2008]).

The creation and maintenance of taxonomies can be very expensive. On the one hand they can be built manually by a group of experts as in the case of the Yahoo! directories on the World Wide Web. The experts have to determine and maintain all the terms and their hierarchical relations with a specific domain. Next, documents have to be indexed based on the taxonomy. On the other hand, taxonomies can be developed automatically with taxonomy-building software (e.g. Autonomy). Such software uses several algorithms (such as Machine Learning algorithms) to analyse the documents created by the employees and, based on the results, it develops and maintains the taxonomy. The documents are automatically classified within the taxonomy. However, buying and implementing software requires a financial investment and can be very expensive (Morville and Rosenfeld [2006]).

3. Hierarchical/exclusive: In the case of a taxonomy, a user has to decide which category is more important than other categories. For example how should an article about a cat species native to Africa to be organized, what will be the first category? Country, mammals or cats? A taxonomy is hierarchical and exclusive, whereas tag-\textsuperscript{12}Total cost of ownership (TCO) includes all the direct and indirect costs that are related with an investment in information technology. (L. Mieritz and B. Kirwin, Defining Gartner Total Cost of Ownership, Gartner Group, 2005.)
ging is the opposite. When you label a resource you do not need to think in which
category you will need to put the resource, you just select all the tags you think
are appropriate to describe it and retrieve it at a later stage (Golder and Huberman
[2006]).

4. Problems such as polysemy, synonymy and basic level variation: Taxonomies, as
well as tagging, face the same problems with semantics. The only difference is
that taxonomies have more semantic meaning than tagging (Golder and Huberman
[2006]) as we just discussed in a previous paragraph. Tags have no conceptual
meaning or hierarchical relations which are added to the tags. As a consequence,
tags have no synonyms or homonyms. Furthermore, specialized as well as general
tags can be used to annotate the same resource (Golder and Huberman [2006]; Guy
and Tonkin [2006]). These weaknesses can be solved by:

(a) giving the users tools that enable them to add more information to their tags
(for example cluster tags as on Delicious)(Guy and Tonkin [2006]) or give
feedback by providing recommendations (Sinha [2005])

(b) trying to generate more information on the tags by employing text mining,
statistical techniques and asking for additional feedback from the community
(Van Damme et al. 2007a).

However, some papers discuss how tags can be enriched to overcome this kind of
problem (Specia and Motta [2007], Van Damme et al. [2008c], Peters and Weller
[2008], Angeletou et al. [2009]).

5. Serendipity search versus exact search: Folksonomies are said to be better to re-
trieve information serendipitously or by accident than by exact search (Quintarelli
[2005]).

In recent few years, we have observed growing attention from researchers in the se-
mantic Web domain for folksonomies. As discussed in the previous chapter, Section 2.7,
researchers try to overcome the problems that current ontology engineering techniques have by using folksonomies as a foundation layer for ontologies.

### 3.3.2 Broad versus narrow folksonomy

We have to make a distinction between two kinds of folksonomies: broad and narrow folksonomies as classified by VanderWal [2005]. The difference between these types lies in the number of people that tag the object. In the case of broad folksonomies, a resource is tagged by many people (e.g. bookmarks on Del.icio.us) whereas in the case of narrow folksonomies there are only a few persons involved, in most situations only the author or creator of the resource (e.g. pictures on Flickr). We will discuss both types of folksonomies in the paragraphs below.

**Broad folksonomy**

In the case where the same resource is tagged by many people in their own tag vocabulary, except for the author, it is called a broad folksonomy (VanderWal [2005]). A visualization of a broad folksonomy is shown in Figure 3.2. The people who use the same kind of vocabulary are on this picture grouped together and are labelled with an alphabetic letter (A, B, C, D, E and F). The arrows, which point out to the numbers, represent the tags given by the group of people. The tags used for searching are represented by the arrows pointing out the opposite direction. Each group uses different tags to annotate a resource and to retrieve it. One of the consequences of a broad folksonomy is that it creates a *power law distribution function*, as we will discuss in Section 3.7. A few popular tags are used a lot and many tags are used only a few times.
Narrow folksonomy

In contrast to broad folksonomies, only a smaller number of people tag the resources in a narrow folksonomy. It concerns objects which are not easily searchable or can only be described by using text. The used tags are, most of the time, “singular in nature” as VanderWal [2005] describes it. This implies that only one tag is used to describe a term whereas in case of a broad folksonomy, there are sometimes several tags used to express the same concept/term. Different groups annotate different terms in the case of a broad folksonomy.

The person who creates the object makes it accessible to other users but, contrary to a broad folksonomy, the creator also gives tags to the object, in this case tag 1 (Figure 3.3). Only groups B and F gave tags to the object. The other groups, except for E, retrieve
information by using one of the tags entered by B, F or the creator. An example of such a folksonomy could be a Web post on a blog.

Vander Wal, however, does not explain how many people have to annotate a resource to consider it to be a narrow folksonomy. He mentions a “few users”, but he does not describe an exact number. There are many examples of situations where resources are annotated by merely one person, mostly the creator of the resource, for example Flickr.

3.3.3 Example broad folksonomy: del.icio.us

In scientific literature, del.icio.us is undoubtedly the most cited example of a website that incorporates a tagging mechanism. It is a website which allows its users to bookmark their favourite web pages on a personal web account that is accessible from every computer.
connected to the Internet. Instead of saving the bookmarks on the browser of a personal computer, it is saved on the del.icio.us web server.

![Pop-up window del.icio.us](image)

**Figure 3.4:** pop up window del.icio.us

There are several ways to add a bookmark to your del.icio.us account. You can go to the del.icio.us website, log on and add the bookmark manually, or click on a shortcut button in the browser to directly annotate the web page in a pop-up window. By installing a browser plug-in, del.icio.us buttons appear on top of the browser. When the user encounters an interesting bookmark he does not have to leave the web page to add it to his del.icio.us account. He can just click on the del.icio.us button in his browser and a pop-up window appears. Some fields are automatically populated, such as title and URL address and the user can decide whether he wants to add some additional information. In addition, the user can choose his own keywords or tags to describe the web page. In case the website is already annotated by other del.icio.us users, del.icio.us gives a number of recommendations, such as popular tags and recommended tags. Popular tags are the tags most frequently used by other users to label this web page. They are ranked in order of importance from left to the right. Recommended tags are tags that lie at the intersection of popular tags and the user’s own tags. The user can also add some metadata such as a summary. When the save button is clicked, the information is stored in the user’s personal
del.icio.us account, the window disappears and the user can continue surfing.

![Image of del.icio.us interface]

Figure 3.5: Overview of de.licio.us bookmarks and related tags.

The user can always access his bookmarks, tags and related tags or tags often used together. As the account is by default public, all the bookmarks are visible for anyone who accesses the account. In case you do not want to share a bookmark, you can always make it private. For every web page which is added to the bookmark collection, we can see how many users share this bookmark and get an overview of all the accounts that share a particular bookmark. By clicking on the number, we can easily click through the bookmark collection of one of these users. This function allows us to search for information in a different way. Instead of using a search engine or directory, we can use tags to find people with similar interests or people who have knowledge on certain topics we are interested in. This is also called social navigation (Morville [2005]) as discussed in Section 3.3. “the folksonomy is most often social so that others that use the same vocabulary will be able to find the objects as well.” (VanderWal [2005]).

To minimize the problems of misspellings, tags can be clustered manually into bundles or clusters.
3.3.4 Tag cloud

A folksonomy can be visualised in a so-called tag cloud. A tag cloud is a collection of tags that have a different font size (Sinclair and Cardew-Hall [2008]). The more often a tag has been used by a user, the larger the text font. For instance, in Figure 3.6, we notice that wedding, California and Japan are frequently used tags. By clicking on one of the tags, we would get an overview of all the resources, in this case pictures, that have been tagged with this tag.

By having a look at the tag cloud of a community one can already have an idea about the topics which are frequently discussed. It can be considered to be a “a visual summary of the contents of the database” as results of interview questionnaires in Sinclair and Cardew-Hall [2008] indicated. Also, the interviews revealed that tag clouds are an interesting means for serendipity search: looking for information which is not very specific.

![Figure 3.6: Visualization of a tag cloud on Flickr (29 October 2009).](image)
3.3.5 Power law distribution function

In case of a broad folksonomy, we obtain a *power law distribution function* with a long tail effect when we plot the unique tags based on decreasing frequency (VanderWal [2005]; Mathes [2004]; Quintarelli [2005]; Halpin et al. [2007]).

Newman [2005] defines a *power law distribution function* as: “*When the probability of measuring a particular value of some quantity varies inversely as a power of that value, the quantity is said to follow a power law.*” Anderson [2006] defines it as a decreasing curve with a long tail at the end. There are several examples of distributions that follow a power law: people’s height, citations of scientific papers, word frequency in a text and so on (Newman [2005]). A characteristic of such a distribution is that when you plot the values on logarithmic axes you obtain a straight line as displayed on the figures below:

![Figure 3.7: power law distribution function (Newman [2005])](image)

A power law distribution function can be expressed as (with $\alpha$ = *exponent of the power law distribution*):

$$p(x) = Cx^{-\alpha}$$

or the equation can also be written as

$$\ln p(x) = -\alpha \ln x + c$$
Except for Halpin et al. [2007], the authors we mentioned in the beginning of this section did not test empirically whether a broad folksonomy follows a power law distribution function. Halpin et al. [2007] selected 500 popular bookmarks of del.icio.us that were annotated by around 2,000 users as well as 250 recent or less-popular bookmarks. For each of these bookmarks they selected the 25 most-used tags. They plotted the data of each bookmark on a logarithmic scale. The standard deviation of \( \alpha \) for the popular bookmarks was very small (std=0.03) whereas for the less popular bookmarks the standard deviation was much higher: 4.63. Therefore, the authors concluded that popular bookmarks follow a power law distribution function, but this is not really the case with less-popular bookmarks.

In addition, Golder and Huberman [2006] analysed how tags on del.icio.us evolve over time. They concluded that a convergence takes place after hundred annotations. More specifically, they noticed that after hundred annotations the frequency of a tag is almost a fixed proportion of the total number of tags used to annotate a resource.

### 3.3.6 Cognitive analysis of tagging

Sinha [2005] made a cognitive analysis of tagging by explaining how people think when they label a resource and how this differs from categorization. She distinguished two stages. In a first stage a number of related concepts are retrieved. When someone sees an object he automatically associates a number of concepts to the object. For instance, when you see a bottle of wine you will automatically come up with a number of concepts derived from your own knowledge and experience, such as *red wine*, *Shiraz*, *Bordeaux*, and so on. A selection of the associated concepts is made in a second stage. In the second stage you have to decide whether it is a bottle of Shiraz or Bordeaux wine. This selection process is called the *cognitive process*.

However, Sinha explains that categorization is far more difficult than tagging. When categorizing an object, the findability of that object becomes a fear. Humans have to spend
some time thinking whether the chosen category is the right one to retrieve the object at a later stage. She calls this the “post-activation analysis paralysis”. This difficulty does not exist in the case of tagging. The user should not be afraid of being unable to retrieve the resource as he can add several tags to it and thus put it into different categories at the same time.

### 3.3.7 Categorization and analysis of tags and tag usage

Golder and Huberman [2006] analysed the dynamics and tag use on the del.icio.us site. They concluded that users on the World Wide Web are more inclined to tag for personal use than for general community retrieval purposes (Golder and Huberman [2006]). However, the authors explain that personal tags (e.g. toread) can sometimes be useful to the whole community. For example, many people can say that they have toread a certain book when annotating a book at LibraryThing\(^\text{13}\). However, this is not always the case, for instance, a movie annotated with the tag funny will not always mean that the movie is funny for everyone.

\(^{13}\text{http://www.librarything.com}\)

They also identified a number of functions performed by del.icio.us (Golder and
Huberman [2006]). We paraphrase them below:

1. **Topics which are being discussed**: mostly nouns or proper nouns

2. **Publication format**: for example article, blog or book

3. **Name of the owner or creator of the bookmark**

4. **Refining categories**: compound words that contain numbers

5. **Adjectives which express opinions**: adjectives such as funny and scary

6. **Idiosyncratic tags**: for example mydog and myhouse

7. **Task-related tags**: tags related to performing a specific task such as buy and read

For his Master thesis Sterken [2008] used the categories proposed by Golder and Huberman [2006] to find the functions of tags in LibraryThing. In contrast to del.icio.us, where content that is directly available on the Web is labelled, a user can annotate his personal book collection. Sterken [2008] manually analysed a sample of 500 tags annotated to a set of 50 books that were obtained from the LibraryThing community and found that the tags on LibraryThing correspond to the categories proposed by Golder and Huberman [2006]. More specifically, he concluded “the tags mostly serve on the first four categories, to categorize and secondly on managing the resources”.

Sen et al. [2006] reorganized the categories proposed by Golder and Huberman [2006] into three more general groups: factual, subjective and personal tags. As Sen et al. [2006] wanted to classify the tags that people use on MovieLens to annotate movies, they reorganized the original categories of Golder and Huberman [2006]. They selected a sample of tags and classified these tags manually. Sixty-three per cent of the tags used on MovieLens are factual, three per cent personal, twenty-nine per cent subjective and five per cent could not be classified in one of the categories.

---


15 http://www.movielens.org
3.3.8 User incentives

Of course, the tags people use are probably biased by their underlying motivations\textsuperscript{16} as Marlow et al. [2006] reasons. Marlow et al. [2006] argue that the incentives to tag can be categorized in two groups: organization of information and social incentives. The authors, however, remark that in most cases users are influenced by more than one incentive at the same time.

In addition to being a possible means for ontology creation, we believe there are a number of other benefits related to tagging and thus folksonomies. We summarize and discuss each of them below:

- **Improving information retrieval:**

  Folksonomies are said to help users find information more easily since they can use their own vocabulary to annotate the resources (Hammond et al. [2005], Marlow et al. [2006]). Furthermore, folksonomies are also said to help people find information they would not retrieve when using a search engine (Begelman et al. [2006]). Hammond et al. [2005] and Golder and Huberman [2006] categorize the information retrieval incentives into two groups: tagging for themselves and tagging for the community. Hammond et al. [2005] classified social bookmarking tools in a diagram based on the kind of content creators (self or others) and the information retrieval purposes.

  Tag search, however, is limited because tags do not take semantical relations into account. Searching for a bookmark annotated with the tag *dog* would not take the bookmarks into account that are only annotated with the tag *Beagle*. Therefore, Begelman et al. [2006] suggested a clustering algorithm to improve information search. The algorithm tries to cluster similar or related tags.

  Hotho et al. [2006] proposed another solution to improve search in a folksonomy:

\textsuperscript{16}This was not tested by the authors. It was more like a logical assumption.
folkRank. FolkRank is an algorithm derived from page rank. In Hotho et al. [2006] they compared folkRank to a derived version from page rank, adapted page rank and concluded that folkRank is far better.

Morrison [2008] compared the information retrieval qualities of folksonomies on del.icio.us to search engines and taxonomies in terms of precision\(^\text{17}\) and recall\(^\text{18}\). Results from his research showed that search engines have the highest scores on precision and recall, but nevertheless folksonomies score rather well. Directories are more precise than folksonomies but their scores were similar on recall. Morrison [2008] suggests that the information retrieval quality of folksonomies can be improved by “better query handling”.

- **Social Navigation**

As most of the annotated resources and tags are public, a new kind of navigation becomes possible: social navigation (Morville [2005]). Tags and the annotated resources allow people to find other people with similar interests as we already explained in Section 3.3.3.

- **Broad classification of resources**

Brooks and Montanez [2006] examined whether similarities exist between blog entries tagged with the same tag. Out of their experiments, the authors of the paper concluded that tags are useful to describe broad content topics of blog entries, but not for a specific topic.

- **Social incentives**: Marlow et al. [2006] provides an overview of a number of social incentives. We summarize them below:

  - Get the attention from other members of the community by using common tags. Common tags are displayed in the tag cloud and this way other people

\(^{17}\)Definition *precision*: “the fraction of the returned results that are relevant to the information need”. (Manning et al. [2008])

\(^{18}\)Definition *recall*: “the fraction of the relevant documents in the collection that were returned by the system”. (Manning et al. [2008])
might look at the resources they have annotated with popular tags.

- Tags that are used to play a game. For example, the ESP game lets users annotate pictures with the tags. Each time both players use the same tag, they get a point.

- Leaving a mark by annotating a resource. For example, annotate an iPhone bookmark with the tag recentlybought.

- Using tags to express their opinion and share it with others within the community.

3.4 Corporate tags

As tagging generates many benefits for users on the Web, companies also started to think of the possible advantages it might generate for them. To discuss folksonomies in a company, some authors use the term enterprise folksonomies, such as Quintarelli [2005]. We prefer to use the term corporate tagging and will use it throughout this dissertation.

3.4.1 Benefits of corporate tags

At the time of research, not so many authors had already discussed the use of folksonomies in a company. Most of the publications were focused on how companies can use and implement bookmarking systems such as del.icio.us (Millen et al. [2006], Millen and Feinberg [2006], Damianos et al. [2006]) and some papers discuss how corporate tags can be used to find experts in the company (Farrell and Lau [2006], John and Seligmann [2006], Farrell et al. [2007], Schmidt and Braun [2008])\textsuperscript{19}.

\textsuperscript{19}A detailed explanation is given in Chapter 8.
Dogear

For instance, Millen et al. [2006] present Dogear, a social bookmarking system can be used in large organizations to let employees annotate their bookmarks from the corporate intranet and the World Wide Web. The system is based on a real authentication, thus the employees cannot use a pseudonym. Next, it is designed in such a way that it can be used behind the firewalls of the company. The Dogear system was implemented at IBM (Millen et al. [2006]).

A year after the implementation, Millen and Feinberg [2006] did a qualitative as well as a quantitative analysis of the Dogear system by analysing the log files as well as conducting a number of interviews. An analysis of the log files showed that community browsing for information annotated by the community was the most frequent way to search for information on the system. The second most frequent form of search was explicit search. Personal search was the less popular way to search for information.

It seems that the employees in this company were not mainly interested in their own annotated resources. Millen and Feinberg [2006] remark that this contrasts with the general idea of why people like to annotate resources. People are said to like to annotate in order to retrieve their own annotated resources; in this case they are also interested in the resources annotated by other community members.

However, we believe this can easily be explained by the fact that a corporate environment differs from a Web environment. To do their job properly, employees often need information from colleagues, in this case bookmarks annotated by colleagues.

Improve enterprise search

In Van Damme [2006], we argued that a folksonomy could help a company retrieve unstructured information stored in information systems. As most information systems do not contain a subject-based classification mechanism, we believed a folksonomy could be
regarded as a low-cost alternative. By making a proof of concept, we showed how easily a tagging mechanism could be integrated within SugarCRM\textsuperscript{20}, a Customer Relationship Management (CRM) system, to annotate notes written by the sales representatives. At that moment in time, SugarCRM did not have a built-in tagging functionality.

To make the proof of concept, we used an open source PHP plug-in Freetag\textsuperscript{21} that contains a code to store and retrieve tags within a MySQL database. As we wanted to have a more extended feedback functionality, we first extended the plug-in and then integrated it within SugarCRM. Recently, SugarCRM integrated tagging as a feature in the SugarCRM module.

Dmitriev et al. \cite{Dmitriev2006} suggest using explicit as well as implicit annotations to improve enterprise search. They consider annotating a web page as creating metadata for that resource. By asking a user to augment a web page, explicit feedback is gathered. Of course, it is time-consuming to ask this of every employee. Therefore, they propose to use the search query of employees as an implicit annotation. Four different scenarios to collect implicit feedback are proposed by analysing the log files which are created when one searches for information. Preliminary testing showed that explicit annotations were a good means for enterprise search. Analysing the explicit annotations revealed that employees often use abbreviations, opinions and descriptions of a web page.

Currently, little research is available on the use and analysis of corporate tags. In the sections below, we summarize some of the main issues that are discussed in these research papers.

\subsection{3.4.2 Analysis of corporate tag use}

Muller \cite{Muller2007} analysed tag usage across four different services in the company. Many users participated in more than one of these services and many of the unique tags were

\textsuperscript{20}http://www.sugarcrm.com/crm/
\textsuperscript{21}http://getluky.net/freetag/
applied in several services. Muller [2007] was interested to know how these tags were used over the different services. He concluded that there were more tags-in-common within a service than across different services. When doing the same test on the level of each employee and tag by comparing the group of users applying a specific tag over different services, he obtained the same results. The amount of tags-in-common over different services was in both situations low. He was surprised by the small overlap in tag vocabulary over the different services as they were all used in their daily work. Having a closer look at the tags themselves made clear that there were a lot of variations of tags which had a huge impact on the number of tags-in-common. However, he was not able to clean-up other problems such as semantic level problems. The results of the tag analysis allowed the authors to formulate a number of approaches to improve tag re-use: when annotating the resource, searching for the resource and storing the tags.

### 3.4.3 Incentives for employees to tag

In Section 3.3.8, we gave a number of incentives for users on the Web to tag. We believe most of these incentives are not applicable to the situation of a corporate environment. The users, their underlying motivations and the environment may be different. For example in a corporate folksonomy the user or employee is known and will not always tag voluntarily.

Thom-Santelli and Muller [2007] interviewed a number of employees to find out what their underlying motivations are to tag. Although most literature focuses on how users are interested to tag mainly for themselves, the interviews revealed that this is not the case in a corporate environment. Many respondents answered that they found it important to create tags that could be of interest to other employees. In such a way, the authors concluded that there is some kind of continuum between tagging for personal and community benefit. They described the phenomenon as the *wisdom of my crowd*. 
3.4.4 Social tagging roles

Thom-Santelli et al. [2008] researched the tag use that existed within enterprise systems. By means of semi-structured interviews they were able to distinguish five different social tagging roles that exist when people use tags:

- *Community-seeker*: creates tags to help other people that share the same field of interest
- *Community-builder*: tries to select tags in such a way that they become an added value for other members of the community.
- *Evangelist*: uses tags to link related content with one another
- *Publisher*: information creators that would like to extend the amount of information on a specific content item.
- *Small team leader*: analyses tags of other people in order to create a terminology that is understandable by every member of the community.

3.5 Conclusion

We provided a general overview on tagging. We positioned folksonomies, and thus tagging, against a taxonomy and concluded that folksonomies have their weaknesses but definitely have strengths.

As in this dissertation we investigate how an analysis of corporate tags can help a company to unlock the knowledge that resides in the organizational memory, we provided some related work on corporate tags. Until now, the number of publications on corporate tags has remained rather low, nevertheless tagging offers many advantages to companies.
Part III

Research
Chapter 4

Quality of tags in broad and narrow folksonomies

4.1 Introduction

As discussed in previous chapters, users can enter any words that enter their minds when they label resources. As a consequence, we may have to question the quality of tags.

Krestel and Chen [2008] make a distinction between two kinds of problems related to the quality of tags: the low semantic value and spamming.

Tagging is often criticized for having a low semantic value since it does not take into account synonyms and homonyms. Therefore it can have a negative impact on the quality of the tags (Krestel and Chen [2008]). In addition, Xu et al. [2006] also mention syntactic variances, for example running and run, which can also be considered as having low semantic value.

Spamming is the second problem Krestel and Chen [2008] consider. They define tag spam as tags that are intentionally used to give a wrong description to the resource. However, we believe that it can also happen unintentionally. In cognitive science, human
beings are said to be far better at analysing texts than machines because of the reasoning power (Sowa [1984]). But this does not imply that humans will always choose the most appropriate and relevant tags to describe a resource.

In a corporate environment, we assume we have to inspect the quality even more, since employees may not directly receive any kind of personal benefit when they create tags as is the case on the World Wide Web.\(^1\)

To investigate our main research question, we need to have a closer look at the quality of tags and, more specifically, the quality of corporate tags. Before we can measure the quality of corporate tags, we need to define the quality of tags and find techniques that will allow us to retrieve high-quality tags. As we will discuss in section 4.3, we need to make a distinction between broad and narrow folksonomies\(^2\) to study the quality of tags. In literature, however, no distinction between the quality of tags in broad and narrow folksonomies is made. Authors of papers on quality of tags implicitly assume a broad folksonomy. However, many examples of a narrow folksonomy exist, such as pictures annotated on Flickr or the dataset of the Company as we will discuss in Chapter 5.

Before, we investigate the quality of corporate tags, we first present some tag quality algorithms we applied on broad folksonomies in order to find high-quality tags. Because we did not have a dataset with corporate tags that belong to a broad folksonomy, as we will discuss in Chapter 5, we applied the algorithms to a del.icio.us dataset, an example of a broad folksonomy\(^3\). In case a company has a broad folksonomy the results we obtained on the Del.icio.us dataset could be interesting as we will discuss in Chapter 10. After we describe the dataset we obtained by the Company in Chapter 5, we discuss the quality of corporate tags in a narrow folksonomy in Chapters 6 and 7.

---

\(^1\)Overview incentives to tag: Chapter 3, Section 3.3.8.
\(^2\)Definition of broad and narrow folksonomies: Chapter 3, Section 3.3.2.
\(^3\)Chapter 3, Section 3.3.3.
CHAPTER 4. QUALITY OF TAGS IN BROAD AND NARROW FOLKSONOMIES

4.2 Quality of information and tags

Quality issues are not something new. They are already extensively discussed in other disciplines such as the information systems domain. An information system with poor quality information can have a huge impact on the organization (Kumar and Ballou [1998]). Suppose that an information system which contains the salary information of all the employees in the company does not have updated salary information. Paying employees incorrect wages would cause disorder in the organization.

In literature on information and data quality, many publications propose different quality dimensions. Ballou and Pazer [1985] identified four data quality dimensions:

- **Accuracy**: data has to be correct,
- **Completeness**: data must contain all the relevant information,
- **Consistency**: a uniform data format should be used, and
- **Timeliness**: date should be saved instantly.

Klischewski and Scholl [2006] proposed eight information quality dimensions that are similar to the ones proposed by Ballou and Pazer [1985], but more extended: accuracy, objectivity, currency, authority, assurance/reliability, relevance/precision/recall, timeliness and the perceived value.

We decided that the one of Ballou and Pazer [1985] are most applicable to the situation of tags. When we apply these dimensions to tags, we obtain the following definitions:

- **Accuracy**: tags should always be a reflection of the content of the resource. Spamming is not allowed,
- **Completeness**: tags should describe all the aspects of the resources they annotate,
• **Consistency**: a uniform format for tag use is needed. For examples no plural nor conjugated tags,

• **Timeliness**: tags should be created instantly after creation of the resources.

We consider all the dimensions to be applicable to the situation of the quality of tags, except for timeliness. Of course, time is an important parameter for tags. However, we believe it is hard to check the timeliness of tags as a quality dimension. It is not because someone bookmarks an article which has been written one year previously that the tags do not have a value anymore. Time can sometimes be an issue, for instance as we will explain in Chapter 10.

### 4.3 Parameters for tag quality

We decided to use the *accuracy*, *completeness* and *consistency* data quality dimensions to explain the definition of *high-quality tags*. In the paragraphs below, we discuss two parameters we believe are important when discussing the quality of tags: *implicitness-explicitness* and type of folksonomy. The definition of *high-quality tags* is explained in the next section.

#### 4.3.1 Implicitness-explicitness

We organize the three tag quality dimensions into two tag quality groups: *explicit* and *implicit* tag quality. We define *implicit tag quality* as being how well tags describe the resource in terms of the tag quality dimension *completeness* and *accuracy*. According to these tag quality dimensions, we need more information than merely the *tag itself* to check the implicit tag quality. This implicit tag quality is related to the other tags as well as the annotated resource.

This contrasts with the *explicit tag quality*. We describe explicit tag quality as being
the quality which measures the quality of the tags in terms of being consistent. We need to determine whether the tags obey a uniform format. To investigate the explicit quality of the tag we do not need additional data. Therefore, explicit tag quality is far more easy to verify than implicit tag quality. Also, explicit tag quality can easily be improved by cleaning up the tags as we will discuss in Chapter 6.

4.3.2 Type of folksonomy

Next, we believe the type of folksonomy should also be taken into account more specifically, when we investigate the implicit tag quality. As we explained in Chapter 3, we can distinguish two types of folksonomies: broad and narrow folksonomies. The latter has only one set of tags whereas the first type of folksonomy has many sets of tags. The difference in the amount of resource annotations does not have an impact on how to evaluate the explicit tag quality, because the explicit tag quality is related to the tags themselves as visualized in Figure 4.1. This is however not the case for the implicit quality of tags. The difference it makes is that in the case of broad folksonomies we have far more metadata created by the users and thus need different techniques to measure the implicit quality of tags than for narrow folksonomies.

Moreover, we believe that the principle of the wisdom of the crowds (Surowiecki [2004]) is applicable to broad folksonomies. This is also argued by Peters [2009] where she explains how collective intelligence is created when different people annotate the same resource.

Both terms, the wisdom of the crowds as well as collective intelligence, were brought into general use by the book of James Surowiecki. The principle of the wisdom of the crowds states “that solving a problem by a group of a hundred people will often be as good as by the smartest from the group”. Therefore, we state that a group of people that labels a resource is as good as letting the resource be annotated by the smartest member of the group (Surowiecki [2004]).
As a consequence, we believe that principle of the wisdom of the crowds is applicable to the case of the implicit tag quality of broad folksonomies. We assume that tags that correspond to the tag quality dimensions accuracy and completeness can automatically be retrieved by analysing the tags used by a group of people to annotate a resource. Of course, we need to know which technique is appropriate to do this. In this chapter, we propose a number of techniques which we tested on a dataset, as we will detail later on in this chapter.

This is, however, not possible in the case of a narrow folksonomy. We do not know whether the person who annotated the resource is the smartest of the group or has proper knowledge to do so. Therefore, we believe tags have to be analysed individually together with the resource to check their implicit tag quality. We assume that is very hard to let this process occur automatically as we will discuss in Chapter 7. However, in Chapter 7 we propose and test an approach to verify the implicit quality of tags automatically.

4.4 High-quality tags

We define high-quality tags as well cleaned-up tags that correspond to the three tag quality dimensions: accuracy, completeness and consistency. To determine high-quality tags, we first need to verify the explicit quality of the tags or check whether the explicit quality of the tags need to be improved. Then, we can measure the implicit quality of tags. Of course, we need techniques to determine whether a tag can be considered a high-quality tag, as visualized in Figure 4.1.

In this chapter, we present a number of algorithms to automatically select high-quality tags in a broad folksonomy. In Chapter 7, we determine high-quality tags in a narrow folksonomy: tags created by the employees of the Company.

The structure of this chapter is as follows. After we have given an overview of related literature on the quality of tags, we present a number of tag quality algorithms for broad
folksonomies. We tested these algorithms on a del.icio.us dataset and the results were evaluated by a group of students.

4.5 Related work

Until now, only a small number of publications discuss the quality of tags used on the World Wide Web.

The problem of low quality tags was first mentioned in Xu et al. [2006]; Guy and Tonkin [2006]. Xu et al. [2006] suggest *a number of criteria for high-quality tags* to improve future information recall:

1. *The number of times a tag is used to label a particular resource.* The more people use a certain tag, the less chance there is that a tag can be considered as a low quality tag.

2. *The number of tags used to describe a resource.* A small number of tags simplifies the information retrieval process.
3. **Internal tag cleaning**: Users can use whatever tag they like, but afterwards the tags need to be cleaned. Guy and Tonkin [2006] suggest giving a sort of tag education to the users to ameliorate the quality of the tags. Guy and Tonkin [2006] argue that the quality of tags can be improved if the users are trained to tag in a consistent way, for example by only using singular nouns and stemmed verbs.

4. **No idiosyncratic tags**: Personal tags cannot always be considered for a community-wide usage.

5. **The number of facets tags describe**: In order to provide a complete view of the annotated resources, tags should describe all the facets of the resource they label.

Next to a number of criteria, Xu et al. [2006] also introduce a reputation score to automatically detect someone who creates high or low quality tags. The reputation score is based on the criteria mentioned above. The higher the reputation score, the less someone is considered as a spammer. However, Xu et al. [2006] do not evaluate a tag individually as Krestel and Chen [2008] argue.

Krestel and Chen [2008] present the Tag-Resource Pair Rank algorithm to automatically detect tag spamming. They describe spammers as users who maliciously assign low quality tags to resources. This algorithm is an extension of the PageRank algorithm\(^4\) and the TrustRank algorithm\(^5\). The Tag-Resource Pair Rank algorithm measures the quality of the tags by analysing the relationship between the resources and tags in a graph. The graph consists of nodes (resources or tags) and edges. The weight of the edges is measured by the number of users who have assigned a certain tag to the resource. In order to measure the Tag-Resource Pair Rank for each node, a set of manually evaluated nodes is necessary to calculate the score of all the other nodes. By separating the tags assigned to a resource, they make a distinction between the quality of each tag resource combination.

\(^4\)The PageRank algorithm was introduced by Brin and Page [1998] to assign a weight to a web page based on the number of times other web pages are referring to this web page. Detailed information in Section 8.4

\(^5\)The TrustRank algorithm is an extension of the PageRank algorithm.([http://pagerank.suchmaschinen-doktor.de/trustrank.html](http://pagerank.suchmaschinen-doktor.de/trustrank.html), retrieved 11 April 2010.)
In contrast, the authors in Sen et al. [2007] as well as in Lee and Han [2007] propose to extend tags with a kind of rating mechanism. In Lee and Han [2007] it is suggested that the user tags a resource with neutral tags and/or tags extended by a context to the tag that is either positive (liked by people) or negative (not liked by people). The positive or negative context is expressed by assigning a positive (+) or negative (-) sign to the tag. For instance, a picture of a Ferrari could be tagged as car, expensive(-), fast (+). When counting the tags as well as all the positive and negative signs, a global overview of the tagged resource can be obtained.

Different tag rating scenarios with implicit as well as explicit mechanisms are tested and discussed in Sen et al. [2007]. The authors conclude with a number of design guidelines for the creators of websites that embed a tag mechanism so that only the tags which the community prefers are displayed.

Next, Sen et al. [2009] want to know which tag selection algorithm, as they define it (an algorithm that tagging systems use to select and order tags), is the best algorithm to display high-quality tags. They explain that not everyone will like to see the same tags. First, they asked a group of users to evaluate a number of tags added to a resource. They combined the features of these annotated resources into a machine learning classifier. For each classifier they calculated the probability that a certain tag was a high-quality tag. They compared the results from the classifiers with those from the offline survey by using a number of metrics. Results showed that the item-top-n metric is the best metric to use when evaluating tag selection algorithms. Based on online as well as offline analyses, they concluded that tag selection algorithms that combine implicit features (properties related to the tag itself) with explicit features (ratings, in this case thumbs up and down) are the best algorithms to select the high-quality tags. They concluded that tagging sites with relatively low activity benefit from adding a thumbs up/down rating mechanism.

---

6 The authors use high-value tags instead of high-quality tags.
7 Sen et al. [2009] define this metric as “the metric which measures the top-n tags tag have a high value.”
4.5.1 How our research differs from existing research

Although some publications have already been written on the quality of tags as discussed in the related work section, we believe the research we discuss in this chapter as well as Chapters 6 and 7 is a valuable contribution to the domain.

At the moment of research, we even did not find any paper that investigated the quality of corporate tags. We must also remark that none of these publications mention that a difference should be made between narrow and broad folksonomies when evaluating the quality of tags. Most of them implicitly assume a broad folksonomy, a folksonomy where the resources are tagged by more than one person, whereas the opposite folksonomy also exists, as we already explained in the previous Chapter.

Next, the algorithms they present were not always tested, such as in the case of Krestel and Chen [2008]. Krestel and Chen [2008] presented one algorithm based on trust rank and page rank, but did not validate their approach by asking the users for feedback. Sen et al. [2009] analysed which tag metric can be used to evaluate the most appropriate tag algorithm. They also concluded that tag selection algorithms that combine implicit as well as explicit features are the best to select high-quality tags. Their research, however, has been done at a later period of time and differs from ours because we did not take implicit features into account to measure the quality of the tags.

4.6 Algorithms to find high-quality tags in broad folksonomies

In this section, we present three tag quality algorithms for broad folksonomies that are largely based on a paper we presented at the i-semantics conference 2009 in Graz, in
We tried to find an algorithm\textsuperscript{9} which automatically selects high-quality tags in a broad folksonomy.\textsuperscript{10}

Therefore, we present three possible tag quality algorithms that we derived from literature on information theory. For each tag quality algorithm, we give a description, explain how it can be calculated, and put forward why we propose it. As a prerequisite to use the algorithms, the explicit tag quality has to be verified before we can use these algorithms as we will detail later on.

Since the tag dataset of the Company belongs to a narrow folksonomy, as we will discuss in Chapter 5, we were not able to test it on this dataset. We decided to apply the algorithms to a del.icio.us dataset. In cases where companies have a broad folksonomy, the results we obtained might be of interest for them, because we assume that the principle of the wisdom of the crowds in broad folksonomies is also applicable to the situation of corporate broad folksonomies.

### 4.6.1 Tag quality algorithm 1: high-frequency tags

1. **Description** We select tags with the highest $n$ frequencies.

2. **Calculation** For each tagged resource, we order the tags and count the frequency of each distinct tag. We then choose the tags with the highest $n$ frequencies.

3. **Motivation** Currently, tagging systems apply simple heuristics for ranking tags, for example when visualizing tags. On most websites, tags are visualized in so-


\textsuperscript{9}Here we use algorithm instead of metric, because it is a more appropriate term.

\textsuperscript{10}In the paper, we originally defined them as intersubjective tags, the degree to which a tag is understood by many members of a group or describes a resource in such a way that it corresponds to the living world of all the members of the community. Intersubjective tags can be considered as tags that correspond to the accurate and complete dimensions of tag quality.
called tag clouds. The text font of the tags in the cloud is an indication of their frequencies: the more frequently used, the larger the text font. Tags which have a low frequency are considered to be less important. To maintain a good visual representation, only tags which exceed a certain frequency threshold are included in the cloud (Sinclair and Cardew-Hall [2008]). This is consistent with the first theory on automatic text indexing formulated by Zipf [1949]. Zipf [1949] stated that determining the frequency of words in a text can be used as a technique to detect relevant keywords for a document. In the case of broad folksonomies, many people express their thoughts about a resource through their selection of tags. This indicates that a cognitive filtering mechanism exists which detects significant from non-significant words.

4.6.2 Tag quality algorithm 2: tag agreement

1. **Description** We define tag agreement for resource \( x \) as the tags that are selected by more than 50\% of the users who have tagged resource \( x \).

2. **Calculation** We first determine the frequency of each unique tag. Then, we calculate the number of users who have tagged each resource. The tag agreement is consequently calculated by dividing the tag frequency by the number of users who have tagged a resource and then multiplying the result by 100 in order to express the result as a percentage. When all the users agree on a certain tag, this number should be equal to 100\%. The closer to 0\%, the less the users agree on that particular tag.

3. **Motivation** Decisions in various areas of human activity are often taken on the basis of a majority: more than 50\% of the people have to agree on a certain proposal in order for it to be accepted. Tagging in the case of a broad folksonomy can be seen as a way of voting for the semantic labelling of a resource and this is why we suggest tag agreement as a second tag quality algorithm.

\[\text{Definition tag clouds: Chapter 3, Section 3.6.}\]
4.6.3 Tag quality algorithm 3: TF-IRF

1. **Description** For each tag we calculate its Tag Frequency Inverse Resource Frequency (TF-IRF) weight and select tags with the highest $n$ TF-IRF scores.

We derive the TF-IRF tag quality algorithm from Term Frequency Inverse Document Frequency (TF-IDF), a common tag quality algorithm in the domain of automatic indexing to find descriptive keywords for a document. When selecting the appropriate tags for a certain document, the TF-IDF formula takes the intra- as well as inter-document frequency of keywords into account, because Salton and Yang [1973] explain that the frequency of a word in a document is not the only indicator of whether or not it is a valuable keyword, which also depends on how often it is used in a set or corpus of similar documents. For example, a collection of documents on Unix applications will probably contain the word *Unix* very often and will therefore not be very descriptive for these documents. However a document on *sam2p*, a command line Unix application to convert images and other kinds of file types into PDF files, will contain the word *sam2p* very often in this document and rarely in the corpus. In this way, it will have a high TF-IDF weight.

The higher the TF-IDF weight, the more valuable the keyword (Salton and Yang [1973]). The formula is calculated by the formula given in the equation below (with $n =$ frequency of the word in the document, $N =$ total number of words in the document, corpus = collection of documents, and $D_n =$ document frequency of the word).

$$TFIDF(word) = \frac{n}{N} \times \log\left(\frac{|corpus|}{D_n}\right)$$

The word’s document frequency is multiplied by the logarithm of its inverse document frequency in the corpus. The second part of this expression computes how common the term is in the corpus (Salton and Yang [1973]). A high TF-IDF weight...
indicates a good descriptive document keyword: one that is rarely used word in the corpus and frequently used in the specific document.

2. Calculation

- **Corpus:** To calculate the TF-IRF weights, we need a corpus of similar resources. This can be obtained through clustering and it is suggested that the Markov Clustering (MCL) algorithm\(^\text{12}\) developed by Van Dongen [2000] should be used, as results in Van Dongen [2000] show that the MCL algorithm is very good for clustering graphs.

To apply the algorithm, we need to visualize the relationship between tags as graphs. We suggest that the co-occurrence\(^\text{13}\) is calculated, because it is an often-suggested analysis technique in literature on folksonomies (Specia and Motta [2007], Schmitz [2006]) to find relationships between tags. For each tagged resource all the tag pairs are determined. The tie strength between a tag pair is increased each time two tags are used together in other annotated resources. By transforming the pairs (as nodes) and their co-occurrences (as weighted edges) into a graph, we can apply the Markov Clustering (MCL) algorithm (Van Dongen [2000]).

- **Calculating TF-IRF:** In order to transform TF-IDF into TF-IRF, we have to make some adjustments to the formula. We have to exclude the textual information or documents from our formula since tagged resources are not always textual (e.g. an mp3 audio file). The only data we can analyse are tags. As a consequence, we suggest the equation below to calculate the TF-IRF weight for a certain tag annotated to a resource. The formula is based on TF-IDF (with \(t_{x,y} = \text{frequency of tag}_x \text{ for resource}_y\), \(T_y = \text{total number of tags for resource}_y\), corpus = sum of resources, and \(R_x = \text{sum of resources that have tag}_x\)).

\[^{12}\text{It is a clustering algorithm based on Markov Chains. A Markov chain is a stochastic matrix where the future state depends on the current state of the system. (Guerry et al. [2004])}\]

\[^{13}\text{Definition of co-occurrence: Glossary.}\]
CHAPTER 4. QUALITY OF TAGS IN BROAD AND NARROW FOLKSONOMIES

\[ TF - IRF(tag_{x,y}) = \frac{t_{x,y}}{T_y} \times \log\left(\frac{|corpus|}{R_x}\right) \]

3. Motivation In the domain of automatic indexing a lot has been written on how to select the most appropriate keywords. Research on automatic indexing dates back to the 1950’s and consequently represents a large body of knowledge. We believe it is interesting to apply TF-IDF, which is one of the common techniques in this area, to broad folksonomies.

4.6.4 Choose an appropriate value for \( n \)

We suggest that \( n=5 \) is used for each tag quality algorithm. Although it is said in literature that humans are able to recall up to 7 items in short-term memory locations (Miller [1956]), we decided not to select \( n=7 \) because the average number of tags people use, for example to describe their bookmarks, is much lower, around 2.5. Calculating the average of those two numbers gives \( n=5 \).

4.6.5 Dataset

For the analysis we used the del.icio.us\(^{14}\) dataset from Laniado et al. [2007]. Laniado et al. [2007] created a script to scrape 3,400,000 unique bookmarks from around 30,000 users and the corresponding metadata\(^{15}\).

Preparatory steps

To compare the results obtained by applying the different tag quality algorithms on the tag dataset, we had to calculate each tag quality algorithm on the same collection of bookmarks. Since the TF-IRF tag quality algorithm requires the creation of a corpus or

\(^{14}\) A detailed description of Del.icio.us: Chapter 3, Section 3.3.3.
\(^{15}\) The data was retrieved in March 2007.
a set of related resources, in this case bookmarks, we needed to create the corpus on a cleaned tag dataset before we could apply the algorithms.

In the paragraphs below, we discuss the different steps used to obtain the tag sets that result from the different tag quality algorithms. As a prerequisite to apply the algorithms to the dataset we needed to verify the explicit tag quality.

1. Cleaning

The data cleaning approach we present here clearly differs from the one we will present in Chapter 6, because we applied it on a Del.icio.us dataset that mainly contains English tags. Next, the research we present in this chapter was done at an early stage of our research and we did not have an extended cleaning methodology at that time.

- We removed all the English stop words\(^\text{16}\) from the tag set, since most of the high-frequency tags of the dataset are in English.
- We stemmed\(^\text{17}\) the remaining tags by removing the suffixes. Words that have the same stem or root are considered to be referring to the same concepts (for example running and run have the same stem, i.e. run).
- We merged duplicate tags since a duplication of tags appears after stemming.
- We disregarded all bookmarks that were tagged by less than 100 users, because we only wanted to include bookmarks that were evaluated by a large number of users. Next, empirical research in Golder and Huberman [2006] showed that a kind of tag stabilization takes place after the first 100 bookmarks. This restriction reduced the number of bookmarks in the collection to 3,898.

\(^{16}\)We used the list which is available on: [http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words](http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words) (retrieved April 2008).

\(^{17}\)We used the Porter stemming algorithm for English as implemented in [http://search.cpan.org/dist/Lingua-Stem/lib/Lingua/Stem/En.pm](http://search.cpan.org/dist/Lingua-Stem/lib/Lingua/Stem/En.pm). Definition stemming: Chapter 6, Section 6.2.
CHAPTER 4. QUALITY OF TAGS IN BROAD AND NARROW FOLKSONOMIES

- We calculated the co-occurrence\(^{18}\) of all the tags pairs we could retrieve in the
tag set of 3,898 bookmarks

2. **Application of MCL algorithm** We applied the MCL algorithm\(^{19}\) on all tag pairs
and their corresponding frequencies obtained in the previous step. We excluded the
tag pairs with frequencies of less than 100. This means that both tags have been
used less than 100 times together to tag a particular resource. A lower threshold
value did not result in clearly distinguishable clusters. We opted for the cluster with
commonly used terms containing the following tags:

\[(entertainment, film and movie)\]

We decided to include a bookmark in the corpus if it had at least one tag with a
frequency of 10 that belongs to this cluster. We opted for a number of 10 since we
wanted to be sure that a link with the cluster existed. As a result, we obtained 127
bookmarks for this cluster.

3. **Application of Three Tag Quality Algorithms** Then, we applied the three algo-
rithms on the tag dataset of the 127 bookmarks that are in the corpus. For each
algorithm, we ordered the tags from the left to the right based on decreasing values.
Some examples of the results are included in Table 4.1.

### 4.6.6 Results

We noticed a close linkage between the tag sets obtained by the first and third tag quality
algorithms. In some cases, the high-frequency and TF-IRF tag quality algorithms only
differed in the order of the tags. In the other cases, there was a close overlap between tag
quality algorithms 1 and 3 because they often shared similar tags.

\(^{18}\)As implemented in http://search.cpan.org/ allenday/Math-Combinatorics-
0.09/lib/Math/Combinatorics.pm

\(^{19}\)We used the code that Stijn Van Dongen made freely available for academic purposes at
http://www.micans.org/mcl/scripts/minmcl
Table 4.1: Tag sets obtained by applying the algorithms

<table>
<thead>
<tr>
<th>URL</th>
<th>Tag Quality Algorithms</th>
</tr>
</thead>
</table>

Table 4.2: Results by applying the algorithms

<table>
<thead>
<tr>
<th></th>
<th>Number bookmarks which follow rule</th>
<th>Average number tags/bookmark</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HQ</td>
<td>101</td>
<td>5</td>
<td>91</td>
</tr>
<tr>
<td>TA</td>
<td>75</td>
<td>0.94</td>
<td>11</td>
</tr>
<tr>
<td>TFIRF</td>
<td>101</td>
<td>5</td>
<td>71</td>
</tr>
</tbody>
</table>

When applying the tag agreement algorithm on the tag dataset of the corpus, we noticed that the average number of tags per bookmark where agreement exists was very low as we can see in Table 4.2. The minimum and maximum values lay between 0 and 3, and the modus and median both had a value of 1. It was therefore not possible to select 5 tags for the tag agreement algorithm since there were on average 0.94 tags per bookmark that corresponded to the definition. There were even 26 bookmarks that did not have any tags confirming to this pattern. After excluding these 26 bookmarks, the mean increased just slightly to 1.18.

4.6.7 Preliminary evaluation

To answer the question of which algorithm generates the best results, we decided to set up an online survey\(^{21}\). To conduct the survey we created a tool in PHP that chooses a bookmark as well as its tag sets randomly from the MySQL database. In each session 10 bookmarks had to be evaluated. There were 101 bookmarks in the database, because we

\(^{20}\)HQ = high-frequency tag quality algorithm ; TA = tag agreement tag quality algorithm; TFIRF = TF-IRF tag quality algorithm

\(^{21}\)The online survey took place in March 2007.
excluded the 26 bookmarks that did not have any tags for the tag agreement algorithm. Before people could participate in the survey, they had to be given a password to access the site and the survey only started after name and age were entered.

On each page a random bookmark from our database was displayed with the corresponding tag sets as shown in figure 4.2. To prevent a user from leaving the survey when scrolling the bookmark, we used iframe, an HTML tag allows another web page to be displayed within the web page.

Since the corpus of bookmarks was based on the tags (entertainment, film and movie), we believed there should be a close connection between the topics discussed on the bookmarks and the interests of a group of international students aged around 20.

After we gave the students a one hour presentation on tagging and folksonomies and introduced them to Del.icio.us, we invited them to a computer room to participate in the survey. We asked them to select the tag set which they considered did the best job of describing a specific bookmark, that is, the one which best corresponds to our definition of high-quality tags. In case of doubt, we told them to take the order of the tags into account. Indeed, a tag placed at the beginning is more important than one located more to
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

the right, as noticed by Golder and Huberman [2006].

Due to randomness, 75 of the 101 bookmarks were evaluated and some of them were assessed several times. In total, bookmarks were evaluated 173 times. We did not obtain the logical number of 200 (20 students doing 10 evaluations), since (1) some of the websites were down during the survey and had to be removed from the result list and (2) not all the students pursued the survey to the end.

In the sample, the students opted in 52.6% ($p_{HF} = 91/173$) of the cases for the high-frequency tag quality algorithm and in 41% of the cases for the TF-IRF tag quality algorithm ($p_{TFIRF} = 71/173$). The tag agreement tag quality scored poorly: in 6.4% ($p_{TA} = 11/173$) of cases they selected this option. A possible explanation for this might be the low number of tags the tag agreement algorithm generated. We tested whether the proportions$^{22}$ from the population can be considered to be equal.

$$\begin{align*}
H_0 & : \pi_{HF} = \pi_{TA} = \pi_{TFIRF} \\
H_1 & : \pi_{HF} \neq \pi_{TA} \neq \pi_{TFIRF}
\end{align*}$$

Results from the Chi-square test$^{23}$ in Figure 4.3 show that these proportions, cannot be considered to be equal ($p$-value=0.000, $\alpha = 0.05$).

<table>
<thead>
<tr>
<th></th>
<th>Observed N</th>
<th>Expected N</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91</td>
<td>87.7</td>
<td>-3.3</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>57.7</td>
<td>-46.7</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>57.7</td>
<td>13.3</td>
</tr>
<tr>
<td>Total</td>
<td>173</td>
<td>87.7</td>
<td></td>
</tr>
</tbody>
</table>

Test Statistics

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>df</th>
<th>Asymp. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.012E1</td>
<td>2</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 57.7.

Figure 4.3: Results SPSS Chi-Square

Based on the results of the descriptive statistics of the sample, we noticed that the

$^{22}$We use $\pi$ as a notation for the average proportion in the population.

$^{23}$The basic assumptions for a Chi-square test are fulfilled: the expected frequencies, $F_i^0$, are larger than 5.
CHAPTER 4. QUALITY OF TAGS IN BROAD AND NARROW FOLKSONOMIES

high-frequency tag quality algorithm was the most often chosen tag quality algorithm ($p_{HF}=91/173$). Therefore, we tested whether the proportion of the high-frequency tag quality algorithm in the population ($\pi_{HF}$) is larger than 0.5.

\[
\begin{align*}
H_0 & : \pi_{HF} \leq 0.5 \\
H_1 & : \pi_{HF} > 0.5
\end{align*}
\]

To test this hypothesis, we used a binomial test. Since a binomial test is only allowed on two categories, we merged the results from the two other tag quality algorithms into one group. Results from doing a binomial test\(^{24}\) in Figure 4.4 show that we have to accept $H_0$ (p-value $= 0.2715$, p-value $> \alpha$, $\alpha = 0.05$). Therefore, we cannot conclude that the proportion of the high-frequency tag quality algorithm in the population is larger than 0.5.

![Descriptive Statistics Table](image)

<table>
<thead>
<tr>
<th>Category</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>V2 Group 1</td>
<td>1</td>
<td>91</td>
<td>53</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>V2 Group 2</td>
<td>0</td>
<td>82</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V2 Total</td>
<td>173</td>
<td>100</td>
<td>53</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Based on Z Approximation.

Figure 4.4: Results SPSS Binomial Test

The results from the sample show that $p_{TA}$ was low. Therefore, we tested whether the proportions from the high-frequency tag quality algorithm and the TF-IRF in the population are together larger than 0.66.\(^{25}\) Again, we did a binomial test and we tested the hypothesis below:

\[
\begin{align*}
H_0 & : \pi_{HFTIRF} \leq 0.66 \\
H_1 & : \pi_{HFTIRF} > 0.66
\end{align*}
\]

\(^{24}\)The basic assumptions for a binomial test were fulfilled: $np > 5$ and $n(1 - p) > 5$.

\(^{25}\)We opted for a value of 0.66, because we wanted to research whether the proportion of those two algorithms together is larger than the third algorithm.
Figure 4.5: Results SPSS Binomial Test

Based on the results we obtained in Figure 4.5, we rejected $H_0$ ($p$-value = 0.000, $\alpha = 0.05$) and concluded that the proportion of the high-frequency tag quality algorithm and the TF-IRF tag quality algorithm together ($p_{HFTFIRF}$) is larger than 0.66. Those two algorithms are more used together than the other algorithm.

However, we are not able to say which of the two algorithms is better. As already discussed in the previous paragraph, there was a close linkage between the tag sets obtained by these two algorithms. Therefore, the experiment should be repeated on a larger group of people by focusing solely on these two quality algorithms techniques.

4.6.8 Limitations of the research

Although we obtained first preliminary results, there were certain limitations that did apply to the online survey.

- We did not ask the students why did they opted for a certain tag set.
- Number of participants was low.

4.6.9 Conclusion

In this chapter, we defined high-quality tags and explained that a distinction should be made between explicit and implicit tag quality, as well as the type of folksonomy. We
presented three algorithms to select high-quality tags in broad folksonomies. We applied them on a del.icio.us dataset and concluded that from the three suggested algorithms, the high-frequency tag quality algorithm and the TF-IRF tag quality algorithms were the most appropriate algorithms to select high-quality tags in broad folksonomies.
Chapter 5

Description of the case study and dataset

5.1 Introduction

To investigate our research questions, we used the dataset of the Company. Currently, the Company is mainly using its tags for information retrieval purposes as we will discuss in Section 5.2. The goal of our research was to investigate how knowledge can be retrieved through an analysis of employees’ tags. As we refer to the dataset in the following chapters, we describe the structure and data elements of the dataset in Section 5.3. In the Sections 5.4 and 5.5, we detail all the difficulties we had with the dataset and elaborate how we handled the large amount of data as well as all the tools we used. An overview of the descriptive statistics of the dataset is given in Section 5.6.
5.2 Tagging in the company

The Company has been tagging all their communication messages for more than 20 years. Messages such as letters and faxes that are not sent electronically are manually scanned, tagged and archived into the database. Tags replace the subject line of the message. Tagging is completely integrated in the corporate culture. The messages can be created manually, automatically and semi-automatically. The automatic and semi-automatic messages have default tags. In case of semi-automatic messages, the author has to add additional tags or replace the default tags. Manually created messages require user-created tags.

Tags have two main functions in the Company: information retrieval and an aid to decide whether or not to read a message. Initially, tags were introduced to solve the information retrieval problem since the Company did not have full text search engines at the time. Until today, tagging is still a part of the communication messaging system to retrieve the messages. However, the ambiguity of the flat tags and the information overload obstruct the search process. The Company introduced the following tag rules to improve the quality of the tags:

- Tag rule 1: Singular nouns
- Tag rule 2: No conjugated verbs
- Tag rule 3: No conjugated adjectives
- Tag rule 4: No abbreviations
- Tag rule 5: At least 6 meaningful tags (no articles or adverbs)
- Tag rule 6: Dates have to be written as YYMMDD

---

1This section is partially based on a section of: C. Van Damme, T. Coenen, T and E. Vandijck. Turning a corporate folksonomy into a lightweight corporate ontology. In Abramowicz, Fensel (Ed.): Proceedings of the 11th International Conference on Business Information Systems (BIS 2008), Innsbruck, Austria, LNBIP 7, pp.36-47. Springer, 2008.

2Since tagging is completely integrated in the corporate culture of the Company, we do not discuss the issue of encouraging employees to tag. Except in Chapter 10, we give some recommendations on how employees can be encouraged to tag.
CHAPTER 5. DESCRIPTION OF THE CASE STUDY AND DATASET

- Tag rule 7: Month and year have to be one tag (for example march2007 or apr09)
- Tag rule 8: Combinations of letters and digits have to be written as one word (for example phase1)
- Tag rule 9: Simple words (no combination of words) (for example for ‘premeeting’ just use ‘meeting’).

As we will discuss in Chapter 6, these rules are not always followed. Next to retrieving messages, employees use the tags to decide whether or not they will read the message. The employees are swamped with messages each day. Hence, they use tags to decide whether a certain message may be of interest to them.

The Company gave us the metadata of all their non confidential messages created in 2006. Due to confidentiality reasons, we only got the content of a few messages for the research we discuss in Chapter 7.

5.3 Data elements in the dataset

In this section, we give an overview of all the elements in the dataset. The set contains more data than just tags. It is important to provide an overview as we often refer to one of these data elements throughout the dissertation.

1. Message-ID. Each message has a unique identifier or ID\(^3\). Each ID consists of one letter, the first character of the Message-ID which represents the message type, and 7 digits.

2. Tags or freely chosen keywords of the employees: the tags are chosen by the author of the message. Each character of the tag must be alphanumerical. Everyone who has read the message can add tags to improve the retrieval process of this message. However, there is no unique identifier in the dataset available which indicates

\(^{3}\text{We will use the term ID as synonym for identifier in the remaining part of this dissertation.}\)
whether tags have been added to the message by the reader. Since we were told that employees are not doing this very often, we assume that the tags correspond to the ones originally assigned by the creator of the message.

3. **Tag line number** Tags are not stored separately. Instead, they are saved all together in one field. When the set of tags exceeds a limiting number of characters the other tags are stored in a field on another row. To organize the set of tags of each message, a tag line number is used.

4. **Employee-ID.** Every employee has an unique employee number that consists of 5 characters.

5. **Dactylo-ID.** The dactylo is the person who has typed and thus created the message. In most cases this is also the author of the message. The dactylo-ID matches with the first 4 characters of the employee-ID of the dactylo.

6. **Volgnummer Dactylo-ID.** It contains only one digit. In most cases it is 0 or 1, but in some situations it has a different value. The concatenation of the Dactylo-ID and Volgnummer Dactylo-ID corresponds to the employee-ID.

7. **Role.** Different roles can be assigned to a message: author, receiver or carbon copy

   - **Author or Afzender.** Normally, the author and creator of the message are one and the same person, but this is not always the case. Sometimes the author records his message and asks a secretary or dactylo to type the message for him. This makes the secretary or dactylo the creator of the message.

   - **Receiver or Bestemmeling.**

   - **Carbon Copy or Kopiant.** It entails the same function as someone who is addressed in the CC field of an email.

Messages sometimes have more than one author or no author at all. We describe a number of situations where one of these situations occur:
• When a customer sends a message to the company, no author is added to the message.

• When employees reply to a message, the employee-ID is added to the list of authors and therefore the message has more than one author.

8. **Message type.** Every message belongs to a certain kind of message type. Below we summarize the frequently used message types:

• Note: is used for longer messages, mostly used to report a meeting.

• Email: incoming and outgoing email

• Quick Internal Message or **SNIV** as they call it in the Company, is used to send someone quickly a brief message. The CCs or **Kopianten** receive these messages only three times a day.

• Fax.

• Letter.

9. **Member-ID.** Every message has a member ID. A member ID can be considered as a classifier: it categorizes a message into an automatically or manually created message.

10. **Creation date.** The date when the message was created and archived into the system.

11. **Confidentiality.** There are three options possible, a message can be

- *not confidential*: the message can be read by everyone in the company;

- *confidential with limited access*: only the tags are visible to everyone

- *have limited access*: the message as well as the tags are only accessible by the author, receiver or carbon copy

We must remark that our dataset only contains meta data on non confidential messages and confidential messages with limited access.
12. **Language** of the message. This corresponds to the language code of the author of the message, in most of the cases Dutch. The author can always change the language when typing the message in another language, but mostly the employee forgets to change this. The result is that some messages are polluted and contain French and English tags.

Furthermore, every employee is part of one or more groups, called functional groups in the company. Functional groups are created based on job functions, projects or other similarities between employees. For each employee, we know the functional groups he or she is part of. We have the following data on functional groups:

- **Functional Group-ID.** Every functional group has an ID that consists of 3 characters.
- **Name.** The name of the group is a collection of words that describe the group.
- **Employee-ID** of the member of the group.
- **Date.** It has a default value of 00-00-00 (MM-DD-YY). When an employee has left the group or the company, the date when he or she has left is filled in.

### 5.4 Tools

The size of the text files, 2.28 Gigabyte of data, forced us to opt for a database that could perform queries on large sets of data.

Since confidentiality was an important issue, we preferred to have a system that was secure and could only be accessed by authorized people. All these features are handled by the MySQL database server (Allen [2004]) that uses SQL as a standard query language. SQL is a standard query language that allows you to interact with the database server and as a consequence manage the database by using a set of statements such as `SELECT`,...
CHAPTER 5. DESCRIPTION OF THE CASE STUDY AND DATASET

INSERT, DELETE (Williams and Lane [2004]). The fact that MySQL is an open source database server with a very large user community is an additional reason to choose this system. There are online tutorials as well as forums freely available that can help us whenever we encounter a problem.

However, when we tried to import the text files into the MySQL database server we ran into a number of problems that were too hard or impossible to handle through SQL. Therefore, we needed an additional programming language that makes data manipulation on large text files possible on an ordinary computer. Since Java caused many performance problems when doing some simple tests on the dataset, we quickly decided to choose another language that is more trustworthy to handle large text files: Perl. Perl is a Unix scripting language that was developed in the late 1980s by Larry Wall and is often mentioned in text mining literature because of its text pattern matching feature called regexp or regular expression. For instance you can use regular expression to search for substrings or words that only contain digits (Bilisoly [2008]). It is an interesting feature that we can definitely use to analyse the tags.

As the non-disclosure agreement we have with the Company does not allow us to store the dataset on a computer that is connected to a network, we could not use grid computing to speed up the processing time when performing some queries on the dataset. Grid computing permits the running of one application on several computers and decreases the processing time a lot.\footnote{http://www.gridcomputing.com/} This implied that we had to run all the queries and software we developed on our own ordinary computer which was not connected to a network. At a certain moment in time we decided to switch to a Linux operating system, Ubuntu, to overcome our limitations. We will explain this later in the next section.
5.5 Problems dataset

The problems we encountered can be classified into three groups: the text files themselves, the size of the database, and the structure of the data in the text files. We will discuss each of them in the paragraphs below.

5.5.1 Loading text files into the database

We had a problem when reading the original text files into the database. The original text files were probably saved on a Unix operating system whereas we first installed the MySQL database on a Windows operating system. The operating systems Unix and Windows use different techniques to end a line in a text file. Whereas Unix uses a line feed, Windows indicates the end of a line with a line feed as well as a carriage return. As a consequence, loading the text file saved onto a Unix machine into our database caused a bad data import. We solved this problem by specifying the ending character of each line in the text file when loading them into the database tables.

5.5.2 Size of the dataset

After solving the first problem, we quickly encountered another one: the size of the database. The database contained the metadata of 8,011,111 messages, which took a lot of processing time. Even simple queries such as a `SELECT` took too much time on an ordinary computer with Windows as the operating system. To overcome this problem we decided to test the performance of MySQL on a Linux operating system, specifically Ubuntu. We quickly noticed that we could improve the performance considerably by letting it run on a Linux operating system, and decided to switch from a Windows operating system to Ubuntu. The fact that the default security levels on a Unix operating system such as Ubuntu are much higher compared to those of Windows was a further motivation to switch operating systems.
5.5.3 Problems with structure of the dataset

The original structure of the dataset forced us to do some data manipulation. There were a number of other problems within the dataset which we discuss below.

- **Tags annotated to a message.** Since the tags of a message were saved as a set of tags on one or more rows, we needed to reorganize the structure of text files that contained the tags annotated to a message. To analyse the tags in a later stage, we needed to split the tags and store them in a separate field. This means we had to split each row (message ID and tags) into as many rows as there are tags for a message. As the message ID and tags were separated by a tab delimiter and the tags themselves by a white space, we developed software that split the tags and put them on separate rows, each containing the message ID and a tag.

- **Employee-ID and dactylo-ID.** When we wanted to select a message where the employee-ID of the author equalled the dactylo-ID, we encountered a problem. As already explained in a previous section, the dactylo-ID consists of four characters whereas the employee-ID has five characters. To overcome this problem, we needed to develop software that merged the dactylo-ID with the Volgnummer Dactylo ID in order to make a comparison with both IDs possible.

5.5.4 Manual and semi-automatic messages

As already explained at the beginning of this chapter, the dataset contained three different kinds of messages: manual, automatic, and semi-automatic. An automatic message is created when the employee for instance accepts an invitation to a meeting. Since each message has a member ID and the list of member IDs that correspond to manual messages is known, we were able to distinguish the manual from the automatic messages rather easily.

This is however not the case with the manual and semi-automatic messages. A semi-
automatic message is based on an existing message, a kind of template message or a SNOT as it is called in the *Company*. Every SNOT contains a number of *fixed* tags that cannot be changed or removed and a number of *variable* tags that have to be replaced by the author of the message. This contrasts with the manual messages where the employee is *free* to use whatever tag he likes. Unfortunately a unique ID that splits these kinds of messages from each other did not exist. The only information that could help us was a list of *SNOTs* with their fixed and variable tags.

To split the messages from each other we needed a way to identify the messages that are based on a SNOT. This implies that for every message we had to verify whether it contained all the fixed tags of one of the SNOTs. We could not take the variable tags into account because these were replaced by the author of the message. Besides this, we had to be careful because we had a huge amount of messages and we needed to write the code as efficiently as possible to prevent our computer from crashing.

To solve the problem, we developed software that verified whether every Dutch message in the database was a manual or semi-automatic message by using the fixed tags of the list of *SNOT* messages. Each time the tags of a message contained the fixed tags of one of the *SNOT* messages, we considered it as a semi-automatic message.

We have to remark that although we applied these rules to the messages, there is no certainty that we did not remove a message from the list that was not a semi-automatic message. For example, a manually created message that contained the same tags as those used in a particular *SNOT* was considered by our script as a semi-automatic message. Besides this, there was no certainty that we were able to withdraw all the semi-automatic messages as the *Company* did not have a unique identifier to check this.

Although we were probably not able to withdraw all the semi-automatic messages, we still believe we were able to remove a significant part of them. Therefore the case study can still be considered a valuable one.

---

5We did not apply this to French messages as we focus our analysis only on Dutch messages.
5.6 Descriptive statistics

In June 2007, the Company gave us a copy of the 2006 dataset which contained all the metadata as described in the previous section. The copy consisted of large sets of data exported from a database to text files. The text files were written on four DVDs and contained 2.28 gigabytes of data.

As the dataset contained metadata of 8,011,111 messages created in different languages, we decided to focus on a specific part of the data to answer our research questions. More specifically, we focused on Dutch messages created and written by the author of the message. In order to be sure that the tags were created by the author of the message, we excluded all the messages that were not manually created nor selected by the dactylo of the message. Next, we did not use tags created by non-Dutch speaking employees because handling tags from different languages can be considered to be a research issue on its own. Therefore, we selected the tags from Dutch manual messages where the employee-ID of the author equalled the employee-ID of the dactylo (2,872,015 messages). We decided to refer to these messages as messages of the Company throughout the dissertation.

In Table 5.1, we provide an overview of all of messages created in the Company.

<table>
<thead>
<tr>
<th>Total number of messages</th>
<th>8,011,111</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch messages</td>
<td>7,341,092</td>
</tr>
<tr>
<td>Manual messages</td>
<td>4,653,191</td>
</tr>
<tr>
<td>Automatic messages</td>
<td>1,694,007</td>
</tr>
<tr>
<td>SNOT manual messages</td>
<td>339,789</td>
</tr>
<tr>
<td>SNOT Automatic messages</td>
<td>654,105</td>
</tr>
<tr>
<td>messages</td>
<td>2,872,015</td>
</tr>
</tbody>
</table>

5.6.1 Messages

In 2006, there were 2,872,015 messages created by 9,419 different employees in the Company. In Figure 5.1, we visualize the number of messages sent each working day in 2006.
We must remark that we did not visualize the dates where no messages were created (for example weekends or official holidays). Based on the dataset we found out that messages in 2006 were only created during 252 days, an average of 11,397 messages a day. Except for one day, the 12th of October, there were only around 2000 messages sent. This might indicate that there was a strike that day because this is not considered as an official holiday in Belgium.

![Number of messages created in 2006](image)

Figure 5.1: Number of messages created in 2006

Different types of messages were sent. We provide an overview of the five most popular types of messages in Table 5.2

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outgoing mail</td>
<td>1,571,136</td>
</tr>
<tr>
<td>SNIV</td>
<td>821,645</td>
</tr>
<tr>
<td>Note</td>
<td>241,585</td>
</tr>
<tr>
<td>Fax</td>
<td>204,828</td>
</tr>
<tr>
<td>Letter</td>
<td>26,080</td>
</tr>
<tr>
<td>Other</td>
<td>6,741</td>
</tr>
</tbody>
</table>

### 5.6.2 Tags

The messages contained 17,719,536 tags in total and 839,722 of them were unique. The total amount of tags was pretty high compared to the average number of words explained
in a dictionary. For example, the Oxford English Dictionary contains around 250,000 words\(^6\).

We selected a set of high-frequency tags attributed to the messages and noticed immediately a number of problems\(^7\) that might explain the difference as we summarize below:

- **Company-specific terminology**: The Company has terminology which is only used and known inside the corporate environment. For example, names of branches and projects of the company we could not retrieve in a Dutch dictionary.

- **Digits**: There were many tags that only contained digits or digits as well as letters as shown in Table 5.3. This is something we do not easily find in an ordinary dictionary. We must remark that those tags have an average frequency that is much lower than the ones that only consist of letters. However, the number of unique tags that only contain letters is still high compared to the number of words in a dictionary.

  ![Table 5.3: Different kinds of tags](image)

- Many tags were not well cleaned up as we will explain in Chapter 6.

### 5.7 Conclusion

In order to answer the other research questions, we presented and described the dataset of the Company that we will use to answer these questions. We explained how the size and

\(^6\)These words do not contain inflections nor technical vocabulary ([http://www.askoxford.com/asktheexperts/faq/aboutenglish/numberwords](http://www.askoxford.com/asktheexperts/faq/aboutenglish/numberwords), retrieved 11 April 2010.).

\(^7\)An overview of all tag dataset problems: Chapter 6, Section 6.1.2.
structure of the dataset caused some difficulties to perform queries on the dataset.
Chapter 6

CorTagCleaning approach to improve the explicit quality of corporate tags

6.1 Introduction

Before we can investigate our research questions and measure the *implicit* quality of the corporate tags, we need to clean the tags within our dataset. More specifically, we need to be sure that the *explicit* quality of the tags, or the quality related to the tags themselves, is verified and checked.\(^1\)

Tags can be cleaned up automatically by using techniques borrowed from the domain of information retrieval as we will explain in Section 6.2. However, each of these techniques has its disadvantages and needs to be extended in order to solve the difficulties related to company-specific terminology.

Another approach is to ask users to follow a number of tag rules (Guy and Tonkin [2006]). Although the *Company* asked its employees to follow a number of tag rules\(^2\), we

---

\(^1\)As we will discuss later on in this chapter, we cannot guarantee to obtain an explicit tag quality that is 100 per cent accurate, because this would incur a very high total cost of ownership. We opted for an approach that significantly improved the explicit tag quality at a low cost and limited human feedback.

\(^2\)Overview of tag rules: Chapter 5, Section 5.2.
noticed problems with the *tags in the dataset* as we will detail in Section 6.1.1.

To overcome these problems, we suggest the CorTagCleaning approach, our *corporate tag cleaning approach*, which cleaned the tags of the *Company* with a limited amount of human resources and a low total cost of ownership.

### 6.1.1 Tag rules in the company

From the nine tag rules\(^3\) formulated by the *Company*, we selected three of them to measure how well they were followed by the employees. We were, however, not able to test all the tag rules:

- Tag rules four and nine could not be measured without additional information and input from the *Company*\(^4\) nor human interference.
- The results of testing tag rule five will be discussed in Chapter 7.
- We were able to retrieve tags that corresponded to tag rules five, seven and eight. However, we could not calculate how many of them were not written in the right way. Regarding tag rule seven: we only found 2,222 tags where the month and year were written as a composed tag (name of month and year).

We developed software to test the first three tag rules\(^5\). The software is mainly based on the approach we present in Section 6.3. A description of how we tested these rules, as well as the restrictions is explained in the Appendix A. In Table 6.1, we present the results of verifying these three tag rules.

We can conclude that the dataset contains conjugated verbs and adjectives, and plural nouns as well as digits that need to be tidied up before we could use them for further

---

\(^3\)Overview of the tag rules: Chapter 5, Section 5.2.

\(^4\)The information required to test these rules was not yet available in the *Company*. Asking the *Company* to collect this information would incur a high cost and take a lot of time.

\(^5\)Overview tag rules: Chapter 5, Section 5.2.
CHAPTER 6. CORTAGCLEANING APPROACH TO IMPROVE THE EXPLICIT QUALITY OF CORPORATE TAGS

Table 6.1: Results of testing three tag rules

<table>
<thead>
<tr>
<th>Tag rule (TR)</th>
<th>unique tags</th>
<th>unique tags that follow the tag rule</th>
<th>( f_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR 1: no conjugated Verbs</td>
<td>6,097</td>
<td>3,095</td>
<td>50.76%</td>
</tr>
<tr>
<td>TR 2: no plural nouns</td>
<td>22,826</td>
<td>6,127</td>
<td>26.84%</td>
</tr>
<tr>
<td>TR 3: no conjugated adjectives</td>
<td>4,632</td>
<td>1,696</td>
<td>36.62%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TR</th>
<th>total tags</th>
<th>total tags that follow the tag rule</th>
<th>( f_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR 1: no conjugated Verbs</td>
<td>513,142</td>
<td>161,166</td>
<td>31.40%</td>
</tr>
<tr>
<td>TR 2: no plural nouns</td>
<td>5,417,407</td>
<td>459,505</td>
<td>8.48%</td>
</tr>
<tr>
<td>TR 3: no conjugated adjectives</td>
<td>737,510</td>
<td>216,689</td>
<td>29.38%</td>
</tr>
</tbody>
</table>

analysis. The low number of tags that follow these tag rules might be explained by the absence of a control mechanism.

6.1.2 Tag dataset problems

In addition, we also did some random queries on the dataset to get an idea of what the data look like, for example we selected a set of high-frequency tags added to the messages. We noticed other problems with tags within the tag dataset. These tags polluted the dataset and they also had to be tidied up before we could use them for further analysis. Below, we list several problems we encountered as well as the ones we obtained when testing the three tag rules:

- There was pollution in the Dutch tag dataset caused by French and English tags. In the Company, the language of a message always corresponds to the language code or author of the message. Although the author is asked to change the default language when typing the message in another language, employees often forget to change this. As a consequence, some messages were polluted with French and English tags. Another cause of this problem was Anglicisms, words that are borrowed from English by Dutch, and Gallicisms, words that are borrowed from French by Dutch.

- Some tags could be classified as stop words, words that do not have a significant meaning such as articles and pronouns.
• **Spelling errors.** In the dataset of the Company, we distinguished two types of spelling errors.

  – *Misspellings* occurred due to typing errors or due to people not knowing how to spell a certain word or tag.

  – *C and k problem.* In the Dutch language, there exists some confusion about the spelling of certain words, especially the ones that begin with a *c* or *k*. The letter *c* and *k* are sometimes pronounced in the same way. In the last decade, the Dutch spelling rules have changed several times, and as a consequence people no longer know whether words should be written with a *c* and *k*. For instance, the Dutch word *contract*\(^6\) is sometimes written as *kontrakt*.

• **Dialect.**

• **Conjugated verbs.**

• **Conjugated adjectives.** In Dutch, adjectives are conjugated. For instance, the Dutch adjective *groen* or *green* in English, depending on its function in a sentence, can be *groen* or *groene*.

• **Singular** and **plural terms.**

• **Company-specific terminology.** We noticed company-specific terminology, such as Message-ID, Employee-ID, file names, names of branches of the Company, first names, and family names.

• **Numbers.** Year or date of a month

• **Synonyms**\(^7\) and **homonyms**\(^8\) We noticed general as well as company-specific synonyms and homonyms.

---

\(^6\) In English: *contract*.
\(^7\) Definition *synonym*: Chapter 2, Section 2.7
\(^8\) Definition *homonym*: Chapter 2, Section 2.7.
6.2 Related work

In literature on tagging, we distinguish two different kinds of tag cleaning methods often proposed and derived from the domain of information retrieval: stemming algorithms and the Levenshtein edit distance (Specia and Motta [2007], Braun et al. [2007], Van Damme et al. [2008c], Peters and Weller [2008], Angeletou et al. [2009]). In the paragraphs below, we discuss these techniques as well as lemmatization, which can be considered as a kind of stemming.

6.2.1 Stemming and lemmatization

Both stemming and lemmatization try to “reduce inflectional forms and sometimes derivationally related forms of a word to a common base form” (Manning et al. [2008]), but use different techniques to do so.

**Stemming** reduces tags to their stems or roots, the core meanings of words. The stemming algorithm removes suffixes. For example, it reduces the words separation and separated to separ (Paice [1994]). Because every language has its own suffixes, a stemming algorithm is language dependent. The Porter stemming algorithm is one of the most commonly used algorithms for stemming English words (Manning et al. [2008]).

Of course, it is important that the removal of suffixes does not have an impact on the semantic meaning of words (Paice [1994]). There are two commonly known errors related to stemming: understemming and overstemming errors. The first error occurs when suffix removal lets two words refer to different stems, although both are related to the same concept, for example division and divide will be respectively stemmed to divis and divid.

In the case where words which are not semantically related are stemmed to the same root, an overstemming error takes place, for example operation and operate will be stemmed to oper (Paice [1994]).

In contrast to stemming, **lemmatization** does not cut off the endings of words. In-
stead, it tries to transform words to their lemmas or their base or dictionary forms. For example, *separates* becomes *separate* and *computers* becomes *computer* (Manning et al. [2008]).

However, some words can be nouns or adjectives, depending on how they are used in a sentence, and can have more than one lemma. Most lemmatization tools contain morphological analysis techniques to derive the grammatical function of a word, for example Tadpole\(^9\). This is a feature which is not used in stemming algorithms (Manning et al. [2008]).

**Problems**

Both techniques are based on rules that are language dependent. This can sometimes cause problems; for example, when a set of tags contains words from another language or company-specific terminology. In such a case, words mostly differ from the general spelling rules. Applying the Dutch stemming algorithm to a corporate-specific tag of the *Company* can destroy its meaning. Before using one of these techniques, we believe a dictionary list should be used to filter out all the words that are not used in a specific language.

### 6.2.2 Levenshtein edit distance

The **Levenshtein edit distance** is a text similarity metric which calculates the distance between two words. More specifically, it counts how many letters have to be replaced, deleted, or inserted to transform one word into the other. The higher the Levenshtein edit distance, the more different two words are. When two words are similar we obtain a Levenshtein edit distance of 0 (Manning et al. [2008]). It is a valuable technique to verify the similarities of two tags as suggested in Specia and Motta [2007]. In order to calculate the distance, we should first make all the combinations of possible tag pairs.

---

\(^9\)An explanation about Tadpole will be given later on in this chapter.
CHAPTER 6. CORTAGCLEANING APPROACH TO IMPROVE THE EXPLICIT QUALITY OF CORPORATE TAGS

In Specia and Motta [2007], the Levenshtein distance is calculated as a relative frequency to make comparisons between tag pairs possible. The values lying between 0 (not similar) and 1 (very similar). Specia and Motta [2007] suggest a threshold value of 0.83 to indicate that two tags are similar. However, tests showed us that a threshold value of 0.83 excludes a number of similar tags. For instance, the Dutch nouns fiets and fietsen or bicycle and bicycles in English, express the same thing but do not agree in number. Both tags are similar and their Levenshtein similarity is 0.71, and thus lower than 0.83.

Problems

We believe it is somewhat risky to automatically consider a pair of tags as similar when a certain limiting value is exceeded. Sometimes tag pairs are not similar at all. A solution could be to increase this limiting value and to reduce the chance of misjudging a tag pair, but then we will have to exclude too many tag pairs from being cleaned up. Therefore, we believe human feedback should be included. A representative employee or someone else who is able to judge the similarity of a tag pair should be asked to confirm or reject the similarity because there is no certainty that two similar words are actually similar. To reduce the time a representative employee spends on verifying the similarity of tag pairs, the results can be saved and automatically be applied on the dataset the next time the tags need to be cleaned up.

We must remark that the Levenshtein edit distance does not solve all kinds of lexical problems, for instance when both tags in the tag pair are related to the same concept, but neither of them are in the appropriate format, for example raining and rains. In such a case, we could consider both tags as similar, but we cannot transform one tag into the other as they are both conjugated verbs. We still need a technique that cleans the pair with similar tags.
6.3 CorTagCleaning approach

To compensate for the problems of the techniques we just discussed, we decided to use an integrated approach which uses lemmatization and the Levenshtein edit distance to tidy up tags. In this way, we tried to take the weaknesses and problems of these techniques into account, as we will explain below.

We preferred lemmatization over stemming because the meaning of a word was sometimes completely lost by the stemming algorithm. In our opinion, the Levenshtein edit distance combined with human feedback is a good technique to clean up the company-specific terminology or the words which could not be retrieved in a Dutch dictionary. Although it might seem time consuming to ask someone to do this, we could save all the tags we evaluated as similar in a list and reuse the list the next time.

In addition, we needed to add some preparatory as well as additionally steps before we were able to start the lemmatization and calculate the Levenshtein edit distance.

In literature on enriching tags with semantics, online lexical resources and existing ontologies are suggested as an approach to improve the enrich tags. For instance, in Specia and Motta [2007], Van Damme et al. [2008a], and Peters and Weller [2008] the use of online resources such as Google, Wikipedia, online dictionaries is suggested as additional mean. The resources are regarded as spelling checkers and as a mean for retrieving concepts. However, we decided not to include these resources into our approach. The company-specific terminology makes it hard to use some of the sources on the internet. For instance, the Company had a gara tag, used as the abbreviation of the Dutch word garage. When using gara as a search term for Google, we did not find any link referring to the correct meaning of the term. On Wikipedia, we found a page describing the term, but the concept or description attributed to it was incorrect. On Wikipedia, gara is a Basque word and the name of a Spanish newspaper. This causes problems. We have to know whether the tag belongs to the specific terminology of the company or not. In order to find this out, human feedback is necessary. However, asking employees to verify the
background of the word can quickly become too time-consuming. Therefore, we decided not to include theses resources in our approach.\textsuperscript{10}

\subsection*{6.3.1 Step 1: preparatory steps}

Before we could apply our cleaning approach, we needed a list of distinct tags. To let the cleaning process run more smoothly, we removed some distinct tags that could never be tidied up. We must remark that we did not exclude these tags from further analysis.

\textsuperscript{10}This section is partially based on a section of: C. Van Damme, T. Coenen, T and E. Vandijck. Turning a corporate folksonomy into a lightweight corporate ontology. In Abramowicz, Fensel (Ed.): \textit{Proceedings of the 11th International Conference on Business Information Systems (BIS 2008)}, Innsbruck, Austria, LNBIP 7, pp.36-47. Springer, 2008.
Removal of tags with a least one digit

We had to remove some tags because some of them can never be cleaned up, for instance tags which contain a digit, such as an employee-ID. Without asking the author of the tag for verification, we can never be sure whether this particular tag is well written or not. Therefore, we decided to exclude these tags from the list of distinct tags that needed to be cleaned up. We could easily find these tags by using a *regexp* or *regular expression* in the MySQL query.

Removal of first names

As we noticed that there were a lot of first names used in the tag set, we decided to remove them from the list of tags. In some cases, first names could be considered as misspellings when we applied our CorTagCleaning approach.

We took a list of all the first names of the Belgian population, published by Statbel\textsuperscript{11} in 2007. Stabel is an official Belgian website which contains several statistics regarding the Belgian population. The list we downloaded from Statbel contains all the first names of Belgian people which are used at least two times to name a newborn. In our analysis, we excluded all the first names with a frequency of lower than 100 because some of them are used as nouns (for example *bloem*\textsuperscript{12}, *kat*\textsuperscript{13}) or adverbs (for example *geel*\textsuperscript{14}) and might lead to confusion. Besides, these names were not so popular and we therefore decided to exclude them from the list. We also replaced all the special characters in the remaining names, because only alphanumerical characters in the tag dataset of the *Company* are allowed. Next, we also removed two names which had frequencies of more than 100 and could be considered as homonyms: *augustus*\textsuperscript{15} and *mei*\textsuperscript{16} used as first names as well as months. The remaining list contained 5,067 first names.

\textsuperscript{11}www.statbel.fgov.be
\textsuperscript{12}In English: *flower*.
\textsuperscript{13}In English: *cat*.
\textsuperscript{14}In English: *yellow*.
\textsuperscript{15}In English: *August*.
\textsuperscript{16}In English: *May*. 
6.3.2 Step 2: removal of stop words

We removed the stop words, words that have little or no meaning. We selected words from the list of Corpus Gesproken Nederlands that is available through TADPOLE\footnote{Tadpole is an open source lemmatization tool for Dutch. This will be discussed later in this chapter.} and selected all articles, conjunctions, infixes, adverbs, and pronouns (1,112 words). We removed 14 ambiguous words and words which we could not retrieve in the official dictionary because the Corpus Gesproken Nederlands also contains spoken language that sometimes differs from written language. We opted for OpenTaal,\footnote{We used the vocabulary list retrieved in November 2008.} an open source Dutch vocabulary list. In July 2008, OpenTaal received the quality label Keurmerk Spelling van de Nederlandse Taalunie. The quality label guarantees that the vocabulary list of OpenTaal corresponds to the official Dutch language. Since the vocabulary list of OpenTaal is freely available, easy to import into a database table, and has a quality label, we decided to use this list. 327 Of the original words in the list remained. We also included the company-specific stop words FW, RW, and READ that are automatically created and added to a message when someone replies or reads a message. The list of stop words contained 330 words in total.

6.3.3 Step 3: adjustment of the c/k words

There are many words in Dutch where c and k words are pronounced in the same way. Spelling rules are often changed which causes a lot of confusion and people have difficulty knowing whether a word should be written with c or k. Because it is not easy to find an official list of c/k words and derivatives, we decided to take a list which we found on the site of Het groene boekje and also had a look at all the high frequency tags in the tag dataset of the Company to find some c/k misspellings. We provided a list of 78 c and k words in Appendix A. Of course, we are aware of the fact that this list was definitely not complete, but a more complete list was not available during our research.
6.3.4 Step 4: Levenshtein edit distance

Before we transformed tags to their lemma, as we will explain in step 3, we removed the tags which could not be found in an official dictionary. We opted for OpenTaal,\(^{19}\) an open source Dutch vocabulary list which also contains all the lexical variations of the Dutch words. There were 4,972,134 tags that could not be retrieved in the dictionary list of OpenTaal and had to be tidied up with the Levenshtein edit distance.

We made all combinations of possible tag pairs and calculated the Levenshtein edit distance for each tag pair. We calculated this distance\(^{20}\) in the same way as presented in Specia and Motta [2007]. Each time a tag pair exceeded the threshold value 0.65\(^{21}\), it was considered as a pair of possible similar tags. It was then listed in a pop-up window as shown in Figure 6.2. Then, we verified the tag pairs ourselves by going through the checklist. When the verification was done, we had to type the correct tag. We did not include functionality which allows the user to check which of the terms in the tag pair is the correct one. In some cases this did not correspond to one of the tags in the tag pair, and therefore we decided to do this manually. Neither lemmatization nor stemming can be used as these tags could not be retrieved from the dictionary list or could be considered to be company-specific terminology or French and English tags.

6.3.5 Step 5: transform tags to their lemma

The tags we retrieved from the OpenTaal dictionary list were used for lemmatization. Doing this manually is of course a time consuming task. TADPOLE is a tool which allows this to be done automatically. It is an open source tool developed at the University of Tilburg that can be used for several linguistic purposes (in Dutch) such as lemmatization. TADPOLE also tries to derive the grammatical function of Dutch words in a sentence.

\(^{19}\)We used the vocabulary list retrieved in November 2008.
\(^{20}\)We use the Levenshtein edit distance as implemented in http://search.cpan.org/ jgoldberg/Text-Levenshtein-0.05/Levenshtein.pm
\(^{21}\)Based on trial and error, we decided this value generated the most appropriate results.
On one hand, it uses existing corpora such as the *Corpus Gesproken Nederlands*, which contains more than nine million Dutch words retrieved from conversations of Flemish and Dutch people. On the other hand, it tries to automatically derive the grammatical function of an unknown Dutch word based on morphological analysis. Since a word can have different grammatical functions depending on its position in the sentence (e.g. *kost* can be a noun or verb), *morphological analysis* tries to derive the parts of speech of a word by analysing the position and relation of the word in a sentence (den Bosch et al. [2007]).

In our situation we did not need this kind of functionality since tags are not used in a sentence and are ordered alphabetically in our table. Therefore, we always ordered the tags in separate rows so that the data could be read as a separate word.

Of course, we had to be prudent with the *Corpus Gesproken Nederlands* because it includes interjections or other types of words that are often used in spoken language and can be confused with misspellings or company-specific terminology. We also noticed that corporate-specific terminology was often lost by using *TADPOLE*. This is why we only selected the tags that corresponded to the Dutch Dictionary *OpenTaal*.
6.3.6 Step 6: add tags we removed in the first step

We added the tags which we previously removed from the list except for the stop words, as those words were not meaningful, for example first names and tags which contain digits.

6.3.7 Step 7: remove duplicate tags

The tag cleaning process sometimes resulted in duplicate tags. Therefore, we had to remove the duplicate tags for each message by using the `Alter Ignore` MySQL syntax.

6.3.8 Overview of results

There were 25.67 per cent of the tags in the dataset which could not be cleaned up because they equalled a first name (2.38 per cent) or they contained a digit in the tag (23.29 per cent). Next, 6.80 per cent of the tags did not have a significant meaning and were considered to be a stop word. As a consequence, we removed 32.47 per cent of the tags from the dataset because these tags could not be cleaned up.

We had a closer look at the remaining 11,966,013 tags and adjusted 68,976 tags based on the c/k error list in Appendix A. As visualized in Figure 6.3, there were 67.14 per cent of the tags remaining that needed to be tidied-up.

Only 13.90 per cent of these tags (total: 6,924,903, unique: 35,109) corresponded to one of the words or associated formats of the dictionary list of OpenTaal. There were 11.084 tags out of the 35,109 which were cleaned up by using lemmatization. We adjusted 850,799 tags in total through lemmatization or 12.29%.

There were 86.10% tags or 4,972,134 in total that could not be retrieved in the dictionary list of OpenTaal. A total of 91.33 per cent of them had a frequency of less than 20 and could probably be considered as misspellings.

---

22 A summary is provided in Table 6.2.
CHAPTER 6. CORTAGCLEANING APPROACH TO IMPROVE THE EXPLICIT QUALITY OF CORPORATE TAGS

Figure 6.3: Results from CorTagCleaning Approach step 1-3.

We noticed that the high-frequency tags in this category mostly corresponded to company-specific terminology, such as the name of branches of the Company, or English and French tags. The large number of tags that could not be retrieved in the dictionary list of OpenTaal indicates that the Levensthein edit distance might be an interesting feature. However, we must remark that we were not able to calculate this step on the remaining set of tags. To calculate the edit distance, the computer had to make a huge amount of combinations \( \binom{4,972,134}{2} \). Therefore, we only applied the latter step on the tags that we used to answer a specific research question. In this way, we reduced the amount of tags and combinations to make. However, we applied this step on tags with a frequency of at least 100 and cleaned 4.63% of the 4,972,134 tags.
Table 6.2: Results of cleaning up tags

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>Unique</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Removal of tags at least 1 digit</td>
<td>4,126,028</td>
<td>582,928</td>
<td>23.29%</td>
</tr>
<tr>
<td>Removal of stop words</td>
<td>1,205,383</td>
<td>229</td>
<td>6.8%</td>
</tr>
<tr>
<td>Removal of first names</td>
<td>422,112</td>
<td>3,941</td>
<td>2.38%</td>
</tr>
<tr>
<td>Adjustment of c/k words</td>
<td>68,976</td>
<td>41</td>
<td>0.4%</td>
</tr>
<tr>
<td>Tags considering Levenshtein edit distance</td>
<td>4,972,134</td>
<td>217,474</td>
<td>28.06%</td>
</tr>
<tr>
<td>Tags considering Lemmatization</td>
<td>6,924,903</td>
<td>35,109</td>
<td>39.08%</td>
</tr>
</tbody>
</table>

6.4 Limitations CorTagCleaning approach

We tried to clean up the tags as much as possible with limited and human resources to reduce the total cost of ownership. However, we must remark that our approach did not tackle all the problems we encountered in the dataset.

Below we formulate a number of limitations of the CorTagCleaning approach:

- **We did not solve semantically related problems.** For instance, we did not solve the synonyms nor homonyms within the tag dataset. This is something we could do by using a thesaurus\(^{23}\) such as EuroWordNet for Dutch. As EuroWordNet is not freely available\(^{24}\), unlike the English WordNet, we tried to find another suitable thesaurus online, but we could not find any. Besides, there were also some synonyms and homonyms on the level of corporate-specific terminology that could probably not be retrieved from a thesaurus. It was also not possible to create a complete list of all the synonyms and homonyms used in the Company, as there was no complete synonym\(^{25}\) and homonym list available in the company and it would be too time consuming to build one. Therefore, we believe that using a general thesaurus such as EuroWordNet would probably not improve the accuracy a lot.

- Tags that were not on the OpenTaal list and did not have a similar tag with a Levenshtein edit distance of more than 0.65 were not cleaned.

---

\(^{23}\)Definition *thesaurus*: Chapter 3.  
\(^{24}\)We asked the organization EuroWordNet for a free version, but one was not available. Even for academic purposes we still had to pay a considerable amount of money.  
\(^{25}\)We define a complete synonym list as a list that contains the synonyms as well as the senses.
• French and English tags were not solved except when similar words were listed in the Levenshtein edit distance pop-up window.

• The cleaning of some tags depends on their function and this is something we can not derive from a tag. This is something which is normally retrieved through morphological analysis.

6.5 Conclusion

In this chapter, we presented the CorTagCleaning approach to verify and improve the explicit quality of tags, because we noticed that the tags within the Company were polluted. We tested three of the tagging rules asked by the Company and concluded that these three tagging rules were not well followed. In addition, we also noticed that there were many other tag problems. Therefore, we suggested to clean-up the tags with the CorTagCleaning approach which we presented in this chapter.

To increase the accuracy of the tags, we decided not to tidy up the tags using only one technique, but by using a mixture of several techniques. We will often refer to this approach throughout this dissertation. Although this approach is based on the tag dataset of the Company, we believe it is also applicable for other corporate datasets, as we will discuss in Chapter 10.
Chapter 7

Implicit quality of corporate tags

7.1 Introduction

Having investigated the explicit quality of tags of the Company. We measured the implicit quality of corporate tags in the Company. In addition, we evaluated a methodology to select high-quality tags automatically in a narrow folksonomy.\(^1\)

7.2 Quality of corporate tags

To investigate the research question whether the analysis of employees’ tags can help a company to unlock the explicit and tacit knowledge that resides in the organizational memory, we had to be sure of whether or not we can rely on the quality of the corporate tags, more specifically the ones from the Company. In contrast to the tags created on the World Wide Web, we believed we had to question the quality of tags even more when the tags are created in a corporate environment. We assumed that tagging in a corporate web environment and on the World Wide Web have slightly different incentive mechanisms.

\(^1\)We must remark that the methodology which we present in this chapter is focused on narrow folksonomies that contain textual resources.
While users on the Web are mostly tagging for personal usage and benefit (Golder and Huberman [2006]; Marlow et al. [2006]), for example future information recall, this might not really be the case in a corporate environment. Although employees might receive some kind of personal benefit when annotating resources in the company, for example\(^2\) retrieving documents more easily, the shareholders of the company get most of the benefits. The better that tags describe a resource, the more effectively the resource can be retrieved, and the better the employee will perform his or her job. Eventually, this might result in a higher efficiency and in a higher corporate profit. As a consequence, it is the company or the shareholders that gain the greatest benefits. We assumed that this may lead to a kind of principal-agent problem, as often referred to in economic literature: the interests of the principal (i.e. owners) conflict with those of the agents (i.e. employees) (Ross [1973]) and might result in low quality tags.

However, research from Thom-Santelli and Muller [2007] showed that people are concerned about their tags as we explained in Chapter 3, because their tags could of interest to other employees. Of course, this does not imply that they will always use high-quality tags to annotate their resources.

To know whether or not employees are more reluctant to choose low quality tags, we needed to measure the quality of the tags, in this case the implicit quality of the tags of the Company which belong to a narrow folksonomy. Next to this, we researched whether high-quality tags can also be automatically selected based on a mining of the textual resource. More specifically, we researched whether a principal-agent problem exists or whether employees are really concerned about their tags as concluded by Thom-Santelli and Muller [2007]. In addition, we also investigated whether suggested tags based on a mining of a text can help an employee to tag a message and to improve the quality of the tags.

In this chapter, we want to answer the research questions as formulated below

\(^2\)In the case they are using it to retrieve information.
1. **How good is the implicit quality of corporate tags?**

2. **Can a mining of a textual resource help to find high-quality tags?**

3. **Can the tags obtained through a mining of a textual resource help an employee to select more high-quality tags?**

4. **How do employees experience the difficulty of selecting high-quality tags?**

In Chapter 4, we investigated how in a broad folksonomy we can automatically derive *high-quality tags* by testing it on a del.icio.us dataset. These algorithms cannot be applied to the situation of a narrow folksonomy\(^3\), since resources are tagged by only one person as we explained in Section 4.3.2. Therefore, in this chapter, we propose a technique to find high-quality tags automatically in a narrow folksonomy, and compare the results we obtained when measuring the quality of tags in the *Company* manually.

This chapter is organized as follows. In Section 7.3, we provide an overview of related literature. We then discuss our research methodology in Section 7.4. We detail the results from the interviews in Section 7.5 and highlight the limitations in Section 7.6. We end this chapter with a conclusion.

### 7.3 Related work

At the moment of research, we were not aware of any paper discussing the quality of tags in a corporate setting. Only a small number of publications discuss the quality of tags used on the World Wide Web as we presented in the previous chapter. Since one of our research questions investigates whether mining of a textual resource can help to find high-quality tags automatically, we focus our related work section on literature that discusses how keywords can be derived automatically from a text and the difference between tags and automatically generated keywords.

\(^3\)Definition of *narrow folksonomy*: Chapter 3, Section 3.3.2 or Glossary.
7.3.1 Automatically generated keywords

In the domain of automatic indexing a lot has been written on how to automatically select the most appropriate keywords from a text. The most well known theory is probably the one which was extensively examined and discussed by Zipf and called Zipf’s law. Zipf showed that the words people use follow a power law distribution function\(^4\), but this theory was earlier discussed by Estoup as explained in Zipf [1949]. Due to the work of Zipf, the law became well known (Zipf [1949]). Luhn [1958] used Zipf’s theory in his work to propose a method to summarize a document. Luhn [1958] stated that the frequency of words in a text as well as the position of the words in a sentence can be used as a technique to detect significant sentences in a document. The approach does not take semantical relations into account; it only tries to find lexical variations of words and removes common words such as articles and prepositions because he considers them as non-significant. Luhn [1958] considers only the words that are in between the upper and lower bounds to be significant.

The term TF-IDF\(^5\) (Salton et al. [1974]) is another theory derived from that of Luhn which is also very common in the domain of automatic indexing and often cited in literature. As we discussed in Chapter 4, the importance of a keyword depends not only on the frequency of the word in the document but also on the intradocument frequency: how many times the term is used in related documents. Based on these theories many other algorithms have been developed.

7.3.2 Tagging compared to automatically generated keywords

Al-Khalifa and Davis [2007] did an experiment where they compared the quality of user-generated tags to the quality of those which are automatically generated by the Yahoo

\(^{4}\)Definition of power law distribution function: Chapter 3, Section 3.7.

\(^{5}\)Definition TF-IDF: Chapter 4, Section 4.6.3.
API\textsuperscript{6} term extractor. First, they explicitly asked two trained human indexers to compare both sets in terms of which had the most semantic value. Second, they automatically calculated the overlap between user tags and the keywords chosen by the trained human indexers. Results showed that the (1) tags generated by human indexers were better than automatic keyword extraction, and (2) the agreement between tags and keywords from human indexers was larger than that found when comparing the automatic versus the manual keywords.

7.4 Methodology

We first discuss the approach we suggested selecting high-quality tags automatically in a narrow folksonomy. Then, we detail the approach we used to measure the implicit quality of the tags manually as well as the interviews we set up.

7.4.1 Approach to select high-quality tags automatically in a narrow folksonomy

Although research from Al-Khalifa and Davis [2007] showed that automatic keyword extraction does not give better results than human indexers, we wanted to research if this is also the case in a corporate environment. More specifically, we were interested to know the extent to which automatic tag generation can help the employees in a company to select high-quality tags for textual resources.

We decided to opt for a simple technique to find high-quality tags, because it is important that these kinds of keywords can be found immediately. Calculating a corpus of similar resources as is the case of TF-IDF\textsuperscript{8} would probably require very powerful

\textsuperscript{6}Application Programming Interface (API) is an interface which allows an interaction with another software program. (Wikipedia)

\textsuperscript{7}This is an API of Yahoo which extracts terms form from documents by taking the context into account (http://developer.yahoo.com/search/content/V1/termExtraction.html)

\textsuperscript{8}Definition of TF-IDF: Chapter 4, Section 4.6.3.
computers or a lot of time to do the calculations on the spot. Therefore, we opted for the law of Luhn, but did not use an upper bound. We assumed that excluding upper bound tags would remove an important property or characteristic of the resource and wanted to keep the tags that were above a lower bound.

For each textual resource, we suggested to break the sentence down to words, select all the distinct tags, and clean up the lexical variations of the words without losing the company specific terminology. We decided to use the CorTagCleaning approach as described in Chapter 6. Based on a trial and error testing, we decided to take 0.50 as a threshold value for the Levenshtein edit distance to consider a pair of words as similar. We noticed that we could obtain a better accuracy when taking a lower limiting value than the one we suggested in Chapter 6 because the set of words in a message was rather small.

After we cleaned up all the words, we calculated the frequency and took all the words with a frequency of at least two. Since we assumed that words with a frequency of one were probably not valuable to be a high-quality tag, we excluded them. Also, we did not want to have a list of suggestions that was too long. A limiting value that was too high, would probably exclude a number of important tags, therefore we decided to take a frequency of two.

7.4.2 Methodology to measure the quality of the corporate tags manually

To know the extent to which high-quality tags can automatically be derived from a textual resource based on our methodology, we first had to measure the quality of the corporate tags manually. In the paragraphs below, we detail the research method we used to measure the quality of the corporate tags, and the sample of messages and tags of the Company.

9Of course, the CorTagCleaning approach contains the Levenshtein edit distance that requires human feedback as we discussed in Chapter 6. To overcome this problem, a company could build-up a list of similar words by testing this on a set of tags annotated to corporate messages.

10We decided to take a lower threshold value than in the CorTagCleaning approach because the number of words in a message is less than applying the approach on set of tags.
All the preparatory steps we had to take, we discuss in the Appendix A.

**Interviews**

As our research method we opted for survey research. This is a method of *collecting, organizing, and analysing data* (De Vaus [2002]). Different techniques of data collection can be used: questionnaires, interviews, in-depth interviews, observations, and so on (De Vaus [2002]). We preferred interviews and decided not to do a questionnaire because interviews are far more flexible. In case a respondent does not understand a question the interviewer can give some additional information (Bernard [1966]). Bernard [1966] distinguishes a difference between three kinds of interviews: standardized, semi-standardized, and unstandardized interviews. We decided to use a combination of standardized and semi-standardized interviews. In a standardized interview the interviewer is not allowed to change the order of the questions nor to ask additional questions. Each respondent has to answer the same questions. Since we wanted to give some additional explanation in case the question was not clear to the respondent, the interview was also semi-standardized. As we will explain in the next paragraph, the importance of the order of the interview questions was another reason why we opted for interviews: we had to be certain that no one cheats. The confidentiality of the message made it impossible to do the interviews electronically.

During the interview each interviewee had to read 10 messages and annotate the message with his own tags: *high-quality tags* which corresponded to the tag quality dimensions *accurate* and *complete*\(^{11}\). Since research discussed in a previous chapter shows that the high frequency algorithm is a good algorithm to select *high-quality tags* in broad folksonomies, tags annotated to a resource, we decided to let every message be evalu-

\(^{11}\)Definition of *tag quality dimensions*: Chapter 4, Section 4.4 or Glossary.
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

ated by several persons.\textsuperscript{12} We opted for six employees.\textsuperscript{13} In this way, we sort of create a broad folksonomy where we can apply the high frequency tag quality algorithm\textsuperscript{14} to find high-quality tags.\textsuperscript{15}

During the second part of the interview, the interviewee had to judge the quality of the original tags in terms of being high-quality tags by selecting a value on a five-point scale

\textit{very relevant}, \textit{relevant}, \textit{mediate relevant}, \textit{not so relevant}, \textit{not relevant}

Finally, he had to select from a suggestions list additional tags that could also be considered as high-quality tags. At the end of the interview, we asked the participants to fill in a small questionnaire\textsuperscript{16} with open and closed or fixed-alternative questions.

7.5 Results

Before we could start the analysis process, we had to do some cleansing on the tags for the interviewees, the authors’ ones, as well as the suggestions to make the process of comparing tags and words smoother. We provide an overview in the Appendix A.

In this section, we first introduce a number of definitions and notations and then detail the results of the interviews.

\textsuperscript{12}In Chapter 4, we concluded that the high-frequency as well as the TF-IRF tag quality algorithms were the most appropriate algorithms from the three proposed algorithms to select high-quality tags in broad folksonomies

\textsuperscript{13}We must remark that the messages were evaluated by six employees on average. In a few cases they were evaluated by only four people or more than six employees, because some participants asked for another message because they were new in the \textit{Company}.

\textsuperscript{14}In Chapter 4, we concluded that the high-frequency and the TF-IRF tag quality algorithms were the most appropriate tag quality algorithms to select high-quality tags in a broad folksonomy.

\textsuperscript{15}However, Thomas Vander Wal did not specify the exact number of people that should annotate a resource to consider it to be a broad folksonomy. In our opinion, we believe that when six people annotate a resource it could be considered as a broad folksonomy.

\textsuperscript{16}An overview of the questions is provided in the Appendix A.
CHAPTER 7. IMPLICIT QUALITY OF CORPORATE TAGS

7.5.1 Definitions and notation

In this section we first define three types of tags and then introduce a number of definitions regarding the quality of a tag.

Type of a tag

- **author tag** is a tag chosen by the author of the message.

- **spontaneous reader tag** is a tag chosen by the reader of the message without any input from the interviewer.

- **suggested tag (type A)**: is a word from the text that has a frequency of at least two.

- **suggested tag (type B)**: a suggested tag (type A) that does not belong to the category of an author tag.\(^{18}\) This is the list of suggested tags which was shown to the interviewee.

- **suggested reader tag** is a tag chosen by the reader from the list of suggested tags (type B).

Quality of a tag

In real life, most decisions are taken on approval of a majority: more than 50 per cent of the people have to agree on a certain proposal in order to get it accepted. As there were only six people to evaluate the quality of the tags of the messages, we decided to take a threshold value that was larger and opted for 60 per cent.

- **author tag of good quality**: at least 60 per cent of the readers of a message qualify an author tag as *relevant* or *very relevant*.

\(^{17}\)We define *reader* as the employee being interviewed or the interviewee.

\(^{18}\)We excluded the suggested tags (type A) that were similar to the author tags, because this was the list of suggested tags we showed to the interviewees in the last part of the interview. We did not want them to see duplicate tags. Also, we did not want to have a list of suggested tags (type B) that was too long.
• **author tag of poor quality**: at least 60 per cent of the readers of a message qualify an author tag as *not so relevant* or *not relevant*.

• **spontaneous reader tag of good quality**: at least 60 per cent of the readers of the message choose this tag after they read the message. We must remark that in some situations a spontaneous reader tag was similar to an author tag.

• **suggested reader tag of good quality**: at least 60 per cent of the readers of the message choose a tag that belongs to the suggested tags (type B) and that was not selected, yet as an author tag of good quality nor a spontaneous reader tag of good quality.

• **suggested tag of good quality**: a tag from the list of suggested tags (type A) is considered to be a *suggested tag of good quality* when it is an author tag of good quality, spontaneous reader tag of good quality, or a suggested reader tag of good quality.

• **high-quality tag of a message**: a tag that belongs to the category of the author tags of good quality, spontaneous reader tag of good quality, or a suggested reader tag of good quality\(^\text{19}\).

• **missing tag**: a high-quality tag of message that is not an author tag.

**Quality of author tags annotated to a message**

• **message with author tag of good quality**: at least 60 per cent of the author tags are author tags of good quality.

• **message with author tags of poor quality**: at least 60 per cent of the author tags are author tags of poor quality.

\(^{19}\)This definition is still compliant with the tag quality dimensions we detailed in Chapter 4.
CHAPTER 7. IMPLICIT QUALITY OF CORPORATE TAGS

Other definitions

- **length** of a message: Each message from the sample has a different length. The length varies from 76 words to 1760 words.

7.5.2 Question 1: how good is the implicit quality of corporate tags?

To answer this research question, we first discuss some descriptive statistics we obtained from the interviews on the sample of messages. In the second part, we discuss the results of a number of hypotheses.

Descriptive statistics obtained from sample

We first calculated the quality of the authors tags and the high-quality tags based on the definition we described in previous section. We summarize these results from the sample of 47 messages in Table 7.1 together with their standard deviation. Sixteen messages (34 per cent) from the sample could be considered to be messages with good-quality tags and only three messages (6.38 per cent) could be classified as messages with poor-quality tags. There were 28 messages that were in-between these classes and thus could not be classified as messages with poor-quality tags, nor as messages with good-quality tags. This means there was only a small amount of messages in the sample that had author tags of a poor quality.

It is not because a message can be considered as a message with a good quality of tags that it contains a sufficiently number of high-quality tags. It could be that the author forgot to add some important tags. Therefore, we calculated the number of high-quality tags of a message as a combination of all author tags of good quality, spontaneous reader tags of good quality, and suggested reader tags of good quality. As there were no suggested reader tags of good quality, we simplified the formula as a combination of all author tags
On average, a message from the sample had 6.32 high-quality tags. When we compared the high-quality tags of a message to the author tags of a message, we obtained 2.62 missing tags per message in the sample.

Based on the results of the sample, we conclude that the quality of the author tags in the sample was good. There were on average only 1.81 author tags of poor quality and 2.62 missing tags in the sample of messages.

To answer our research question, we verified the significance of the results obtained from the sample by testing some hypotheses.

**Hypothesis 1: a message has on average more author tags of good quality (ATGQ) than author tags of poor quality (ATPQ)**

The results in Table 7.1 show that there is indeed a difference between the author tags of good quality ($\bar{x}_{ATGQ} = 3.57$) and author tags of poor quality ($\bar{x}_{ATPQ} = 1.81$). However, to

---

20 We must remark that there was often an overlap between author tags of good quality and spontaneous reader tags of good quality.
conclude that in the population of messages, a message has on average more author tags of good quality than author tags of poor quality, we had to do a paired t-test (one-tailed).

\[
\begin{align*}
H_0 & : \mu_{AT_{GQ}} \leq \mu_{AT_{PQ}} \\
H_1 & : \mu_{AT_{GQ}} > \mu_{AT_{PQ}}
\end{align*}
\]

![Figure 7.1: SPSS results two-sample paired t-test](image)

Results from SPSS in figure 7.1, show that we have to reject \( H_0 \) \((p= 0.000 ; p < \alpha, \alpha = 0.05)\). We can conclude that a message has, on average, significantly more author tags of good quality than author tags of poor quality.

---

\( \mu_{AT_{GQ}} \) = average author tags of good quality in the population of messages, \( \mu_{AT_{PQ}} \) = average author tags of poor quality in the population of messages.

22 The results are for a two-tailed test. To obtain the results for a one tailed, we need to divide the p-value by 2.
Hypothesis 2: a message has on average more high-quality tags than author tags of a good quality

Although a message has, on average, significantly more author tags of good quality than author tags of poor quality, we wanted to research whether the average number of author tags of good quality was sufficient. Therefore, we tested whether the average number of high-quality tags of a message is higher than the author tags of a good quality:

\[
\begin{align*}
H_0 & : \mu_{HQT} \leq \mu_{ATQQ} \\
H_1 & : \mu_{HQT} > \mu_{ATQQ}
\end{align*}
\]

![Figure 7.2: SPSS results two-sample paired t-test](image)

Results in Figure 7.2 show that we have to reject \( H_0 \) \((p=0.000, p < \alpha, \alpha = 0.05)\) and we can conclude that a message has, on average, significantly more high-quality tags than
author tags of a good quality. Based on the results of testing hypotheses 1 and 2, we can conclude that the implicit tag quality of messages is good, but can still be improved.

Hypothesis 3: a message has on average 6 author tags of good quality

We wanted to test whether tag rule five of the Company was obeyed. This rule requires a message to have at least six meaningful tags. Therefore, we tested in SPSS whether a message from the population has on average 6 author tags of good quality:

$$\begin{align*}
H_0 &: \mu_{AT\text{GQ}} = 6 \\
H_1 &: \mu_{AT\text{GQ}} \neq 6
\end{align*}$$

![Table](image)

**One-Sample Test**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author_tag_good_quality</td>
<td>47</td>
<td>3.57</td>
<td>1.956</td>
<td>.270</td>
</tr>
</tbody>
</table>

![Table](image)

**One-Sample Test**

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author_tag_good_quality</td>
<td>-8.987</td>
<td>46</td>
<td>.000</td>
<td>-2.426</td>
</tr>
</tbody>
</table>

![Table](image)

**One-Sample Test**

<table>
<thead>
<tr>
<th></th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author_tag_good_quality</td>
<td>-2.37</td>
<td>-1.86</td>
</tr>
</tbody>
</table>

Figure 7.3: SPSS results one-sample t-test

Results from a paired sample test (two tailed) in SPSS (Figure 7.3) show that we obtain a p-value = 0.000 and we reject $H_0$. We conclude that the tag rule five is not followed: the average author tags of good quality for a message are significantly different from 6 (p=0.000, p< $\alpha$, $\alpha$ = 0.05). The output of SPSS also gives us information regarding the confidence interval. We can conclude with a certainty of 95% that the average author tags of a message in the population are between, 3.03 and 4.12 and thus lower than 6.
Hypothesis 4: the relative number of high-quality tags of a message (RHQT) is influenced by the length of a message.

To test this hypothesis, we introduced a new variable: *relative number of high-quality tags* (RHQT). This variable is calculated by dividing the number of high-quality tags per message by the length of a message. In this way, we can measure the influence of the length on the number of high-quality tags of a message better, because a few messages were very long.

We did a regression analysis where we studied the linear relation between the dependent variable, RHQT, and the independent variable, length. We can estimate the regression equation of the population as follows:

\[
\hat{\mu}_{RHQT} = \beta_0 + \beta_1 \text{length}
\]

Again, we used SPSS to test the estimated regression equation. Results from the regression analysis can be found in Appendix D.

Regression analysis requires two assumptions to be tested: normality and homoskedasticity. To test the first assumption, we plotted the observed relative cumulative frequency of the *relative number of high-quality tags* versus the expected cumulative frequency in case of normality. On the plot, we notice that many observations are on the diagonal or nearby the diagonal and therefore they can be considered to come from a population with a normal distribution.

We tested the homoskedasticity by plotting the standardized predicted values versus the standardised residuals. Since the standardized residuals do not have a systematic increase or decrease, we can conclude that the second assumption is also met and we were allowed to do a regression analysis.

Then, we had a look at the Pearson correlation and noticed that the correlation between the dependent and independent variable is negative (\(r = -0.523\)) and significant.
(p=0.000, \( p < \alpha, \alpha = 0.05 \)). Based on the results from the model summary, we can conclude that 27.4 per cent of the variance of HQT average is explained by the model \( (R^2=0.274) \).

Results in the ANOVA model show that the vector of the regression coefficients \( \beta \) is significant different from 0 \( (p=0.000, p < \alpha, \alpha = 0.05) \). We can conclude that the independent variable explains the variance of the dependent variable. In the coefficient table, we find more information regarding each regression coefficient. Both regression coefficients are significant different from 0 \( (p=0.000, p < \alpha, \alpha = 0.05) \). We can rewrite our model as follows:

\[
\hat{\mu}_{HQT\text{average}} = 0.34 - 2.535E^{-5}\text{length}
\]

This implies that when the length of a message increases, the expected number of relative number of high-quality tags will slightly decrease.

Taking the results from Table 7.1 as well as the results of the hypotheses into account, we can conclude that the implicit quality of the tags is good. Messages have, on average, significantly more author tags of good quality than author tags of a poor quality \( (p=0.000, p<\alpha, \alpha = 0.05) \). However, the quality of the tags can still be improved: a message has, on average, significantly more high-quality tags than author tags of a good quality \( (p=0.000, p<\alpha, \alpha = 0.05) \). Also, the average author tags of good quality significantly differs from 6 \( (p=0.000, p<\alpha, \alpha = 0.05) \) and are consequently not compliant with tag rule five of the Company. We can conclude with a certainty of 95% that the average author tags of good quality in the population of messages is between 3.03 and 4.12. In addition, we can conclude that the length of a message significantly influences the relative number of high-quality tags: long messages will slightly decrease the relative number of high-quality tags.
7.5.3 Question 2: can a mining of a textual resource help to find high-quality tags?

To answer this question, we measured the quality of the suggested tags (type A) by counting the number of times the suggested tag (type A) corresponded to a high-quality tag of a message.

The descriptive statistics of the sample show that on average, 3.83 tags of the high-quality tags of a message could be retrieved in the suggested tags (type A). We must remark that the list of suggested tags (type A) also contained many tags that could not be considered as high-quality tags of a message: on average 24.13 of the suggested tags (type A) per message in the sample. Results\textsuperscript{23} from doing a paired t-test showed that, on average, the number of suggested tags that cannot be considered as high-quality tags was significant larger than the number of suggested tags of good quality (p=0.000, p< $\alpha$, $\alpha = 0.05$).

Suppose we would use the suggested approach to select high-quality tags automatically for this sample of messages, the list of suggested tags would contain a lot of tags that cannot be considered as high-quality tags. As the approach generated too many irrelevant tags for this sample of messages, we conclude that our suggested approach is not appropriate to select high-quality tags automatically.

7.5.4 Question 3: can the suggested tags (type A) help an employee to select more high-quality tags?

However, we believe that the approach is valuable to generate suggested tags of good-quality because the list of suggested tags (type A) contained on average 3.83 tags of the high-quality tags per message in the sample. Results\textsuperscript{24} from doing a t-test show that

\textsuperscript{23}SPSS output can be found in Appendix D.
\textsuperscript{24}Output from SPSS can be found in the Appendix D.
the number of suggested tag of good quality is, on average, significantly larger than 3 (p=0.0195, p< α, α = 0.05).

Therefore, we were interested to research whether the interviewees liked the suggested tags (type B) or not as we discuss in the next section. Of course, we have to be aware of the fact that the list of suggested tags (type B) contained less tags than the suggested tags (type A) as we explained in Section 7.5.1.

In the last part of the interview, we asked the interviewees to answer a questionnaire. Regarding suggested tags, we first asked the interviewees whether they liked the idea of suggested tags in general: 72 per cent of the interviewees were in favour of suggested tags in general. Then, we also asked them whether they liked the suggested tags (type B) and asked them to explain their answer. We classified their answers into five groups as shown in Table 7.2. Based on the answers provided by the interviewees we can conclude, that 21 of the interviewees agreed with the suggested tags (type B) and 8 of the interviewees who did not explicitly liked the suggested tags (type B).

<table>
<thead>
<tr>
<th>It depends on the message</th>
<th>Number of interviewees</th>
</tr>
</thead>
<tbody>
<tr>
<td>I already selected the suggested tags myself</td>
<td>8</td>
</tr>
<tr>
<td>Some of the suggested tags were interesting for me because I did not think about them myself</td>
<td>8</td>
</tr>
<tr>
<td>The suggested tags did not help me because the list of suggested tags was too long, contained too many detailed words, etc.</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 7.2: Answers from interviewees to the question whether they liked the suggested tags (type B)

Based on the results of measuring the quality of the suggested tags (type A) in the sample as well as the answers from the interviewees we can conclude that the suggested tags (type A) can be used to help the employees when selecting high-quality tags.
7.5.5 Question 4: how do employees experience the difficulty of selecting high-quality tags?

In the questionnaire, we asked the interviewees how they experienced the process of selecting tags in general. Five of the respondents told us that they had difficulty in selecting appropriate tags when they had to label a message. “I always try to imagine who will be the audience for my message? What kind of tags would they use to retrieve it? This is not so easy for me”. It is very important to them that a colleague will be able to retrieve the message based on their tags. We could say that they sort of select tags for the community, in this case the employees, and not for themselves.

Someone else argued: “Some words can be written in different ways and I never know which one I have to choose”. Two employees said that it depended on the content of the message and therefore they answered that it is sometimes difficult for them to tag, but not all the time.

In all, 76 per cent of the participants did not consider tagging a message as a difficult job. “We cannot write a message before writing down our tags. So, the first thing I always do is think about the content I am planning to write down.” Before sending the message, the author of the message can always revise the chosen tags before sending the message. Some of readers adjust the tags, when they finished the message. Or as someone else explained: “Once I have finished writing down my message, I can easily choose my tags because I know the content of the message very well”.

Based on the answers of the interviewees, we can conclude that employees do not consider selecting tags as a difficult job. However, the quality of the corporate tags can still be improved. Since the number of high-quality tags are, on average, significantly larger than the author tags of good quality (p=0.000,p<\alpha, \alpha = 0.05). Based on their answers, it seemed like the interviewees have mainly selected words from the message. To have more certainty on this issue, we tested whether a relationship existed between the evaluation of an author tag and being a word in the message.
CHAPTER 7. IMPLICIT QUALITY OF CORPORATE TAGS

Hypothesis 5: there is a relationship between the evaluation of the quality of an author tag and being a word in the message.

$H_0$: there is no relationship between the evaluation of the quality of an author tag and being a word in the message.

$H_1$: there is a relationship between the evaluation of the quality of an author tag and being a word in the message.

For each author tag, we counted the number of interviewees who evaluated the tag as an author tag of high/low quality or mediate relevant and we determined whether the tag was used as a word in the message. We collected all the values and put them in a Table 7.3.

<table>
<thead>
<tr>
<th>Author tag quality</th>
<th>Yes (word in message)</th>
<th>No (not in message)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author tag of a good quality</td>
<td>880</td>
<td>330</td>
</tr>
<tr>
<td>Author tag evaluated as mediate relevant</td>
<td>195</td>
<td>178</td>
</tr>
<tr>
<td>Author tag of a poor quality</td>
<td>260</td>
<td>533</td>
</tr>
</tbody>
</table>

Table 7.3: Data table hypothesis 5

To test the significance of association in this table, we did a chi-squared test. We were allowed to calculate chi-squared because all the expected frequencies had a value larger than five. We obtained a $p=0.000$ and could therefore reject $H_0$. We can conclude that there is a significant relationship between the evaluation of an author tag and being a word in the message.

7.6 Limitations

We have to be aware of a number of limitations of this research

1. As we were not able to involve the recipients of the messages in the interviews, we noticed that some interviewees had difficulties evaluating the author tags of the message because they did not know the context of the message. This was especially
true for employees that had not been in the company for a long time. Some of the interviewees rated some author tags as not relevant, although they might have evaluated them as relevant if they had known the context.

2. For a few employees it was hard to apply the definition of high-quality tags in terms of the tag quality dimensions accurate and complete as, for some of them, high-quality tags are tags they use to retrieve the message. In this way, there was sometimes a confusion.

3. There was no limitation on the maximum number of spontaneous reader tags.

7.7 Conclusion

We manually measured the quality of tags in a sample of messages obtained from the Company by letting 29 employees evaluate and annotate 47 messages. We concluded that the implicit quality of the tags was good: messages have, on average, significantly more author tags of good quality than author tags of a poor quality ($p=0.000$, $p<\alpha$, $\alpha=0.05$). Therefore, we cannot conclude that there is a principal-agent problem as we assumed in the introduction of this chapter. However, the quality can still be improved: messages have significantly more high quality tags than author tags of good quality ($p=0.000$, $p<\alpha$, $\alpha=0.05$).

Since the suggested approach to select high-quality tags generated on average significantly more non high-quality tags than high-quality tags ($p=0.000$, $p<\alpha$, $\alpha=0.05$), and similar results were obtained in Al-Khalifa and Davis [2007], we concluded that it is not an appropriate approach to replace the tags through text mining. However, more complex text mining techniques might generate better results, but it was out of the scope of our research to further investigate this.

We concluded that the suggested approach to select high-quality tags would be a valuable approach to improve the quality of the tags. The list of suggested tags contained,
on average, significantly more than 3 high-quality tags per message and a majority of the interviewees liked the quality of the suggested tags.

Only a minority of the interviewees (5 people) considered selecting tags as a difficult job. Most of them did not experience selecting tags as a difficult job. Many of them try to take the audience of the message into account and select the tags from the message they have written. Indeed, we tested and concluded that there was a significant relationship between the evaluation of an author tag and being a word in the message (p=0.000, p< α, α = 0.05).
Chapter 8

Unlocking tacit knowledge through tags: expert finding

8.1 Introduction

We investigated the quality of the corporate tags of the Company in the previous chapters. Since the Company is mainly using tags for information retrieval purposes, we researched the quality of the tags from an information retrieval perspective as we will discuss in Chapter 10. Since the quality of tags is not only important for information retrieval purposes, but also for knowledge retrieval purposes we researched how employees’ tags can help a company to unlock the tacit knowledge that resides in the organizational memory, in this chapter. We focus our research on how we can use tags to get the tacit knowledge which resides in employees. As an approach, we propose expert finding or collecting the expertise of the employees through an analysis of tags. We support this approach by literature which argues that:

1 More information: Chapter 10, Section 10.2.2.
2 The research we discuss in this chapter was performed with the help of a Master Thesis student Sarah Berghmans whom we supervised during her thesis: S. Berghmans. Het lokaliseren van experts binnen de onderneming aan het hand van een analyse van tags. Master Thesis, Vrije Universiteit Brussel, Belgium. 2009
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

• Experts are considered as a source of information or knowledge which is not (yet) documented (Yimam-Seid et al. [2003]).

• Tags are said to be a reflection of people’s interest and knowledge (John and Seligmann [2006]).

To answer the main research question of this chapter, we introduce a number of related research questions:

• Do tags that are used to label corporate resources reflect the expertise of the employee?

• Which of the tag analysis techniques we present in this chapter is appropriate to reflect the expertise of the employees?

• Do employees who use the same tags have the same knowledge field?

To answer the main as well as all the related research questions, we present three tag analysis techniques which we applied to a tag data set of the Company to try to unlock the expertise of the employee. Next to these techniques, we visualized the social networks which implicitly exist within the Company through the use of tags. By analysing these social networks, we tried to find similar experts in the Company. As a last step, we asked a group of employees to evaluate the tag sets we obtained and to verify the social networks.

The structure of this chapter is as follows. After we motivate why it is important for a company to retrieve its experts in Section 8.2, we provide in Section 8.3 an overview of literature on expert recommender systems, discuss some related work on how tags are used to deduce expertise from employees, and elaborate on the social network which arises through tag use in Section 8.4. The rest of the chapter is organized as follows: we discuss the research methodology we used in Section 8.5, give an overview of the techniques in Section 8.5.1, and in Section 8.6 we describe the answers of the interviewees.

3 A description of the data set is provided in Chapter 5.
We provide a conclusion at the end of this chapter and point out the limitations of this research.

8.2 Motivation for expert finding

Many authors focus on the importance of unlocking the knowledge which is within people. For example, Yimam-Seid et al. [2003] note that a knowledge management approach should not only contain the means to make information stored in documents accessible, but also reveal the knowledge stored in people.

Hertzum and Pejtersen [2000] investigated how engineers are searching for information and experts, and more specifically, how both activities are related to each other. As previous research in this domain already revealed that engineers are more likely to first ask their colleagues for information, Hertzum and Pejtersen [2000] wanted to know the processes that exist among both sources. The results from their research showed that engineers seek documents in order to find the experts within the company. The answers of the experts will help them to retrieve the necessary documents. This observation corresponds to one of the reasons why employees try to find experts, as explained by Yimam-Seid et al. [2003].

According to Yimam-Seid et al. [2003], experts are mainly needed for two reasons: to retrieve information and to help execute tasks. When people want to retrieve specific information, they have to find out who knows what about a certain topic. In most cases, people cannot find the information by themselves or do not have time to perform a search. In cases where people are looking for an expert to execute a specific task, they are more interested in knowing how much the expert knows about a specific topic (Yimam-Seid et al. [2003]). The larger the organization and the more the organization is geographically dispersed, the more internal experts become important. Next to internal experts, external experts are also important as they can foster the collaboration between different actors, for example between customer and supplier.
8.3 Expert recommender systems

As information about people’s expertise is hard to gather manually, automated support becomes necessary. Collecting information in a database on who knows what might be a solution but again it remains a time-consuming task, especially to keep the information updated. Letting people maintain personal websites showing their expertise is another option, but it is hard to retrieve the website from the Web and a manual interpretation is still required. Therefore, Yimam-Seid et al. [2003] propose to use automatic expert finders to collect the knowledge of employees.

Automatic expert finders try to overcome the problem of collecting tacit information. Based on the overview given by Yimam-Seid et al. [2003] on the automatic expert finders, we can conclude that implicit information is retrieved by analysing a number of secondary sources such as documents created by the employee himself and intelligent agent systems. In literature, there are a lot of expert recommender systems, for example Demoir, Mitre, Kean and Hermes.

8.4 Related work

In this chapter, we discuss literature on how tags should be used to find experts and the social networks that arise around the creation of tags.

8.4.1 The use of tags and experts

Some papers discuss how tags can be used to find experts. We notice two kinds of research directions. On the one hand, researchers suggest explicitly asking users to annotate the expertise of a person (Farrell and Lau [2006], Coenen [2006], Farrell et al. [2007], Schmidt and Braun [2008]), and on the other hand researchers discuss how expertise from people can be derived based on the tags they use to annotate resources (John and Seligmann
CHAPTER 8. UNLOCKING TACIT KNOWLEDGE THROUGH TAGS: EXPERT FINDING

[2006]). In the paragraphs below we elaborate on these research directions.

Instead of tagging a bookmark, Coenen [2006] suggests building expert location systems by letting people annotate other members of the community. In this way, it will be easier to retrieve experts.

Farrell and Lau [2006] and Farrell et al. [2007] also suggest tagging people, more specifically employees. The authors present a system, called Fringe, where employees can describe their own expertise or that of a colleague with tags. To make the barriers to tagging as low as possible, they created a number of plug-ins that allow people to tag in several applications or as they describe it, tag people in context. For instance, there is a plug-in that gives the employees the option to tag the colleague with whom they are chatting without needing to leave the chat application. Alongside creating a number of plug-ins, they took a number of privacy actions to prevent employees from adding tags that exaggerate the expertise of the employee. To evaluate the approach, they did a user survey as well as interviews. Based on the user survey, they concluded that most of the respondents were satisfied with tagging and being tagged by their colleagues. The results from the interviews show that:

- most people use the tags to organize their contacts,
- tags given by their colleagues reflect the expertise of the employee well,
- tags which are not updated may become a problem, and
- tags can be used for ranking experts although a more extended technique is required than merely ordering the experts by tag frequency (Farrell and Lau [2006], Farrell et al. [2007]).

The tags in Coenen [2006], Farrell and Lau [2006], and Farrell et al. [2007] are, however, explicitly created to describe someone’s expertise and thus require additional input from the employees. We believe it is less time consuming to derive people’s expertise through implicit tag use and more profitable in the long run.
Schmidt and Braun [2008] observe that tagging systems, such as Fringe (Farrell et al. [2007]), 43people,4 and Xing,5 have shortcomings since users can employ whatever tags they like. They state that there is no common vocabulary nor common topics used to describe the expertise of the employees/people, which they consider as a weakness. Therefore, Schmidt and Braun [2008] present an approach to improve the expert finding systems based on tagging. Instead of using flat tags, ontology concepts should be used to describe people’s knowledge and at the same time users can participate in maintaining and extending the competence ontology.6 By letting users collaboratively participate in the creation process of competence concepts, they explain how they can come to an agreed vocabulary and a competence ontology in the long run. Having an agreed vocabulary or a competence ontology to describe people’s expertise generates more value than using flat tags. The approach that Schmidt and Braun [2008] present is based on the ontology maturing process, an approach which is described in Braun et al. [2007] to collaborative build ontologies based on tagging and formal ontologies. We agree with Braun et al. [2007], but we assume that it will be hard to implement this in a company as it will require a good incentive mechanism to persuade employees to participate in creating the ontology. Furthermore, we should also remark that this approach is not yet tested.

John and Seligmann [2006], on the other hand, discuss how the tags people use to annotate resources can be analysed to derive the expertise of this person. In the paper, John and Seligmann [2006] introduce the concept of Expertrank and propose two kinds of models to calculate the Expertrank. The first model does not take the relationships between the tags into account and is calculated by counting the number of times an employee has used a certain tag to label a bookmark. The second calculates the expertise of the employee by also taking into account the related tags in a cluster of strongly connected tags. This model assumes that when someone has knowledge on topic or tag x, he also
has knowledge on tags that are strongly related to topic or tag x. To take the relation into account, they present a formula based on the Page Ranking Algorithm. The Page Ranking Algorithm was introduced in Brin and Page [1998]. It is a technique which takes into account the citations or hyperlinks of one Web page to another Web page when calculating the results of a query entered in a search engine. The Page Ranking Algorithm is one of the algorithms which is used to calculate the search results in Google.

The research which we present in this chapter is closely related to the work of John and Seligmann [2006]. However, our research differs from John and Seligmann [2006] in three ways:

1. In contrast to the research of John and Seligmann [2006], we present three techniques to analyse the tags. Two of them differ from the ones proposed in John and Seligmann [2006].

2. Whereas the two techniques presented in John and Seligmann [2006] have not yet been tested, we applied our techniques to a real case dataset and validated the results by interviewing a group of employees.

3. Before using the tags for further analysis, the authors did not apply a tag cleaning approach to improve the explicit tag quality.

8.4.2 Experts and social network analysis based on tags

Tagging delivers more information than just tags: Mika [2005] states that a social dimension is connected with the tagging process. When we look at tagging systems on the World Wide Web, we can see that there are four groups or entities involved in the tagging process: (1) tags, (2) resources, (3) actors, and (4) the tagging systems themselves (Gruber [2005], Van Damme et al. [2008c]). As a consequence, social networks are formed around tags (Mika [2005], John and Seligmann [2006]).
A social network is “A specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved (Mitchell [1969])”.

A social network can be considered as a real-world graph as mentioned in Dorrow [2006]. Each graph consists of nodes and edges: the nodes represent people and the edges visualize the social ties between them. Different social networks can be derived from tags as is shown in Mika [2005]. Out of a tripartite model of tags, resources, and actors, Mika [2005] explains how a number of bipartite graphs can be generated based on the co-occurrence of its elements: the AC (actor-tag) graph, AI (actor-resource) graph, and the CI (tag-resource) graph. In his work, Mika [2005] focuses on how the folding of these graphs into one-mode networks generates implicit social networks, a network of instances and lightweight ontologies.

Social networks are said to be valuable in finding experts (Coenen [2006]). Ehrlich et al. [2007] present SmallBLue, an application that helps to find experts in the company. The system uses data mining techniques to retrieve experts by mining emails and chat conversations written by the employees. Next to text mining, it also uses social network techniques to visualize the position of the user versus the expert he is looking for. More specifically, the system provides information regarding the degrees of separation from the experts or it measures the social distance between the user and employee: how many people are between them. In this way, users are also able to seek sub-communities working on a certain topic in the company.

The social networks that arise around tags in a community may also be used as a technique to detect expertise in the community (John and Seligmann [2006], Coenen [2006]). Implicit social networks in a tagging system may return additional information regarding the expertise of a person, since the tags that people choose are said to express and reflect the actors’ subjective levels of knowledge on and their interest in the respective

---

7 In his work, Mika [2005] uses the term concept, but we prefer to use the term tag.
8 In his work, Mika [2005] uses the term instance, but we prefer to use the term resource.
9 Definition of lightweight Ontology: Chapter 2, section 2.6.
resource (John and Seligmann [2006], Van Damme et al. [2008c]). Consequently, actors are indirectly linked with other actors by sharing the same tags and/or resources (John and Seligmann [2006], Van Damme et al. [2008c]). John and Seligmann [2006] assume that people who use the same tags have the same field of knowledge.

One of the models used to calculate the Expertrank as proposed by John and Seligmann [2006], suggests calculating the expertise of the employee by also taking into account the related tags in a cluster of strongly connected tags. This means that when someone has knowledge on topic x, he also has knowledge on tags that are strongly related to this particular tag. To obtain this information, clusters are extracted from the social network.

8.5 Methodology

To answer the research questions we formulated in the introduction, we first tested three different tag analysis techniques and visualized the social networks that are formed around the tags of a sample from the Company. Through interviews we then asked a group of employees for feedback. We discuss each of the tag analysis techniques in Section 8.5.1, we explain which dataset we used in Section 8.5.2 and detail how we used interviews to answer our questions in Section 8.5.3.

8.5.1 Description of the techniques

In this section, we present the different techniques we applied on employees’ tags to find their expertise. For each of these techniques, we give a description, explain how we will calculate it and give a motivation. Since research in Farrell et al. [2007] shows that taking the high frequency tags that are explicitly created to describe people’s expertise requires a more complicated technique, we suggest also two other techniques to find the expertise.
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

High Frequency

1. **Description** We selected tags with the $n$ highest frequencies for each employee.

2. **Calculation** For each employee we determined their distinct tags, calculated the frequency of each unique tag, ordered them by frequency and selected the tags with the highest frequency. We did this by performing some MySQL queries and exported the result to a text file ordered by highest frequency. We then selected the top $n$ tags for each employee in the file.

3. **Motivation** Tags are said to be a reflection of the interests and/or knowledge of the actor (John and Seligmann [2006]). For instance, someone who uses quite often the name of a project as a tag, probably implies that this person has knowledge on or interest for this project. This is one of the techniques proposed in John and Seligmann [2006]. As we also obtained good results with this technique in chapter 4, to select high quality tags in broad folksonomies, we decided to test the same technique to find out whether we can seek the expertise from employees based on a tag high frequency analysis.

Cumulative Frequency

1. **Description** We selected the tags that were between an upper and lower bound percentile or an inter percentile. As upper bound percentile, we took the percentile, just after the one which contains the high frequency tags. As lower bound percentile, we took the percentile that contains the tags with a certain threshold value.

2. **Calculation** For each employee we determined his distinct tags, calculated the frequency of each unique tag and ordered them by frequency. To be able to select the tags that lay between a certain inter percentile, we needed a way to rank the tags on an ordinal scale. At first sight, tags belonged to a nominal scale, but we ranked them on an ordinal scale by their frequency. We put the frequency categories on

---

10 In John and Seligmann [2006], they refer to it as “expertrank in an unstructured tag collection”.

166
CHAPTER 8. UNLOCKING TACIT KNOWLEDGE THROUGH TAGS: EXPERT FINDING

the x-axis and the cumulative frequency on the y-axis. We divided the cumulative frequency on the y-axis in 100 different percentiles.

3. **Motivation** Tags which are used a lot but not as much as high frequency tags, might also be descriptive. Therefore, we opted for a technique that selects tags between an upper and lower bound. The same technique was also suggested in Luhn [1958] to automatically detect the keywords in a text. We must, however, remark that we did not take the interpercentile between both boundaries, but a smaller interpercentile that was located near the upper bound.

**TF-IPF**

1. **Description** We introduced a variation to TF-IDF and TF-IRF, which we presented in Chapter 4.

2. **Calculation** To calculate the TF-IPF, we first needed to define TF-IPF. The formula is based on TF-IDF\(^{11}\) (with \(t_{x,y} = \text{frequency of } tag_x \text{ for } person_y, T_y = \text{total number of tags for } person_y, \text{corpus} = \text{sum of persons and } R_x = \text{sum of persons that have used } tag_x\).)

\[
TF - IRF(tag_{x,y}) = \frac{t_{x,y}}{T_y} \ast \log(\frac{\text{corpus}}{P_x})
\]

To calculate TF-IPF, we needed to build a corpus\(^{12}\). In this case, we needed to build a corpus of people instead of resources as we proposed in chapter 4. When we got a set of clusters, we took a cluster that looks interesting to investigate. Then, we selected all the people that share at least one of these tags with a frequency of 20.

3. **Motivation** As previously discussed, the TF-IDF technique was often used in the domain of information retrieval. Just like the TF-IDF score, TF-IPF takes the inter as well as intra person frequency into account. This implies that a tag which is

\(^{11}\)Description of **TF-IDF**: Chapter 4, section 4.6.3.

\(^{12}\)We used the same methodology as we described in chapter 4.
often used by one person but also by many people in the corpus is not so good to describe the expertise of this person. When this situation occurs, the tag should be considered as a more general tag and not as a specific one to describe the expertise of the person. Also, results in Chapter 4 showed that the high-frequency as well as the TF-IRF tag quality algorithm, also derived from TF-IDF, were suitable to select high-quality tags in broad folksonomies. Therefore, we were interested to know whether this technique would be suitable to unlock the expertise of the employees.

**Choose an appropriate value for \( n \)**

We opted for \( n=10 \) because employees have varied expertise that can be related to the projects they have been working on in the company as well as more general expertise that is not company-related. This means that it would be very hard to describe someone's expertise in a few tags.\(^{13}\)

**8.5.2 Dataset**

We selected the messages created by all the employees working at the functional group 001 in the *Company*. The functional group 001 represents all the employees working at the IT department. There were 452 employees in functional group 001 that corresponded to these criteria. Together they created 394,838 messages and labeled them with 1,837,370 tags in total of which 114,323 were unique.

\(^{13}\)As we will point out in the limitations, we did not test whether 10 is an appropriate value to visualize someone's knowledge.
8.5.3 Interviews

As research method we opted for semi-standardized interview questionnaires. Before we started the actual interviews, we first had a conversation with our contact person at the Company. As our dataset of the Company contained tags created by the employees in 2006 to annotate messages, we wondered whether the time difference would be an issue. We discussed our concerns with our contact person and she told us that this should not be a problem. The policy of the company is to expect employees to be always approachable concerning the projects on which they have worked in the past. We discussed the rest of the interview questions and adjusted some of the questions according to her feedback.

To answer the research questions that we formulated at the beginning of the chapter, we needed to test the values we obtained, and we opted to conduct interviews. Each interview took around 30 minutes and consisted of two parts.

In the first part, we showed each interviewee three tag sets which we received when applying the techniques. We asked them which one reflected their expertise best, taking into account that it concerned the expertise they had in 2006. Next, we showed them a tag cloud and asked them whether a tag cloud visualization would help them to find an expert in the company.

In the second part of the interview, we talked about the interviewee’s position in the general social networks and verified whether his position is a good reflection of reality. Finally, we showed the interviewees their personal social networks and ego networks, and asked them whether colleagues in their ego networks have the same fields of knowledge.

An overview of all the preparatory steps is provided in Appendix B.

---

14 The list of questions which we used as a guideline to structure the interview can be found in Appendix B.
15 The interviews took place on 2nd and 3rd of April 2009.
16 18th of March 2009.
17 A complete overview of the preparatory steps for the interviews are described in Appendix B.
18 Definition of tag cloud: Chapter 3, Section 3.6.
19 An ego network is a network which visualizes all the nodes to which a central node is directly connected in a network and it also includes all the ties that exist between these other nodes (Everett and Borgatti [2005]).
8.6 Interview results analysis

We discuss the answers from the interviewees in the sections below.\textsuperscript{20} Since we were only allowed to do ten interviews, we did not do any statistical analysis on the answers received from the interviewees.

8.6.1 Question 1: which technique reflects the expertise of an employee best?

There was no agreement upon the techniques which reflect the knowledge of the employee the best. Seven of the interviewees opted for the high frequency technique, two for the cumulative frequency technique and only one person chose the third technique. However, all of the respondents preferred a combination of techniques instead of selecting one of the techniques because each of the lists contained tags that were too general\textsuperscript{21} and were not interesting enough to specify the expertise of someone. For instance, five out of the ten interviewees liked to merge the results from high frequency technique with the ones of the cumulative frequency technique because some lists contained tags that were too general. We visualized the answers to this question in Figure 8.1 and 8.2.

One of the interviewees remarked that it would be more interesting to first clean up the tags that are too general before applying the techniques.

Based on the answers from the interviewees we conclude that general tags should be withdrawn before analysing the tags chosen by the employees to label their resources. When general tags are not removed, a combination of the high-frequency and TF-IPF tag analysis techniques is the most appropriate technique to reflect the expertise of the employee.

\textsuperscript{20}We ordered the answers from the interviews to correspond to the research questions formulated in the introduction of this Chapter.

\textsuperscript{21}Some of the tags were probably automatic generated.
8.6.2 **Question 2: does tagging of corporate resources well reflect employees’ knowledge?**

Ten of the employees believed that an analysis of tags reflected the expertise of the employees very well. When asking the interviewees whether they wanted to add a tag or change the order of the tags, eight of the interviewees only wanted to add one (five interviewees) or two tags (three interviewees) to one of the tag sets they believed are the most appropriate. Five of the interviewees wanted to change the order of the tags.

Although the company assumes that an employee should always keep the knowledge he has on a certain topic he worked on in the past, one of the interviewees remarked that it was important to take a time factor into account. As projects change and evolve, they are not always aware of all new developments and therefore the time factor should be taken into account when deriving employees’ expertise by analysing tags. Currently, they have a system within the *Company* that describes the expertise of each employee based on tags provided by the employee. However, the interviewee explains that most of these data were not up to date and almost no one updates the tags on a regular basis.
As a related question, we asked the interviewees whether they would like to see the expertise of a colleague in a tag cloud instead of a vertical or horizontal list. Based on an example of how a tag cloud looks like, all of the interviewees believed that the representation of the expertise in a tag cloud would be an interesting feature.

Based on the responses of the interviewees, we conclude that an analysis of tags created by the employees reflects their expertise very well and that these tags should be visualized in a tag cloud. When doing an analysis, however, the time factor should be taken into account.

8.6.3 Question 3: do employees who use the same tags also have the same knowledge field?

Ten of the interviewees agreed upon the position of the employees in the social network visualized in Figure 8.3. The size of the nodes in the network is based on their degree factor: the number of incoming and outgoing relationships between nodes (Cross and Parker [2004]). The higher the degree, the larger the size of the node. In Figure 8.3,
we noticed that nodes situated at the centre have a higher degree than employees at the boundaries.

Figure 8.3: Social network formed around tags

One of the interviewees remarked that the employees situated at the centre of the network were mostly employees with a leading position or help desk position within the company whereas employees at the boundaries had a more specific knowledge. In Figure 8.4, we added two circles to the network to separate these groups.

Most of the respondents said that this network did not generate an added value for themselves, but it could give new employees an idea about the dynamics of the group.

The ego network could be an interesting tool to retrieve employees with a certain expertise, as five out of eight interviewees stated that the ego networks reflected with whom they are really in contact with. Three out of eight employees have colleagues in their ego network whom are having a similar knowledge. This is very interesting because it allows you to retrieve a colleague with a similar knowledge in case of absence. But, the

22Two of the interviewees did not have an ego network.
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

Figure 8.4: Social network circles

results showed that only three out of eight interviewees believed that the ego network can be used to find similar experts.

Based on the answers from the interviewees, we conclude that social networks formed around tags can be used to visualize the position of an employee in the organization. Although the social network did not give additional information to the interviewees, we believe such visualizations could be interesting means for new employees as well as for the board of directors and executives. For the board of directors and executives it is important to know which employees are central and would have an impact on the performance of the company when they are hired away. The ego networks reflect well those with whom the employees are in contact, but there were only three out of eight employees who stated that ego networks could help them find colleagues with similar expertise. Again, these ego networks could be an interesting to for the board of directors and executives as well as new employees.

\[23\] For two of the interviewees, we did not have any ego networks. The reason why two of the interviewees did not have any ego networks is explained in the Appendix B.
8.7 Limitations

There are a number of limitations related to the research we presented in this chapter:

- The techniques we presented in this chapter were tested on a dataset of 2006 while the interviews took place in 2009.
- Only 10 employees were interviewed and therefore we did not include any statistical analysis. Of course, it would be better to have more interviewees, but in a corporate setting it is not so easy to do large scale interviews. Interviewing employees implies a high cost for a company.
- A limited number of tag analysis techniques were proposed and tested.
- We did not test whether ten tags is an appropriate value to describe an employees’ knowledge.

8.8 Conclusion

Although we are aware of the limitations related to this research, we were able to formulate answers to the research questions described in the introduction.

Based upon the interviews we had with the employees in the Company, we may conclude that an analysis of tags that are used to label resources are an interesting means to find experts in the company. Also, the tags obtained through the analysis should be visualized in a tag cloud.

Answers from the interviewees indicated that the high-frequency technique is a good technique to analyse tags for employees’ expertise, but all of the interviewees preferred to select a combination of two techniques instead of one: five out of ten interviewees opted for a combination of the high-frequency and TF-IRF tag analysis techniques. The fact that the tags contained some general tags could explain why none of the employees preferred
to select only one technique. The interviews also revealed that the time factor should be taken into account when analysing employees’ tags in order to unlock their expertise.

The social networks formed around tags can be used to visualize the position of an employee in the organization. However, the social network did not give additional information to the employees, except when they had been recently hired by the company. The ego networks reflect well those with whom the employees are in contact, but there were only three out of eight employees who stated that ego networks could help them find colleagues with similar expertise. Although only three employees believed that this would give them new knowledge regarding experts, we believe this kind of knowledge could be very interesting for recently hired employees, as well as the board of directors and executives.

The tacit knowledge which is unlocked through tags could be a valuable input for the business strategy of a company, more specifically the human resources strategy of a company. For instance, the visualization of the social network that arises around tags provides the board of directors and executives in a company insight on who are central people in the organization. When central people are hired away, this could have a huge impact on the performance of a company. Therefore, a company should be aware of this kind of knowledge in order to make sure some kind of back-up strategy in case these employees would leave the company.
Chapter 9

Enabling business value by unlocking explicit knowledge through tags

9.1 Introduction

We discussed in Chapter 8 how tacit knowledge can be derived through an analysis of tags. In this chapter, we study how explicit knowledge can be unlocked from the organizational memory through an analysis of tags.

Different approaches currently exist to unlock explicit knowledge, such as business intelligence (BI) and business performance management (BPM). However, these approaches deal with information overload in a company. To implement BI or BPM adequate tools are required that analyse structured as well as unstructured information. The latter type of information is an important one: Forrester research showed that 80% of the corporate information is stored in unstructured information that remains untapped (Orlov and Ramos [2004]).

For a company, it is important to have access to the corporate knowledge to measure its development accurately. By measuring the performance the board of directors and executives know whether business value is created and if it is necessary to adjust the
corporate strategy. The business strategy of a company is formulated by the board of directors and executives and is translated into actions and objectives (Kaplan and Norton [1996b]).

We assume that an analysis of tags could be an interesting alternative to analyse unstructured information and to overcome the information overload, because we can consider tags as the key elements or summary of a document. Instead of analysing a document as a whole, the analysis could be limited to the tags themselves to unlock the explicit knowledge. In addition, tags can also be used to describe the content of scanned documents and images included in corporate documents.

Although many software tools already exist to analyse text, computer programs are not yet as intelligent as human beings. A similar situation exists with image recognition. Currently, optical character recognition (OCR) software are still not good at recognizing images and reading images of words. ReCAPTCHA and the ESP-game are two examples of web applications that use human intelligence to help computers with OCR (Zittrain [2008]).

Since dashboards and scorecards\(^1\) are used as performance measurement tools which provide an overview of the activities that took place in a company (Eckerson [2005]), we wonder whether dashboards can also be generated through an analysis of tags. More specifically, we want to know whether tag dashboards can also be used as a tool to give an overview of the activities that took place in a company and thus to unlock explicit knowledge which resides in the organizational memory.

In this chapter, we investigate:

- whether tag dashboards derived from the corporate strategy and business objectives, can be used as a tool to unlock the explicit knowledge which resides in the organizational memory.

\(^1\)In section 9.4, we explain that only a small difference between the terms dashboards and scorecards exists.
CHAPTER 9. ENABLING BUSINESS VALUE BY UNLOCKING EXPLICIT KNOWLEDGE THROUGH TAGS

- how tag dashboards should be implemented in the company in order to be used as a performance measurement tool to support the business strategy.

To answer these research questions, we present a tag dashboard approach, detail the four components of the tag dashboards, and present the results of 5 in-depth interviews with managers\(^2\) of the functional group of Analysts in the Company to verify our tag dashboard approach. In addition, we present a step-by-step plan to implement dashboards in the company.

The structure of this chapter is organized as follows. We first discuss the problem of information overload in a company and the impact it has on the organization. Then, we highlight the different approaches companies use to extract the corporate knowledge and to measure the knowledge development. In Section 9.4, we provide an overview of the literature of scorecards and dashboards. Next, we propose our tag dashboard approach and detail related work on unlocking knowledge through tags. In the last section, we discuss the results of our interviews at the Company.

9.2 Information overload and the impact on the organization

As already discussed in Chapter 2, there is no generally accepted definition on the term information overload.

A lot of publications, such as Schroder et al. [1967], Eppler and Mengis [2004] and Ceglar and Roddick [2006], stress that the performance of a person depends on the amount of information he has to deal with. The performance is expressed in terms of adequate decision making. Human beings are said to be able to process a certain amount of information, but when a critical amount of information is reached, additional information

\(^2\)In the Company, a manager is an employee who is responsible for a subgroup within the functional group.
processing becomes a burden because it has an impact on the decisions someone has to take. This process is often visualized with an inverted U-curve as conceived by Schroder et al. [1967].

The problem of information overload is not something which only exists in libraries or on the World Wide Web; it is also a problem that is present in companies across different management disciplines, such as marketing, accounting, and management information systems (Eppler and Mengis [2004]).

Figure 9.1: The inverted u curve (Eppler and Mengis [2004])

Eppler and Mengis [2004] did an extensive literature research on information overload across different management disciplines and concluded that the effects or symptoms of information overload in an organization can be classified into four groups: limited information retrieval possibilities, problems in organizing all the information, impact on performance of the individual, and impact on the decision-making process.

Hence, we can conclude that information overload has a huge impact on the organization. More specifically, a company has difficulty being aware of all the useful information and of course the knowledge which resides in the organizational memory and in taking appropriate decisions. It is important that a company has the means to measure all the actions that take place in the company to know whether it achieved its business objectives or not.

---

3Chapter 2, Section 2.5.
9.3 Evolution of approaches to information management and decision making

Over the last decades, different approaches have been used to manage information and to support decision making: decision support systems, executive information systems, data warehouses, business intelligence, and business performance management (Marr and Schiuma [2003]).

Back in the 1970s, companies used decision support systems (DSS) to help decision makers to make better decisions. Computer systems combine data with insights from the decision makers and try to visualize them in a user-friendly interface (Turban and Aronson [2001]).

Later on, these computer systems were replaced by Executive Information Systems (EIS) which put more focus on the needs of the top executives (Marr and Schiuma [2003]). Turban and Aronson [2001] explain that literature on DSS shows that those systems provided more information for middle-level managers than for top managers.

As information became more and more spread over the company, a system that provides a unified view became necessary and hence data warehouses found their entrance into the market.

The entrance of digitization increased the amount of information enormously. A more extended approach was required and business intelligence was born. Business intelligence provides the necessary tools to analyse large amounts of data to improve decision making in the company. Nowadays, several business intelligence tools are available on the market, but most of them only focus on analysis of the structured information and neglect the unstructured information (Spangler and Kreulen [2007]). Therefore valuable business information remains untapped. Some business intelligence tools do leverage unstructured
information. For example, the mining tool SAS Text⁴ and Clarabridge BI Search⁵ mine structured as well as unstructured information, but these systems are often very expensive.

At present, there is a new approach: business performance management. Also called corporate performance management or enterprise performance management (Frolick and Ariyachandra [2006]), it can be defined as a top-down discipline that helps executives understand what processes are needed to achieve strategic objectives and then measure the effectiveness of those processes to deliver the desired results (Eckerson [2005]). The term business performance management has a different meaning from business intelligence. Whereas business intelligence is a technique to mine large amounts of information to improve decision making, business performance management is a much broader concept. In addition, business performance management also contains the means to check, control, and manage the business objectives (Frolick and Ariyachandra [2006]).

9.4 Dashboards and scorecards

In Marr and Schiuma [2003], a citation analysis on the literature of business performance management reveals that the balanced scorecard of Kaplan and Norton dominates the literature on business performance management.

A company needs a “balanced presentation of financial as well as operational measures”, as Kaplan and David [1992] describe it, because there is no single measure which summarizes all the critical information of the performance of a company. After doing a research project in 12 companies, Kaplan and David [1992] came up with a balanced scorecard that contains four perspectives based on both measures. The operational measures are split into a number of categories: customer perspective, innovation and learning perspective, and internal processes. For each of these measures, goals must be set and specific measures should be identified.

⁴http://www.sas.com/technologies/analytics/datamining/textminer/index.html
⁵http://goliath.ecnext.com/coms2/summary0199-5418813ITM

182
Although the terms dashboard and scoreboard are often used as synonyms, a difference between them exists, as Eckerson [2005] explains. Scorecards help a company to control tactical and strategic goals whereas dashboards are used to visualize the operational processes. Since a clear distinction between the two terms does not really exist, Eckerson [2005] says that it is not wrong to use the terms interchangeably as long as they display performance information which can be used to supervise the processes and goals.

Eckerson [2005] classifies the dashboards into three groups: operational, tactical, and strategic dashboards. We discuss each in a little more detail.

- **Operational dashboards**: visualize the main operational processes in a company for front-line workers and supervisors.
- **Tactical dashboards**: give managers an overview and analysis of all the activities, processes and projects that take place in the company.

---

• Strategic dashboards: are created for executives in the company and provide performance information related to their strategic objectives.

9.5 Tag dashboard approach

Since dashboards are used to provide an overview of activities and processes that take place in the company and are populated with data from different information systems or databases, we wonder whether dashboards can also be generated through an analysis of tags. Tags are used to describe resources and can therefore be regarded as key elements to describe the resource.

Although software exists to analyse text, computers do not yet have the same reasoning power and intelligence that human beings have to summarize a text or to distinguish significant terms from non-significant ones (Sowa [1984]). Computers first have to screen a whole text and then apply the rules that are created and implemented by human beings.\textsuperscript{7} An empirical study presented in Al-Khalifa and Davis [2007] showed that human beings are better indexers than the text indexation algorithm they presented.

Research in the domain of cognitive science explains that the simplest things for human beings are the hardest for computers and vice versa. For instance, recognizing a face is very easy for a human being whereas counting the number of characters a document contains is very time consuming for humans (Sowa [1984]).

The ESP-game and ReCAPTCHA are two examples of image recognition applications that use human intelligence to make computers smarter. In the ESP-game people are asked to tag pictures. Since humans are not always willing to do this on a voluntarily basis, they made a game out of it. When two anonymous players use the same tag to label the picture, they score a point and the image is labelled with this particular tag. Recently, Google licensed this technology to improve image search: Google Image Labeler. A sim-

\textsuperscript{7}At the moment, computers do not yet have algorithms that give them reasoning power that human beings have. As better algorithms are developed for computers, this might change in the future.
ilar principle is implemented in ReCAPTCHA. On one hand ReCAPTCHA is used for validation to make sure that the web surfers are humans and not robots and on the other hand they are used to help to recognize words from scanned texts (Zittrain [2008]).

Results from an experiment done by J. Sachs showed that human beings are good at memorizing the general and semantic meaning of a topic but have difficulties in recalling the exact words of a text they have heard or read. This means that human beings have the intelligence to analyse unstructured information and to detect the important elements in it but fail to do this for all the unstructured information which resides in a company as computers can do (Sachs [1967]).

We believe that those two worlds, humans and computers, should be more intertwined so that we can leverage the intelligence of human beings and combine it with the calculation power of computers. The intelligence of human beings can be tapped by letting humans use their own-user-created keywords or tags to describe the unstructured information they have created or read, and the computation power of computers can be used to apply existing mining algorithms on the tags.

### 9.5.1 Related work

At the moment of research, we have not found any related work on how tags can be used to unlock the explicit knowledge in the organizational memory.

Elsewhere, we described a methodology to derive a lightweight ontology from corporate tags (Van Damme et al. [2008a], Van Damme et al. [2008b]). By applying our methodology to a dataset of the Company and discussing the results with some of the employees, we briefly validated that the lightweight ontology\(^8\) reflects the corporate terminology and concepts of the company. We suggested that the visual output of the methodology might have possible applications such as helping managers in their decision-making process since it allows them to detect irregularities in the obtained lightweight ontology.

\(^8\)Definition of lightweight ontology: Chapter 2, Section 2.6.
In Van Damme et al. [2008c], we briefly explained how text mining and time series analysis techniques can be used to reveal relevant business information: information which can help a decision maker in the company to take the right decisions, but the approaches were not tested yet.

However, we found other related work that can be used for our research. Anjewierden and Efimova [2006] discuss how the digital traces people leave on an online community can be analysed to find interesting information regarding this community. In this paper, the authors present a framework that analyses the digital traces of an online community. They explain and discuss the five dimensions that can be extracted from these traces: people, documents, terms, links and time. The dimensions can be used separately or in combination to answer a number of questions about the community itself by using existing techniques such as social network analysis and text mining techniques. For instance, they explain how the topics of a community can be revealed by analysing the co-occurrences of terms in documents or blog posts in this case. Another example they give is how topics discussed in a community evolve over time.

In the same way, we believe these principles can be applied perfectly to the situation of tags to unlock the explicit knowledge that is hidden in the organizational memory. Therefore, we want to analyse tags to reveal this knowledge, more specifically by building tag dashboards.

### 9.5.2 Tag Dashboards: description and motivation

We propose the main components for the tag dashboards. Of course, more components could be included in the tag dashboard, but we restrict the dashboard to only four. Each of these components is logically connected with the others as we will explain in the paragraphs below.
For each component, we describe how to extract the information and put forward why we choose that component.

**Component 1: evolution of the tags over time**

- **Description:** We calculated the frequency of each tag over time and plotted the data in a scatter plot. We plotted tag use over time in Gnuplot, an open source command-line-driven tool that generates plots. Since we needed custom made plots, for example plots which automatically display the names of the tags in the titles of the plots, we collected all the Gnuplot commands into a bash shell script which runs Gnuplot and executes each Gnuplot command in turn. To repeat this script several times, we added a loop to the script that terminates when all the tags in the input file have been used to generate a plot. We must remark that we excluded from the plot the days on which no messages were created.

- **Motivation:** It is interesting to find out how much a certain tag is used over time as we describe in Van Damme et al. [2008c]. A sudden decrease in tag usage might indicate that something is wrong or an opportunity is arising. It might also reveal certain patterns in buying or selling behaviour, for instance by a sudden decrease in the tag *order* during summer. Another situation could be that the tag *stock* which is always used in January is suddenly annotated to resources in March.

**Component 2: descriptive statistics of tags**

- **Description:** To calculate descriptive statistics about tags, we used the standard statistical functions that are available in MySQL. MySQL offers a large variety of built-in functions which are present in a default installation of MySQL. We opted for the functions: mean, variance, standard deviation, etc. Again, we gathered all these queries into a bash shell script that writes the output into a text file.

---

9http://www.gnuplot.info/
• **Motivation:** Although a plot may visualize an evolution of tag use over a period of
time, it is not always easy to grasp detailed information from a graph immediately,
such as mean, minimum, average or standard deviation. In this way, providing
additional information about the graph makes it more easy to understand the graph.

Component 3: descriptive statistics about the employees’ tags

• **Description:** We calculated additional information regarding the employees and
their tag use, such as which employees and or departments have used the selected
tag subsets, the number of employees, and so on. Again, we used MySQL functions
(e.g. calculate the exact number of employees) as well as the built-in mathematical
operators of the bash shell script to calculate these values. We wrote the results in
a text file.

• **Motivation:** As a manager it might be interesting to know exactly how many em-
ployees are working on a certain project or how many employees are using a certain
tag when annotating a resource. If the number of employees using a particular tag
is too high, this might indicate an inefficient use of resources or employees. There-
fore, we need some statistics on the number of employees who use a particular tag.

Component 4: co-occurrence of tags

• **Description:** We selected the tags of all the resources that have a specific tag in
common and their resource ID, made combinations or tag pairs for each message,
and counted the frequency of each unique tag pair. To create the unique tag pairs,
we used the Perl package or module Combinatorics,\textsuperscript{10} which creates combinations
out of a set of input values. We included the package in our Perl file and added
some code that counts the occurrences of each tag pair. In a next step, we used
Graphviz\textsuperscript{11} to visualize the co-occurrence between tags. The tie strength between a

\textsuperscript{10}http://search.cpan.org/ allenday/Math-Combinatorics-0.09/lib/Math/Combinatorics.pm
\textsuperscript{11}http://www.graphviz.org
tag pair is increased each time that two tags are used together. We decided to restrict the number of tag pairs to four as it is too difficult to visualize more relations in a graph in Graphviz. To allow this step to be executed in one step, we wrote a shell bash script that makes a call to our Perl file, selects the tag pairs with the highest frequency, and runs Graphviz to display the visualization.

- **Motivation**: If a sudden increase or decrease in the use of a tag occurs, it might be interesting to know how this term is used together with other tags to get more insight. The co-occurrence technique is an often proposed technique in the literature on folksonomies (Schmitz [2006]; Specia and Motta [2007]), as already explained in Chapter 4.

### 9.5.3 Example and interpretation of a tag dashboard

In this section, we will discuss the tag dashboard of ProjectX\(^{12}\) shown in Figure 9.3.\(^{13}\)

On the plot, we can easily see that there was some tagging activity on this project at the beginning of the year, but after a while it clearly stopped. During the rest of the year, the tag was only used a couple of times around summer and at the end of the year. It seems like the tag ProjectX is a project which is always postponed to a later point in time.

It is not easy to read the frequency of the tags immediately on the graph. Therefore, it is interesting to look at some descriptive statistics that give additional information. In this case, the tag was used 33 times with a maximum of 5 times a day. The variance and standard deviation are in this case not so interesting, as the tag was not used many times. The fact that the tag was used on average by 12 different employees, but only 33 times in total, is a more interesting thing to know. Many people are aware of the project or are working on this project.

Next to this, we believe that as a manager it is also interesting to know how the tag

---

\(^{12}\)For reasons of confidentiality, we do not refer to the name of the project.

\(^{13}\)More information on the dataset in Section 9.5.6.
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

is used together with other tags. In this case it is mostly used with the tags food, SO,\textsuperscript{14}, \textit{afbakening}\textsuperscript{15}, and \textit{assortiment}\textsuperscript{16}. From the co-occurrence component (Figure 9.3), we can deduce that it was probably a project where an adjustment of the business processes should be taken care of. Also, some marking outs on the project had to be done and it was probably related to the assortment of products, more specifically food.

9.5.4 Implementation of tag dashboards as a performance measurement tool in the company

Because knowledge development is an essential part of the corporate strategy, tag dashboards should be aligned with the corporate strategy so that they can be used as a performance measurement tool. The corporate strategy, which is inferred from the company’s vision, is formulated by the board of directors and executives. To implement the strategy it is translated into a number of actions as well as objectives. However, to create business value from the corporate strategy, information technology (IT) is required (Doughty and Grieco [2005]).

In the past, IT was never considered a priority nor the IT-department was involved in the allocation of their resources. This effected the performance of a company negatively and increased the total cost of ownership. Companies realized that the risks as well as all the goals involved in the implementation of an IT-project had to be managed. Therefore, the IT-governance model became a crucial part of the corporate strategy. The IT-governance model makes sure business value is created for the stakeholders\textsuperscript{17} of the company by managing all the goals as well as the risks related to IT (Doughty and Grieco [2005]).

The IT-governance model is visualized in Figure 9.4 and starts with the stakeholders

\textsuperscript{14}The group of employees that have to evaluate and improve the business processes in the company.
\textsuperscript{15}In English: \textit{marking out}.
\textsuperscript{16}In English: \textit{assortment}.
\textsuperscript{17}Stakeholders are all the people that have expectations from the company as well as those who have responsibility such as shareholders, directors, users and employees (IT Governance Institute [2003]).
in the company who want to attain a maximum business value. To achieve the business objectives and thus create business value, IT is required and should be closely aligned with business. Also, the risks involved in the IT-projects need to be well managed. Finally, we can measure whether the business objectives were met or not (IT Governance Institute [2003]). In literature, the Balanced Scorecards of Norton and Kaplan are often mentioned as a measurement tool to measure or control the objectives (Kaplan and Norton [1996b]).

Tag dashboards could be seen as another tool to measure performance for the board of directors and executives in the IT-governance model. Therefore, in the next section we propose a step-by-step plan of how tag dashboards should be implemented in the company. We believe that tag dashboards can also be used as a performance measurement tool for other stakeholders than the board of directors and executives, because company objectives always have to be translated into objectives for other units and individuals (Kaplan and Norton [1996b]) as we will explain in the next section.

9.5.5 Step-by-step implementation plan

Based on an analysis of literature, we can formulate a number of steps that are required to implement tag dashboards in the company. We use the creation of the balanced scorecards as a starting point to build our tag dashboards. Therefore, we first discuss the steps proposed to create balanced scorecards and then detail our steps to create a tag dashboard.

Steps to create balanced scorecards

There are four steps involved in the creation of a balanced scorecard. We paraphrase them below as described in (Kaplan and Norton [1996a]).

1. **Clarify and translate vision and strategy**: The senior executive team should set up a number of specific objectives derived from the company’s strategy. These objectives, as well as measures for these objectives, should be created for each of
the four dimensions of the balanced scorecard. Kaplan and Norton stress that it is important that consensus on the objectives exists because this is not always the case. Sometimes managers are biased by their previous working experience when setting the objectives. Involving several senior executive managers when deriving the objectives and measures from the strategy makes the process more transparent and simplifies the consensus process.

2. **Communicate and link strategic objectives and measures** It is important that the general company objectives are communicated through the organization. In this way, the general objectives can be translated into a number of local objectives.

3. **Plan and set targets and align strategic initiatives** In order to know whether or not the objectives are met, a number of targets should be set.

4. **Enhance strategic feedback and learning** It is important for managers to receive feedback on their strategy. The balanced scorecards offer them ways to monitor whether or not the strategy is achieved.

Based on the literature of how balanced scorecards should be built (Kaplan and Norton [1996a]) we propose a step-by-step plan\(^\text{18}\) of how a tag dashboard approach should be implemented in a company.

**Step-by-step plan to create tag dashboards for board of directors and executives**

As the balanced scorecard of Kaplan and Norton is already implemented in several companies around the world, we decide to use the steps they proposed to build balanced scorecards as a starting point for our step-by-step plan to create tag dashboards as visualized in Figure 9.5.

The dashboards presented by Kaplan and Norton [1996a] are created from a top-down point of view. There is no clear link or feedback from the employees at the bottom

\(^{18}\)A visualization of the step-by-step plan is shown in Figure 9.5.
to enrich the top-down approach with a bottom-up approach. The only feedback board of directors and executives get from their strategy is the evaluation of the results of the implementation of the strategy. We believe that this top-down approach could be enriched by including an analysis of employees' tags within the creation of these dashboards. As discussed in Chapter 3, tagging is considered as a top-down categorization approach that includes people in the creation of the elements of the categorization. We discuss each of the steps in a little more detail:

- **Step 1: define the business strategy and objectives**
  The creation of a balanced scorecard needs the formulation of a business strategy and objectives/goals for the four perspectives of the scorecard. The same holds for the creation of the tag dashboard: a clear description of the strategy and the objectives derived from the strategy are required. To create a tag dashboard, we need a direction concerning which kind of tags we have to analyse. Therefore, the board of directors and executives need to clearly define the company strategy and objectives, and set targets.

- **Step 2: list of tags**
  Once the business strategy is cleared, the board of directors and executives have to make a list of tags that correspond to the business strategy and objectives defined in the previous step.

- **Step 3: tag cleaning**
  The explicit tag quality can be improved by using the CorTagCleaning approach as we described in Chapter 6.

- **Step 4: remove duplicate tags**
  We merge all the duplicate tags that emerged after cleaning up the tags.

- **Step 5: select relevant resources**
  We query the database to retrieve resources annotated with one of the tags in the
Figure 9.4: IT governance model (IT Governance Institute [2003])
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

tag list. If a resource shares one of the tags we select the message and its tags for further analysis.

• **Step 6: creation of tag dashboard**
  We create the tag dashboard that consists of the four components we elaborated in Section 2. A detailed description of the creation of the dashboard is provided in Appendix C.

• **Step 7: Evaluation tag dashboard**
  The board of directors and executives have to evaluate the four components of the tag dashboards. They have to analyse the results from the tag dashboard and see whether the objectives are attained.

• **Step 8: redefine business strategy**
  Based on the evaluation of the tag dashboards, the board of directors and executives will decide whether or not they will adjust the business strategy.

• **Step 9: push tagging behaviour** An evaluation of the tag dashboards can also reveal that an analysis of additional tags is required or that employees who create the tags should be asked to use some particular tags to improve the quality of the tag dashboards.

**Tag dashboards for other departments**

The strategy formulated by the boards of directors and executives needs to be communicated throughout the organization as well as the objectives they want to achieve. To implement the strategy, it must be translated into concrete actions for every department within the company as well as all the different hierarchical levels within a company. For each department as well as every hierarchical level, objectives are derived from a higher level. This implies that every department as well as every hierarchical level requires their own tag dashboards.
9.5.6 Methodology

To research whether tag dashboards can be used to unlock explicit knowledge, we created tag dashboards on a sample of the dataset and did five interviews to answer the following related research questions:

- Can tag dashboards help to unlock the explicit knowledge that resides in the organizational memory?
- How can the tag dashboards support a manager\(^{19}\) to do his work?
- How suitable are the components of the dashboard?

In this section, we detail the components of the dashboard. Then, we describe the dataset we used to prepare the interviews. In the last section we entail the interviews we did to test the approach as well as the results.

Dataset

We took the tag dataset of the *Company* to create the tag dashboards. In 2006, messages were created on 252 days during 2006. First, we only used these 252 days to plot the activity of tags. Second, we took only the tags created by two functional groups (2AF and K38) to create the tag dashboards.

To determine the number of employees who used a certain tag, we assumed that the employee who actually created and annotated the messages was the owner of the tag. As already explained in Chapter 5, a message can have more than one author in the *Company*. For example when someone replies to a message, that person also becomes the author of that message. When this occurred we took only the employee-ID which corresponded to

\(^{19}\)For the interviews, we used the term *managers* because the interviewees in the *Company* have the job title *manager*
the dactylo ID\textsuperscript{20} and did not take the other employee-IDs into account in calculating the employees’ statistics.

**Interviews**

To answer the research questions which we formulated at the beginning of this chapter, we conducted five semi-structured interviews with managers\textsuperscript{21} within the Company. A detailed description of the preparatory steps for the interviews as well as the description of the creation of the tag dashboards is provided in Appendix C.

**Results**

In the paragraphs below, we discuss the answers given by the interviewees to our questions.

- **Do you believe these dashboards could help you do your daily work?**

  Four of the interviewees were not aware of our research and regarded tags as a tool to retrieve the messages from the communication system. Therefore, we first had to explain to them how tags can be used and analysed from a different point of view. It was, however, not always very easy for them to answer our questions.

  For instance, at first one of the interviewees did not believe that the tag dashboards were a good idea because he thought he had other means of doing the same thing. However, after looking at a number of other dashboards he realized that his first reaction was wrong and that this could indeed be an interesting means for him. It provides information he currently does not have or is not available in the company.

  There was only one interviewee who was not convinced by the tag dashboards at all, not even by the end of the interview. He said, “\textit{I can easily retrieve the information}...”

\textsuperscript{20}A description of the dataset of the Company is provided in Chapter 5.

\textsuperscript{21}In the Company, a manager is an employee who is responsible for a subgroup within the functional group.
which is displayed in such a dashboard by searching for messages in the communication system: “I only need to select some keywords and the name of an author to find the messages. Once I get the messages I read them to get the information I need”. The fact that such a dashboard is created after something has happened is not a good thing, he argues. “My employees should inform me immediately when there is some kind of a problem or I should already be aware of the problem. A tag dashboard is created when it is already too late”.

The three other interviewees also considered the tag dashboard a valuable way of helping them be aware of certain issues that are going on in the company. They believe the dashboards can help them do their jobs as managers. Thus, out of the five interviewees there was only one interviewee who disagreed.

We conclude that a majority (four out of five) of the interviewees agreed that the tag dashboard could help them do their daily work, because it provides them knowledge on issues they do not currently have.

- **How do you think these dashboards could help you?**

Every interviewee gave different suggestions about how the tag dashboards could help him to do his daily work.

For instance, one of the interviewees remarked that tag dashboards could help them to overcome the overload of messages they get as managers each day. In the Company, every message which is distributed to employees in a department should also be sent as a copy to the manager. Sometimes managers receive more than 150 messages as copies and do not have the time to read them all. “The tag dashboards could help us indicate what is actually going on and help us to decide which messages we definitely have to read”.

---

22 Since one interviewee was not interested in the tag dashboards, we exclude this person from the results when discussing the answers of the other interviewees throughout Section 9.5.6.
Two other interviewees answered that tag dashboards could help them to verify whether employees are actually using appropriate or high-quality tags. “In my department, I asked the employees to use some specific tags to label messages to make it more easy to retrieve the messages later, but I noticed that not everyone obeys this rule. These tag dashboards could help me monitor the tag use.” Alongside this, they remarked that it would also help them to check whether expectations correspond with reality. “When my team is working on a specific project, I expect to see some activity on the tags related to the project in the dashboard. If I cannot find any or the tag frequency is below expectations, it might help me to identify that there is a problem”. The other interviewee also added: “I always know when I should expect to see more or fewer reports on specific projects, because I always keep a plan of the projects my employees are working on. These dashboards could help me verify whether their tagging activity corresponds with the plan I made”.

Another interviewee suggested that these dashboards could be interesting for a couple of purposes. As a first use, he suggested project planning: “It gives a manager information on how a project is going on”. As a second use, he suggested efficiency measurement: “How many people are using a certain tag? How long are they using a certain tag for?”. Next, they could also be used to measure the collaboration between departments by looking at the tags they are using in combination. Mining for important topics in the company that need to be further investigated was the fourth purpose he mentioned. He also added that the CEO of the Company was looking for a tool which could help him to measure the level of activity around topics discussed by employees in the Company.

We summarize their answers in Table 9.1

- **What do you think of the components of the tag dashboard?**

Four of the interviewees agreed with the components of the tag dashboards, although some interviewees considered specific components to be more important
Help managers decide which kinds of messages/topics they need to keep up with
Monitor tag use regarding tag rules
Monitor status of projects and project planning
Tag efficiency measurement
Measure collaboration between departments
Measure the activity level activity around topics discussed in the company

<table>
<thead>
<tr>
<th>Table 9.1: Purposes for the tag dashboards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help managers decide which kinds of messages/topics they need to keep up with</td>
</tr>
<tr>
<td>Monitor tag use regarding tag rules</td>
</tr>
<tr>
<td>Monitor status of projects and project planning</td>
</tr>
<tr>
<td>Tag efficiency measurement</td>
</tr>
<tr>
<td>Measure collaboration between departments</td>
</tr>
<tr>
<td>Measure the activity level activity around topics discussed in the company</td>
</tr>
</tbody>
</table>

The answers to this question were closely related to the answers we received to the previous question.

One of the interviewees argued that the co-occurrence component is a vital part of the dashboard. “It allows you to check whether people are using tags in combination or whether an important pair of tags is missing.”

All the interviewees suggested making some changes to the components, and we summarize them below:

– Visualize the tag use not on a daily basis, but on a weekly or monthly basis.
– Detail the use of tags on the level of the team instead of on the level of the functional group.
– Compare the results of the tag dashboards with the previous periods and indicate whether there is an increase or decrease in tagging activity.
– Create a tag dashboard based on a cluster of tags.
– Include a drill-down visualization that takes into account the employees who have used a particular tag frequently. By clicking on a specific tag in the co-occurrence visualization, a network of employees who have expertise on this tag should be shown. Then, the manager can drill down to get a network of tags that are frequently used by this particular employee.

Instead of letting managers evaluate every tag dashboard, two interviewees suggested creating only the tag dashboards from the tags that exceed a certain kind
of lower or upper bound value. “By using some kind of trend analysis, the system could send an alarm message when a tag exceeds a certain upper or lower bound value. Instead of evaluating every tag dashboard manually, which can sometimes be very time-consuming, an evaluation can be done automatically.”

- **Would you consider implementing a tag dashboard in your department?**

Except for one interviewee, all the participants liked the idea of implementing it in the company. One of them answered: “It is a cheap system that allows you to create some metrics very easily. Based on these metrics you can decide whether or not you want to build expensive systems that are able to generate more accurate results.”

- **Do you agree with our step-by-step plan to create tag dashboards?**

We briefly verified our step-by-step plan with an interviewee who not only was a manager of a group of employees, but also supervised the other managers participating in the interviews. He was more concerned with the strategy of the company than the other interviewees were. Therefore, we will discuss only his answer.

He believed that the framework could be implemented in two ways in the Company. Tag dashboards could be implemented as a tool to measure the activity around certain topics on the one hand and to monitor the implementation of projects on the other hand.

To measure the activity around certain topics, we can easily use Steps 2 to 7 of the framework, the interviewee explained. We do not need any additional steps. This is however not the case in the other situation.

The strategy within the Company can be split into three types of goals: strategic, tactical, and operational, and each set of goals has a different time frame. Strategic goals are set for a period of five years, tactical goals for a period of one to three years, and operational goals for a period of one year. A number of projects are always derived from the operational goals. For each project a set of tags could be
derived to describe the project. Once we got the list of tags, he agreed with the use of Steps 3-7 to create and evaluate the tag dashboards. However he did not believe that an evaluation of the tag dashboards would result in a change of the corporate strategy. He believed it could influence the operational goals; for instance: “we will extend the project by one or two years”, or “we have to stop the project, but we will not influence the strategy at all”. It could also be that the list of tags will be changed because a part of the information is polluted.

Limitations

We provide a number of limitations related to our research:

- Most of the interviewees were not aware of the research and it took some time for them to understand everything.

- The interviews took only thirty minutes and only five managers were involved. However, we should mention that it was not easy to schedule a lot of interviews with managers because they were very busy. Also, interviewing employees in a company implies a significant cost.

- The approach was not derived from the corporate strategy, because it is not yet implemented in the Company.

9.6 Conclusion

We discussed how tag dashboards can unlock explicit knowledge in the company and how they can be used as a performance measurement tool in the IT-governance model for the boards of directors and executives to measure a company’s objectives. In addition, we explained that tag dashboards could be of interest of every department as well as hierarchical level.
To test the concept of tag dashboards, we had interviews with five managers in the Company. Four out of five interviewees agreed that the tag dashboard could be used to unlock explicit knowledge. There was only one out of five interviewees who was not in favour of the tag dashboards. However, we can easily explain why one of the interviewees did not believe these tag dashboards could be an added value to him. In case the board of directors and executives would use the IT-governance model to communicate their objectives top down to all the other employees as we proposed in this chapter, the employees would be well aware of why and how they have to use these tag dashboards. The tag dashboards are more than merely a result of analysing employees’ tags: they are closely related to the company’s objectives. Employees could be stimulated to use these dashboards by using appropriate incentives.

The other managers gave several ideas of how these dashboards could help them, as we summarized in Table 9.1. Moreover, they liked the idea of implementing such a tool in the Company and four out of five interviewees agreed that the dashboards could provide them with information or knowledge they did not currently have. Regarding the components of the dashboards, the four managers agreed on the components of the dashboards, but they suggested some small changes to each of these components.

We conclude that the tag dashboards can help to unlock the explicit knowledge that is in the organizational memory to enable business value. However, as we mentioned in the limitations section, we only briefly validated the approach. A more profound testing is only possible once the dashboards are implemented in the Company. We believe this is interesting material for future research as we will elaborate in Chapter 10.

In addition, we conclude that the tag dashboards could be used as an information retrieval tool instead of merely an explicit knowledge retrieval tool. Employees could use the tag dashboards as a tool to retrieve information more easily for instance by drilling-down one of the components of the dashboards. This is an interesting topic for future research.
CHAPTER 9. ENABLING BUSINESS VALUE BY UNLOCKING EXPLICIT KNOWLEDGE THROUGH TAGS

Step 1: Define the business strategy and objectives
Step 2: List of tags
Step 3: Tag cleaning
Step 4: Remove duplicate tags
Step 5: Select relevant resources
Step 6: Creation of tag dashboards
Step 7: Evaluation tag dashboard
Step 8: Redefine business strategy
Step 9: Push tagging behavior

Figure 9.5: Step-by-step plan to implement tag dashboards in a company.
Part IV

Results
Chapter 10

Results

10.1 Introduction

In this section, we first discuss a number of prerequisites that we believe are required when a company wants to use its tags to unlock corporate knowledge. We detail them and reason why they are important\(^1\) in the following sections. For each prerequisite we formulate a number of recommendations\(^2\) in order to help a company attain the prerequisites. At the end of the chapter, we also discuss some additional recommendations for companies that want to unlock corporate knowledge that resides in the organizational memory through tags.

\(^1\)These prerequisites are based on the ones we suggested in the paper we presented at the OBI2008 workshop in conjunction with the ISWC2008: C. Van Damme. Approaches to Analyse Corporate Tags for Business Intelligence Purposes. In *Proceedings of the First International Workshop on Ontology-supported Business Intelligence (OBI2008) in conjunction with ISWC2008*, Karlsruhe, Germany, ACM Digital Library. 2008

\(^2\)Figure 10.1 contains a visualization of all the prerequisites and their recommendations.
10.2 Prerequisite 1: quality of tags

The first prerequisite a company should take into account is the quality of tags. As we discussed in Chapters 2, 3, 4 and 7, tagging does not have a built-in control mechanism which checks the quality of the chosen tags. When a company wants to use the tags to unlock corporate knowledge, it has to be certain that the tags created by the employees are reliable. For instance, in Chapters 6 and 7 we discussed the results of measuring respectively the explicit and implicit tag quality. We concluded that many of the tags in the dataset were polluted and thus had a low explicit tag quality. However, we were able to improve the explicit quality of tags by using the CorTagCleaning approach. After we improved the explicit quality of the tags, we measured the implicit tag quality of tags annotated to a sample of messages. We concluded that the implicit quality of the tags
in the sample was good. In addition, we concluded that the messages have significantly more author tags of good quality than author tags of a poor quality. However, the implicit tag quality can still be improved: messages have significantly more high-quality tags than author tags of good quality. Based on these results, we will formulate a number of recommendations for explicit as well as implicit tag quality (Figure 10.2).

10.2.1 Prerequisite 1a: explicit quality of tags

As we explained in Chapter 6, there are two ways to improve the explicit tag quality: introduce tag rules as is the case in the Company or tidy-up the tags automatically.

Recommendation prerequisite 1a: tag rules

Although the Company asked its employees to obey a number of tag rules, many of the tags were polluted as we concluded in Chapter 6. Many tags did not correspond to the tag
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

Prerequisite 1:

Quality of tags

Explicit quality of tags

Tag rules

Clean tags automatically

Implicit quality of tags

Type of folksonomy

Fitness for tag use

Broad folksonomy

Narrow folksonomy

Recommendations:

1) minimize the number of tag rules: only when it can be applied automatically
2) Implementation of a tag rule control mechanism

Recommendations:

CorTagCleaning approach

1) revise and adjust the approach if necessary
2) extend with subject-based classification system → create lightweight ontology through analysis of corporate tags

Recommendations:

a company should be aware of the fitness for tag use: the motivation for an employee to tag

Recommendations:

a company should use the high-frequency or TF-IDF tag quality algorithm to select high-quality tags

Recommendations:

1) implicit quality of narrow folksonomies should be measured manually or narrow folksonomies should be made broad
2) create suggestions based on words in the resource to improve implicit quality of tags

Prerequisite 2:

Integration of tagging into the working processes

Recommendations:

Tagging should be completely integrated into the working processes of the company

Recommendations:

1) give training to the employee
2) transparency: explain why tagging is important
3) social control as incentives
4) financial incentives provided that there is control mechanism

Prerequisite 3:

Involving all the employees

Figure 10.3: Overview of prerequisites and recommendations
quality dimension *consistency* as we explained in Chapters 6 and 7.

In our opinion, tag rules are not always required because many of the polluted tags can easily be cleaned up by using the approach we presented in Chapter 6. In addition, the free spirit of tagging can be considered as a strength of tagging. Asking people to think about the format of the tags they choose, for example *do not use conjugated verbs* implies an additional cognitive cost for them and of course incurs a higher total cost of ownership for the company. Therefore, the number of tag rules should be minimized and mainly be used when an automatic approach is not possible.

However, not all of the polluted tags can be cleaned up automatically. In such cases tag rules would be an added value. For example, in Chapter 7 and Appendix A we discussed some of the post-cleaning steps we had to take to make an analysis of the spontaneous reader tags and author tags possible because an automated approach for some of the problems was not possible. For example, we noticed that every employee had his own way of writing down a *date* or *compound words*. Some of them used several tags and others just wrote down one word. A tagging rule which asks the user to use a consistent format of writing down a *date* or a *compound word* would be recommended.

Therefore, we advise that a *company should minimize the number of tag rules it wants to implement*. It should only take a tag rule into account that cannot be applied automatically on the tags. Next, it is also important that there is *some kind of control mechanism*, because we noticed that the tag rules of the *Company* were not well followed. Therefore we assume that a control mechanism is required to verify whether users obey the rule or not. A control mechanism could be implemented by letting a script check the tags before they can be saved, for example, checking whether tags only contain digits. In cases where the tags are not in the appropriate format, a warning message could be shown and the tags could not be saved.
Recommendation prerequisite 1a: clean up tags automatically

In Chapter 6, we presented the CorTagCleaning approach to clean up corporate tags automatically and discussed how many tags we cleaned-up. Although the approach was based on the dataset of the Company, we believe it could easily be applied on the tag dataset of other Dutch companies because it does not contain steps that are company-specific. We assume that a revised approach will be necessary to apply it to other languages because the approach contained some steps that were specific for the Dutch language. For example, the c/k rule. It could be that these rules should be replaced by other rules that are language-dependent.

We recommend companies to use our CorTagCleaning approach as a starting point and check whether the language-dependent rules are applicable to their situation and adjust it in case it is needed. By letting a script run on the tag dataset every night, the tags can be tidied up without having an impact on the performance of the employees nor on the tag application.

However, we must remark that our CorTagCleaning approach did not remove all the polluted tags. For instance, we did not handle the synonyms and homonyms, as we explained in Chapter 6. Solving the synonyms and homonyms within the Company would require a subject-based classification type such as an ontology.

In Specia and Motta [2007], the authors explain how tags can be enriched with one of the elements of an ontology: concepts, instances or properties. First, they suggest finding out whether there exists an ontology that contains both tags of the tag pair. Then, all the elements of the ontology have to be selected as well as all the additional information such as the parents and domain or range. Based on the information extracted from the ontology, the relationship between both tags can be retrieved. A similar approach could be used to enrich the corporate tags. As every company has its own company-specific terminology, a company needs its own ontology. Of course, other subject-based classification

---

3 Of course, except for the synonyms used in Chapter 7. An overview of the synonyms used in Chapter 7: Appendix A.
systems such as a thesaurus or a taxonomy could be applied in a similar way. **We would recommend companies to extend the CorTagCleaning approach with a subject-based classification system if one is available.**

In case a company does not have any subject-based classification type, we believe one could be created through an analysis of tags. In Van Damme et al. [2008c], we presented the FolksOntology approach to derive ontologies from folksonomies. The approach includes several steps such as statistical techniques on the data generated through the tagging process, exploiting existing online resources in several ways, and involving the community to achieve and maintain consensus. Such an approach could help to build a lexicon base, a set of elementary building blocks that can be reused to build an ontology. This is consistent with the Business Semantic Modeling (BSM) approach proposed in De Leenheer [2009].

However, the FolksOntology approach cannot be directly applied to corporate tags as we explained in Van Damme et al. [2008a]. For instance, labor costs are very high and therefore the number of employees involved with the feedback process should be minimized. In contrast to web communities it is far easier to ask the cooperation of the community: community members have a different mindset than employees and are more willing to participate in additional processes. In most cases they are anonymous. Company-specific terminology is mostly used in a closed company environment which makes it hard to include web resources in the ontology construction process. The terminology may contain terms which have a specific meaning for only a small group of employees.

- **Create a lightweight corporate ontology through tags**

Therefore, in Van Damme et al. [2008b], we presented an approach to create a lightweight corporate ontology through an analysis of tags. However, we must remark that this approach did not yet contain an extended approach to tidy-up tags.

---

4 This section is largely based on paragraphs of: Van Damme et al. [2008b]
as we proposed in this dissertation. The research presented in Van Damme et al. [2008a] was done at an early stage of our research and we did not have an extended methodology to take the quality of the tags into account. We briefly discuss some of the other steps we suggested in Van Damme et al. [2008b] to create a lightweight ontology:

1. **Calculate the co-occurrence**

2. **Finding broader/narrower relations through calculating the conditional probability.** The conditional probability is calculated by dividing the co-occurrence of the tag pair by the frequency of the individual tag’s. Results vary between 0 and 1. The higher the result, the more the term is used in combination with the other term and consequently the more dependent it is of the other term. When the difference between the two results exceeds a certain threshold value, a subsumption relationship is found. Finding broader and narrower terms is important to derive hierarchical relations.

3. **Transitive reduction and visualization.** Transitive reduction reduces the edges of a graph G to a graph G’ by keeping all the paths that exist between the nodes in Graph G (Aho et al. [1972]). The edges are consequently removed because of the implied transitivity.

Of course, this was a basic approach that needed to be enriched in order to create a more complete ontology. We tested this approach on a sample of the Company and asked our contact person in the Company for feedback. The obtained results reflected the terminology used in the Company (Van Damme et al. [2008b]) but still needed to be extended in order to become a rich ontology.

*We recommend to use this approach as a starting point to create a corporate ontology and to extend it with the CorTagCleaning approach. Next, we suggest saving the results obtained through the tag pair verification in the Levenshtein*

---

5Definition **Co-occurrence**: Glossary.
to create a synonym list. In this way, the number of tag pairs in the Levenshtein pop-up window would decrease a lot when similar words would be saved.

To create a more rich ontology, we believe human feedback is required. Of course, a trade-off should be made by the company between the total cost of ownership to enrich the ontology and the hidden benefits it can create.

10.2.2 Prerequisite 1b: implicit quality of tags

As defined in Chapter 4, high-quality tags are not only well-cleaned up tags, but the tags also have to be in line with the tag quality dimensions: accuracy and completeness. We explained in Chapter 4 that a distinction between broad and narrow folksonomies should be made to measure the implicit quality of tags.

Recommendation prerequisite 1b: broad folksonomies

Research discussed in Chapter 4 showed that the high-frequency and TF-IRF tag quality algorithms were the most appropriate techniques (from the set of three proposed algorithms) to automatically select high-quality tags in a broad folksonomy. Of course, other algorithms might be valuable, but we only tested three of them on a del.icio.us dataset. We believe these results are also applicable to a corporate tag dataset. Based on the results of our research, we recommend that a company should use the high-frequency and TF-IRF tag quality algorithms to automatically select high-quality tags in a broad folksonomy.

Recommendation prerequisite 1b: narrow folksonomies

In the case of a narrow folksonomy, we believe it is very difficult to select high-quality tags automatically. Human feedback and input is still required. Our results showed that
we were not able to automatically select high-quality tags based on a mining of the textual resource. Of course, we could make the narrow folksonomy broad by letting several employees annotate the same resource. But again, this could increase the total cost of ownership for the company.

In Chapter 7, we concluded that suggested tags retrieved from words of the message are an interesting feature to improve the implicit quality of tags in the sample. However, we believe that the use of suggestions should be implemented with some care. Research from Sen et al. [2006] explains/showed that people are influenced by the tags they see. Therefore, we recommend companies to show the suggested tags after the user has given his own tags. It is important that the suggestions are based on a corporate ontology or another subject-based classification technique.

**Recommendation prerequisite 1b: fitness for Tag Use**

We did not take fitness for tag use into account when we measured the implicit quality of tags. The fitness for tag use is a term which we derived from the information systems domain. In the information systems domain the quality of data is commonly described as fitness for use (Kumar and Ballou [1998]). This implies that the quality of data depends on its use. Data which is considered to have a good quality for one purpose may have a low quality for another use. For instance, the human resources department in the company will need all kinds of personal information about the employee, for example, address and bank account number, whereas the company’s restaurant is more interested to know the food preferences of the employees. We believe that the same thing holds for the quality of tags and therefore we want to introduce the term fitness for tag use: the incentive or motivation of a user to tag might have an impact on the type of tags used to annotate a resource as we described in Chapter 3.

We were not able to test whether fitness for tag use had an impact on the corporate tags of the Company. Currently the Company uses its tags for retrieving its messages
more easily. Asking the employees to select tags for different purposes (for example finding experts or creating tag dashboards) would probably be very hard for them to do. At the moment of research, we did not find any publication that has measured this effect on corporate tags or tags created on the World Wide Web. It would be an interesting topic for future research: the impact on corporate tags when taking the tag purpose into account.

Therefore, we believe that a company should be aware of the fitness for tag use when it decides to implement tagging as it might have an impact on the type of tags that employees select.

10.3 Prerequisite 2: integration of tagging into the working processes

To unlock corporate knowledge through tags, it is important that there is a close linkage with the working processes. All the information that is shared with other employees or customers should be tagged. In that way a complete picture of the reality, or of what is going on in the company, can be reflected by the use of tags. We recommend that tagging should be completely integrated into the working processes of the employees.

10.4 Prerequisite 3: involving all the employees

In Chapter 9, we studied how tag dashboards derived from corporate strategy and objectives can be used to unlock explicit knowledge in the company. To create these tag dashboards employees have to be involved in this process, because they have to annotate the documents with these tags.

An important problem related to corporate tagging is how employees can be con-
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

vinced to annotate a resource. This is an issue we did not address yet in this dissertation because tagging is completely integrated into the corporate culture of the Company. Employees simply have to annotate a message and there is no way for them not to participate. A message cannot be sent without tags. There is, however, no control mechanism to verify the quality (explicit and implicit) of the tags as we already discussed. We assumed that there was a principal-agent problem that would result in tags of a poor quality, but this was not the case as we concluded in Chapter 7.

When we asked the employees for their tag experiences\(^6\), five out of 29 interviewees explicitly stated that they consider tagging as a difficult task because they always try to keep the audience of the messages into account. In this situation, we could say that social control creates incentives for these employees to choose high-quality tags. Of course, additional interviews would be required to conclude that social control is indeed an incentive in the Company. This was not a question we explicitly asked during the interviews. But these results correspond to the ones obtained by Thom-Santelli and Muller [2007].

Therefore, we believe that social control should be recommended as an incentive for employees.

Transparency is another incentive we recommend. Employees should be informed about the importance of annotating corporate resources. In addition, they should be convinced that they benefit from tagging the resources, as well as the other stakeholders of the company. Letting employees label corporate documents increases the total cost of ownership and the cognitive cost for an employee. Moreover, the working processes of an employee are effected and the employees have to adjust to the new processes.

To implement tagging more smoothly in a company, we also recommend to give training. It is important that everyone in the company knows how they have to tag documents. In case of a question, the help desk should always be available. The usability of the tagging feature, is also an important recommendation. A poor usability will disencourage employees to tag.

\(^6\)Interviews discussed in Chapter 7.
Since most companies use financial incentives to stimulate employees to increase their performance, we assume that financial incentives are interesting means to encourage employees to tag. However, we recommend to use financial incentives together with control mechanisms to prevent employees from cheating.

10.5 Recommendations

In this section, we provide some additional recommendations specifically related to unlocking knowledge in the company.

10.5.1 Recommendations to unlock tacit knowledge from tags: expert finding

Based upon our findings we discussed in Chapter 8, we are able to formulate a number of recommendations:

- To find experts based on an analysis of tags, we also need to withdraw the general tags. Cleaning tags is not sufficient.

- A time factor should be taken into account when analysing the tags. We believe that a time factor can be easily taking into account by using some weights, for example giving more weight to more recent tags and vice versa.

- Based on the results of the research, we suggest social networks to be an interesting means for new employees in the company to give them an idea about the employee structure in the company.

- Unlocking tacit knowledge through tags could help a company to adjust its human resources strategy. It helps the board of directors in a company provide information.
such as who are the central people within the company and whether or not a back-up strategy, in case they would be leaving, is required.

**10.5.2 Recommendations to unlock explicit knowledge from tags: tag dashboards**

In Chapter 9, we discussed how explicit knowledge can be unlocked through tags by creating tag dashboards derived from the business strategy. We also detailed how the tag dashboards can be part of the IT-governance model as a performance measurement tool to enable business value. *Therefore, we recommend a company to use the step-by-step plan we presented in Chapter 9 to implement tag dashboards. In addition, we recommend a company to integrate the tag dashboards as a part of the IT-governance model.*

*We also recommend a company to start from the components we suggested in the dashboard and revise them in order to fit with the company structure.* During the interviews, the managers gave us some interesting adjustments that could be made, but we believe that these adjustments are related to the company itself.

**10.6 Conclusion**

We presented three prerequisites that we believe are required to implement tagging as a tool to unlock knowledge in a company. For each of these prerequisites we formulated a number of recommendations that can help a company to attain them. In addition, we also formulated other recommendations that are directly related to unlocking knowledge through tag analysis.
Part V

Conclusion
Chapter 11

Conclusion and future research

11.1 Introduction

We researched how corporate tags can be analysed to unlock knowledge in the organizational memory and how this affects the business strategy. To do this, we formulated a number of related research questions and answered them by doing a case study research at the Company. We discuss the answers to each of these research questions in more detail below. At the end of the chapter, we give a number of research topics interesting for future research.

11.2 Question 1: how can we select high-quality tags in a broad and narrow folksonomy?

We defined high-quality tags as “well cleaned-up tags that correspond to the three tag quality dimensions: accuracy, completeness and consistency”. To retrieve tags that are in line with this definition we need to make a distinction between explicit and implicit tag quality, as we discussed in Chapter 4. More specifically, we have to be certain that
the tags are cleaned up well, or the explicit tag quality is being checked, before we can measure the implicit tag quality.

In Chapter 6, we presented our CorTagCleaning approach to verify and improve the explicit quality of corporate tags. It is a 7-step approach that is specific for the Dutch language. As we explained in Chapter 10, this approach can easily be revised in order to be applicable for other languages and other companies. We opted for an approach to tidy-up tags, because the tags in the dataset were polluted as explained in Chapter 6 even though the Company had specified seven tagging rules in order to have clean tags.

To measure the implicit tag quality, we explained that we have to make a distinction between broad and narrow folksonomies (Figure 4.1). In Chapter 4, we explained that high-quality tags can be automatically selected in broad folksonomies because of the effect of wisdom of the crowds. Therefore, we investigated in Chapter 4 three tag quality algorithms: high-frequency tag quality algorithm, tag agreement tag quality algorithm and TF-IRF tag quality algorithm. By doing a survey with 20 students, where each student had to evaluate 10 sites, we concluded that the high-frequency as well as the TF-IRF tag quality algorithm were the most appropriate ones to select high-quality tags automatically in a broad folksonomy.

In the case of a narrow folksonomy, we tried out whether high quality tags can automatically be selected by mining a text for frequently used words\(^1\). We tested this approach on a sample of messages of the Company and compared its results to the ones we obtained through measuring the tags manually. Even though, the approach enabled us to automatically select high-quality tags, the list of suggested tags also contained many tags that cannot be considered as high-quality tags. The number of suggested tags that cannot be considered as high-quality tags was, on average, significantly larger than the number of suggested tags of good quality (p=0.000, \(p<\alpha, \alpha = 0.05\)).

We believe that manual feedback will always be required to verify the implicit quality

\(^{1}\)We opted for a frequency of at least two.
of tags of a narrow folksonomy because:

- We believe computers do not have the same reasoning power that human beings have (Sowa [1984]). Research discussed in Al-Khalifa and Davis [2007] also showed that humans were better than the text mining technique they presented to retrieve keywords.
- only one user annotates the resource. There is no guarantee that this person is the smartest of the group as defined in the wisdom of the crowds (Surowiecki [2004]).

11.3 Question 2: how good is the implicit quality of corporate tags?

We measured the implicit quality of the tags annotated to a sample of messages of the Company in Chapter 7. By means of 29 interviews we asked every interviewee:

1. to tag a sample of 10 messages\(^2\) with high-quality tags
2. to evaluate the original tags chosen by the author of the messages in terms of being a high-quality tag. Therefore, they had to select a value on a five-point scale: very relevant, relevant, mediate relevant, not so relevant, not relevant,
3. to select additional high-quality tags from a list of suggested tags,
4. to answer a questionnaire of 10 questions concerning their experiences with tagging.

Based on the answers from the interviewees we calculated a number of measures as explained in Section 7.5.1. Results from the sample are discussed in Table 7.1. To obtain more information regarding the population of messages we tested a number of hypotheses.

\(^2\)Every message was annotated by six interviewees on average.
We concluded that the implicit quality of the tags was good: the average number of author tags of good quality was, on average, significantly higher than the average number of author tags of poor quality ($p=0.000$, $p<\alpha$, $\alpha=0.05$). However, the quality of the author tags can still be improved because the number of high-quality tags per message were, on average, significantly higher than the number of author tags of good quality ($p=0.000$, $p<\alpha$, $\alpha=0.05$). In addition, the number of author tags of good quality was significantly smaller than 6 ($p=0.000$, $p<\alpha$, $\alpha=0.05$). Results from doing a regression analysis showed that the relative number of high-quality tags was slightly negatively influenced by the length of a message.

Since the approach to select high-quality tags automatically in a narrow folksonomy was insufficient, we investigated whether this approach would be suitable to help employees when they have to label a resource and thus increase the implicit (and of course the explicit) quality of the tags.

We concluded that on average more than 3 high-quality tags were retrieved in the list of suggested tags for messages of the population ($p=0.0195$, $p<\alpha$, $\alpha=0.05$), and 72 per cent of the interviewees liked the quality of the suggested tags. Therefore, we concluded that the suggested approach might be an interesting means to help the employees select high-quality tags for a message and thus improve the quality of the tags.

11.4 Question 3: how can an analysis of employees’ tags help a company to unlock the tacit knowledge that resides in the organizational memory?

To unlock the tacit knowledge, we suggested finding experts as an approach, more specifically finding experts through an analysis of tags. We opted for this approach since in literature experts are considered to be a source of tacit knowledge. To test our approach
we used three tag analysis techniques (high-frequency, cumulative frequency and TF-IPF) as well as the visualization of social and ego networks that are formed around tags to derive experts within the company.

To validate the approach, we first applied it to a sample of tags created by 492 employees of the Company and then presented the results to ten employees whose tags belonged to the sample. In the first part of the interview they had to choose the tag analysis techniques that reflected their expertise best. Then they had to evaluate the social network as well as their ego networks.

Based on their answers, we concluded that our approach enables employees to find experts in the company. We briefly summarize our findings below:

- All of the employees agreed upon the fact that the expertise of employees is well-reflected through an analysis of tags.
- Seven interviewees preferred the high-frequency technique. However, every interviewee preferred to select more than one tag analysis technique when they were asked to select the set of tags that reflected their knowledge best. Five out of ten preferred a combination of high-frequency and TF-IRF tag analysis techniques. The fact that some tags could be considered to be more general tags was the reason why they preferred a combination of techniques.
- The ego networks reflects well those with whom the employees are in contact, but there were only three out of eight employees who stated that ego networks could help them find colleagues with similar expertise.
- All of the interviewees believed that the tags obtained through the analysis should be visualized in a tag cloud.

Based on the results from the interviewees, we concluded that unlocking tacit knowledge through tags provides valuable knowledge for the board of directors and executives.
to improve their HR-strategy. It provides them interesting knowledge on who are the central employees and whether or not these employees could easily be replaced by similar experts

11.5 Question 4: how can employees’ tags help a company to unlock the explicit knowledge that resides in the organizational memory and how does this influence the business strategy?

In Chapter 9, we presented a tag dashboard approach, an approach derived from the one of Kaplan and David [1992]. In literature, dashboards are often proposed as a means to get an overview of all the processes that are going on in a company. It can be considered to be a visual summary populated with structured information from different information systems. However, most of these systems are not capable of capturing the knowledge hidden in unstructured information. And when this is the case, computers are said not to have the same reasoning power as human beings have. Since tags can be considered to be a summary of the annotated resource and are created by a human being, we proposed a tag dashboard approach that is built upon results from tag analysis methods. Figure 9.3 shows an example of such a tag dashboard. We opted for four different components. Each of them was based on a different tag analysis method and was intended to focus on another aspect.

To test the validity of the tag dashboard approach as well as determine the importance of the different components of the dashboard, we did five interviews with managers within the Company. For each of these managers, we created tag dashboards around tags that were in line with their job function and description.

Based on the answers from the interviewees, we concluded that the tag dashboards
can help to unlock the knowledge that resides within the organizational memory. Four out of five managers agreed with the tag dashboard approach and would like to see the approach implemented in the Company. There was only one manager who disagreed, because he did not think such a system would provide new knowledge.

We summarize the key findings obtained by the four other managers below:

- The tag dashboards could help a manager in his daily work for different purposes such as: tool to decide which kinds of messages/topics they need to keep up with, project planning, a tool to measure collaboration between different departments.

- They all agreed upon the components of the dashboards, but they suggested some small changes to each of these components.

In addition, we discussed how tag dashboards should be derived from the corporate strategy. More specifically, we detailed how the tag dashboards can be part of the IT-governance model as a performance measurement tool to enable business value. Based on the results presented in the tag dashboards the board of directors and the executives in a company can verify whether the corporate objectives are met. It also provides them insight whether the corporate strategy should be adjusted.

In Figure 9.5, we presented a step-by-step plan of how tag dashboards should be created and linked with the corporate strategy. We explained also how the outcome of the dashboards might effect the corporate strategy.

Tag dashboards are not only a means for the board of directors and executives. Since the business strategy is communicated throughout the organization, other departments have to formulate their own strategy and set their own objectives. Consequently, every department will have to create their own tag dashboards to evaluate their objectives and of course on every hierarchical level, and this will influence their strategy. In short, tag dashboards are an interesting means for every employee who has to attain certain objectives. Each one of them can select their own tags and generate the tag dashboards.
In addition, the results from unlocking tacit knowledge could be used in combination with the tag dashboards in order to improve the general business strategy as well as HR-strategy. Of course, this needs to be implemented in order to test the validity of the approach better. Therefore, this topic is interesting material for future research.

The tag dashboards could also be used as an information retrieval tool instead of an explicit knowledge retrieval tool. By drilling-down one of the components of the tag dashboards, the employees could use the tag dashboards as a tool to retrieve information.

11.6 Prerequisites and recommendations

In Chapter 10, we provided an overview of a number of prerequisites we believe are important when using tags to unlock knowledge in the company. We derived these prerequisites through our case study research and for each of them we gave a number of recommendations. We visualized them in Figure 10.3.

11.7 Future research

In this section we discuss a number of topics we believe are important for future research. We discuss each of them in little more detail.

11.7.1 Tag dashboards

In Chapter 9 we presented the tag dashboard approach and tested the validity by doing 5 interviews with managers in the Company. This approach, however, should be extended and validated on a larger scale, but this is only possible when the approach is implemented in the Company or another company. It would also be interesting to measure in what extent the tag dashboard can influence the business strategy. In addition, it would valuable
11.7.2 Tag purpose

We measured the implicit tag quality of tags of the Company without taking the tag purpose into account. In the Company employees mainly use tags to retrieve information and to decide whether or not they will read a particular message as we discussed in Chapter 5. Therefore, we were not able to take the tag purpose into account. As future research, it would be interesting to test whether employees would select other tags when employees would be asked to annotate the resources for other purposes, for example to unlock knowledge stored within the corporate information.

11.7.3 Corporate ontology through tags improves the information retrieval problem

In Chapter 10 we suggested how a corporate ontology can be built through an analysis of corporate tags. The question, however, comes up how extended the ontology should be. For instance, in Chapter 2 we explained that two types of ontologies exist: lightweight and heavyweight ontologies. Of course, the latter type of ontologies are more complex and thus more time-consuming to build. As future research it would be interesting to test how extended the ontology should be in order to be an added value for the company. For instance whether it would help employees reduce the time to retrieve information or whether it can improve the process of unlocking knowledge that resides in the company. We assume that the more extended the ontology should be the more human input it will require and thus incur a higher total cost of ownership. It would be interesting to find the point where a small amount of input generates a large amount of profit.
11.8 Final conclusion

In this thesis we showed how an analysis of employees’ tags can be used to unlock the tacit as well as explicit knowledge that is in the organizational memory. We have clearly indicated what might be the pitfalls for a company when it decides to implement tagging in the company and we also formulated a number of recommendations. In addition, we explained how tag analysis can effect the business strategy as well as how it can be used to measure the objectives set by the board of directors and executives as well as any other manager that needs to measure his actions and objectives. We hope that this work will be interesting and valuable for companies.
Appendix A

Appendix chapters 6 and 7

A.1 Introduction

In this chapter, we discuss all the technical issues and preparatory steps which were needed for our research discussed in Chapters 6 and 7.

A.2 Methodology to verify tag rules in Chapter 6

We applied step 1, 2, 3 of the CorTagCleaning approach and then selected all the tags that could be retrieved in the Dutch dictionary list in order to lemmatize all these tags. This implies that we only took the tags into account that could be retrieved in the dictionary list to calculate the tag compliance.

A.3 Preparatory steps for interviews in Chapter 7

In the paragraphs below, we discuss all the preparatory steps we had to take before the interviews could take place.
A.3.1 Sample of messages

To be able to draw conclusions from the whole collection of messages, we needed a large sample or we had to focus on a specific group of messages. In statistics, the general law of “the more, the better” holds: the larger the sample, the more valuable the results will be, but the more resources it requires. Since we are in a corporate environment, resources are limited: every minute an employee has to spend on a survey implies a cost for the company. To reduce this problem, we focused on a specific group of messages created by the employees in the Company and took a sample of it.

We concentrated on messages of message type notes created by two functional groups\(^1\) of the Company: analysts that focus on making the working processes in the Company more efficient. A total of 17,199 messages corresponded to these requirements. Although it would have made more sense to take a sample of these messages and only interview the recipients of the messages, we decided not to do this. Due to organizational reasons in the company, it was not possible to interview the recipients of these messages. Since we wanted every message to be evaluated by six employees, we would have had to involve far more employees in the interviews to obtain a random sample and the interviewees would have had to judge different amounts of messages. For organizational reasons, the company preferred to let a smaller group of people evaluate the same number of messages.

To make the content of the interviewees more closely related to their working environment, we excluded the messages that did not have any recipient that was part of these two functional groups. In this way, we reduced the number of messages to 7,300. Then, we excluded the confidential messages with limited access\(^2\) since the company asked us not to include these messages in the sample. The set of messages scaled down to around 4,000. Then we randomly selected 200 messages from the set of 4,000 messages. To be certain that the messages were valuable for the interviews (i.e. not too short and with meaningful content), we asked our contact person at the company to evaluate the mes-

\(^1\)Description of functional groups: Chapter 5.
\(^2\)Description of dataset: Chapter 5.
sages. About 60 per cent of the messages were rejected because most of them were too short and irrelevant to be included in the interviews. As a result, we obtained a total of 95 messages that were suitable for the interviews and took a sample of 50 messages.

A.3.2 Participants

We took all the Dutch-speaking employees, 138, who were part of both functional groups and created manual note messages in 2006. We randomly selected 40 people as possible candidates for the interviews. Again, we asked our contact person to evaluate our list because in the meantime some employees might already have left the Company. Eventually we obtained a list of 26 people and asked our contact person at the Company to extend the list with an additional four employees. Later two employees were replaced because they had other obligations. One interviewee was ill at the time of the interview and it was decided not to replace him.\(^3\)

A.3.3 List of suggested tags (type B)

For each message, we created a list of suggested tags based on the technique which we explained in Section 7.4.2. In a last step, we removed all the tags originally chosen by the author of the message because a tag which was already chosen by the author of the message could not be in the suggested tags.

A.3.4 Test interview

It was important to test the questions before the actual interviews or questionnaire started (Hart et al. [1998]). The authors in Hart et al. [1998] explained that there are several ways to do this: answering the questions yourself, judging the questions by using a number of

\(^3\)A criticism of this approach could be that these people were not all randomly taken, but we had to deal with a number of organizational limitations, as explained previously.
checklists, asking for feedback from the colleagues or doing some test interviews. We
decided to do some kind of test interviews with our contact person at the *Company*. More
specifically, we wanted to know:

- How well the questions were formulated
- How much time it took for an employee to answer all the questions related to one
  message
- If it was difficult for an employee to judge a message when he or she was not a
  recipient of that message

First, we asked our contact person at the *Company* to evaluate our questions and to
do the test for two different kinds of messages of almost equal length: a message where
she was a recipient of the original message and one where she was not a recipient of the
message. In both cases, it took her less than four minutes to judge each of the messages.
She told us that, whether or not she was a recipient of a message, it did not make any
difference to her answering questions regarding that message. We repeated the test with
four other colleagues on one message and, on average, it took them five minutes to answer
all the questions and none of them were recipients of the message. They told us that it
was not difficult for them to do the test.

Based on the feedback we were able to reformulate some of our initial interview
questions.

### A.3.5 Post processing: tags and word cleansing

Some interviewees wrote compound words in one word and others split the word into
two words (e.g. *laptop* and *lap top*). For each message we made an alphabetical list
of all the tags given by the interviewees as an answer to the first question⁴, the original

---

⁴A *spontaneous reader tag*, as we will define it in the next Section.
tags provided by the author of the message\(^5\), and the list of suggestions we generated for each message\(^6\). Although we had already cleaned up the list of the suggestions before the interviews, we needed to tidy up the three kinds of tag categories after the interviews.

Each of these categories contained many lexical variations, for instance *order* and *orders*. Since some words were company-specific terminology we decided to clean up the words manually. Next, we also tried to reduce the number of synonyms. During the interviews, we received feedback from the interviewees on a number of synonyms (e.g. *systeemontwerper* and *SO*) and we tried to take as many synonyms\(^7\) as possible into account. However, we did not consider nouns derived from verbs as synonyms (for instance betalen\(^8\) and betaling\(^9\)). In some cases the meaning from these two formats can deviate a lot and we decided not to consider these terms as synonyms. We also tried to use a consistent format. For instance, every user had their own way of writing a date and therefore we replaced all the dates by a generic term called *date*.

In Table A.1, we give an overview of the results of cleaning up the tags.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>author tag</em></td>
<td>407</td>
<td>275</td>
<td>↓ 34.43%</td>
</tr>
<tr>
<td><em>spontaneous reader tag</em></td>
<td>1329</td>
<td>1174</td>
<td>↓ 11.66%</td>
</tr>
<tr>
<td><em>suggested reader tag</em></td>
<td>1316</td>
<td>1184</td>
<td>↓ 10.03%</td>
</tr>
</tbody>
</table>

Table A.1: Post processing

A.3.6 List of synonyms used to clean-up tags

1. Orpi → Orderpicker

2. Systeemontwerp → SO

3. MF → Mainframe

\(^5\) An author tag, as we will define it in the next Section.

\(^6\) A suggested reader tag, as we will define it in the next Section.

\(^7\) We provide a list of all the synonyms we used in section A.3.6.

\(^8\) In English: *to pay*.

\(^9\) In English: *payment*. 

239
4. DWH → Datawarehouse

5. CS → name of branch

6. Str → Stroppen

7. info → informatie

8. Pom → Pommeroeul

9. SW → Software

10. name of branch → COGO

11. Syso → SO

1. kosten → kost

2. vragen → vraag

A.3.7 List of c/k words used to clean-up tags

- ACCOORD → AKKOORD
- ACTIES → ACTIE
- AKKORDEREN → ACCORDEREN
- AKTIE → ACTIE
- AKTIES → ACTIE
- AKWARIUM → AQUARIUM
- BIOSKOOP → BIOSCOOP
- CATALYSATOR → KATALYSATOR
- CLASSEREN → KLASSEREN
CLASSIEC → KLASSIEK
CLASSIEK → KLASSIEK
COKET → KOKET
COLONIE → KOLONIE
COMIEK → KOMIEK
CONCURRENTEN → CONCURRENT
CONKURRENTIE → CONCURRENTIE
CONTACT → CONTACT
CONVOOI → KONVOOI
CORREKTIE → CORRECTIE
COSMOS → KOSMOS
CRITIEC → KRITIEK
CRITIEK → KRITIEK
CRITISCH → KRITISCH
DEKLASSEREN → DECLASSEREN
ELECTRICITEIT → ELEKTRICITEIT
ELECTRISCH → ELEKTRISCH
ELECTROCUTIE → ELEKTROCUTIE
ELECTROKUTIE → ELEKTROCUTIE
ELEKTROKUTIE → ELEKTROCUTIE
ELEKTRONIKA → ELEKTRONICA
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS

- ETICET → ETIKET
- FACTUREN → FACTUUR
- FAKTUREN → FACTUUR
- FAKTUUR → FACTUUR
- FYSIKA → FYSICA
- INSEKT → INSECT
- KATASTROFE → CATASTROFE
- KATEGORIE → CATEGORIE
- KATEGORISCH → CATEGORISCH
- KLASSIEC → KLASSIEK
- KOALITIE → COALITIE
- KOCET → KOKET
- KODE → CODE
- KOLLEGE → COLLEGE
- KOMIEC → KOMIEK
- KOMIEK → KOMIEK
- KONCLUSIE → CONCLUSIE
- KONCURRENT → CONCURRENT
- KONCURRENTEN → CONCURRENT
- KONCURRENTIE → CONCURRENTIE
- KONKLUSIE → CONCLUSIE
• KONKURRENT → CONCURRENT
• KONKURRENTEN → CONCURRENT
• KONKURRENTIE → CONCURRENTIE
• KONTACT → CONTACT
• KONTAKT → CONTACT
• KORRECTIE → CORRECTIE
• KORREKTIE → CORRECTIE
• KRITIKUS → CRITICUS
• KRYPTISCH → CRYPTISCH
• KWASI → QUASI
• LOKATIE → LOCATIE
• LOKOMOTIEF → LOCOMOTIEF
• MIKROFOON → MICROFOON
• OCTOBER → OKTOBER
• OKTOPUS → OCTOPUS
• PRAKTIJK → PRAKTIJK
• PRODUCTEN → PRODUCT
• PRODUKT → PRODUCT
• PRODUKTEN → PRODUCT
• PROJECTEN → PROJECT
• PROJEKT → PROJECT
A.3.8 List of questions used in Chapter 6

An overview of the questions asked in the questionnaire

1. How do you choose tags?

2. Do you think it is difficult to choose tags? Please explain your answer

3. Do you think that suggestions could help you choose tags?

4. Do you think suggestions could help to reduce the time to choose tags?

5. Are you influenced by seeing the tags of other people? (by reading the message and tags from your colleagues)

6. Do you easily retrieve the messages based on the tags?

7. Did the suggestions in the interviews help you?

8. How would you like to see the suggestions on your screen?
Appendix B

Appendix chapter 8

We provide an overview of the preparatory steps we had to take to prepare the interviews in Chapter 8. At the end of this Section, we provide a list of interview questions we used.

B.1 Preparatory steps to apply tag analysis techniques

Before we could apply the proposed techniques on the dataset of the Company, we had to clean the tags. We used the tag cleaning approach we presented in Chapter 6 and applied it to tags that were used at least 20 times by the employees in this department. We reduced the tags from 114,323 to 112,761 unique tags.

To be able to compare the different techniques, we had to apply them on the same set of employees’ tags. As the TF-IPF technique required a cluster of employees that share a set of tags, we first built the clusters. Once we finished building the clusters, we chose a cluster and selected the employees that are part of this cluster. For each employee, we tested whether he had used a tag that belongs to the cluster at least 20 times; if this is the case, the employee is considered as a part of the cluster. There were 58 employees that belonged to this cluster. Then, it was possible to select a group of 10 employees that
could participate in the interviews. Because some employees might already have left the Company, we needed to ask our contact person at the Company to select 10 employees from this group.

B.2 Preparation social networks

Since we wanted to test whether colleagues who use the same tags also have the same knowledge field, we needed to build the social networks that arise around tags. As the edges in a graph or network represent the relation which exists between the nodes, we suggested counting the number of tags they share as the tie strength. We used the co-occurrence¹ to determine the tie strength. Instead of making pairs of tags for each unique document ID, we ordered the dataset by unique tags and made pairs of employees who share the same tag.

There are many visualization tools that allow us to visualize a social network. We decided to opt for NetDraw.² To use NetDraw, we need to create a VNA file that consists of three parts: node data, node properties, and tie data. As the VNA file requires a specific file format, we developed some software to create the VNA file.

Of course, employees share many tags and to increase the visibility of the network we used a threshold value of 320 based on trial and error. This implies that we only included employees in the social network who shared at least 320 tags. As a consequence not all the employees from the cluster were part of the social network.

Since we were interested in researching whether employees who share the same tags have the same field of knowledge, we created ego networks in NetDraw. An ego network is a network which visualizes all the nodes to which a central node is directly connected in a network and it also includes all the ties that exist between these other nodes (Everett and Borgatti [2005]). For each of the interviewees, except for two people, we visualized

¹Definition of co-occurrence: Glossary.
their ego network. Because of the threshold value of 320, two of the interviewees were not part of the social network and thus did not have an ego network.

B.3 Interview questions

We used several questions to structure the interview.

B.3.1 Part I: evaluation of the tag analysis techniques

• Which tag analysis technique reflects your expertise the best, given the fact that the results are obtained from a dataset of 2006?

• Do you think some important tags are missing? Which ones would you add?

• Would you change the order of the tags in the set that reflects your expertise the best?

• Are there tags that you want to remove from the list?

• Would you prefer to visualize these tags in a tag cloud?

B.3.2 Part II: social network

• Is the social network we obtained a reflection of the reality: do you agree with the employees that are in this network?

• Do you know these people? How do you know them?

• Do you think the visualization of the social network would be an added value to doing your job? Explain why.
Do you think the visualization of the social network would be an added value for new employees or employees who changed from another job in the company? Explain why.
Appendix C

Appendix chapter 9

C.1 Creation tag dashboards

To let these four components of the tag dashboards be automatically generated in one dashboard, we combined all bash shell scripts into one large script. The script performs MySQL queries, plots the data retrieved by the MySQL queries, and visualizes the co-occurrences with Graphviz. An example of such a dashboard is shown in Figure 9.3.

To collect the four components into one dashboard, we needed to transform each individual file (with file extension txt, gif, ps) into pdf files. To do this, we used different Unix tools. For instance, we used sam2p\(^1\) to convert a gif file to a post script file and ps2pdf\(^2\) to save it as a pdf file. In the end, we merged all pdf files into one pdf file while forcing them to be displayed all on one A4 page.

The creation of a dashboard where all the information is concentrated on one page was not so easy to make. For instance, we had some difficulties with Graphviz which does not use a bounding box to display its graphs in a PostScript file. A bounding box draws a rectangle around a PostScript image which fits the figure and text on this file perfectly.

\(^1\)http://www.inf.bme.hu/pts/sam2p/
\(^2\)http://www.ps2pdf.com/
It is needed to crop a PostScript file smoothly into a pdf file without having more white space than needed\(^3\). To solve this problem, we increased the size of the image by using the unix program \textit{psresize}\(^4\).

### C.2 Preparatory steps for interviews

As our research method, we opted for semi-structured interviews with a fixed list of questions. First, we had two informal conversations separately with our contact person at the Company as well as with the manager of the analysts to discuss the dashboard. Our contact person and the manager were both interested in the tag dashboards. However, they told us that the tag dashboards would be more interesting for employees with more responsibilities than for employees without any supervisory responsibilities. This is a quite logical observation because employees who do not have any management position mostly work on a smaller number of projects and are thus very well aware of all the issues going on in the project on which they are working. This corresponds with our general motivation for creating these kinds of tag dashboards: helping managers to achieve their strategy and objectives and at the same time to be able to monitor all the processes. Therefore, we decided to interview only employees who held management positions.

Of course, it was not easy for us to select the appropriate set of tags for each of the interviewees as we did not know which projects they were working on nor the strategy and objectives in 2006. Therefore, we skipped the first two steps of the approach presented in Figure 9.5 and asked our contact person at the Company to give a set of tags that described the projects on which the interviewees were working in 2006.

After we cleaned up all the tags based on the approach described in Chapter 5, we selected the messages and their metadata that contained one of these tags. We repeated this step for every interviewee. In this way, we gathered a collection of messages and tags

---

\(^3\)http://amath.colorado.edu/documentation/postscript/bb.html (retrieved 22nd of July 2009).

\(^4\)http://www.tardis.ed.ac.uk/ ajcd/psutils/psresize.html
that corresponded to the field of interest of every interviewee.

### C.2.1 Description of interviews

We invited five managers to participate in these interviews. Each of them had a certain responsibility in a functional group. Within the *Company* they are called *afdelingschef*\(^5\) but we will refer to them as managers.

Each interview took around 30 minutes. In the first part, every manager had to evaluate a few tag dashboards. To let this run more smoothly, we first showed them a dashboard and explained what kind of information we could get from the dashboard. We then asked them whether our interpretation was right and whether it was a good reflection of the reality. After being given an example, they were able to do the same thing with the dashboards that corresponded to the projects and themes which they were supervising.

In the second part, we asked them about the value of every component of the dashboards and whether they would consider implementing the tag dashboards in the *Company*.

At the end, we presented one interviewee, the manager who supervised the other managers participating in the interviews, with our framework for implementing the tag dashboards in the *Company* and asked them how it would fit into the working processes of the *Company*.

\(^5\)In English: *Head of Department.*
Appendix D

Appendix SPSS

Figure D.1: Spss output
### Syntax

REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT RHQT
/METHOD=ENTER length
/SCATTERPLOT=(*ZRESID, *ZPRED)
/RESIDUALS HIST(ZRESID) NORM(ZRESID)
/SAVE PRED ZPRED ZRESID.

### Resources

- **Processor Time:** 0:00:01.642
- **Elapsed Time:** 0:00:02.000
- **Memory Required:** 1548 bytes
- **Additional Memory Required for Residual Plots:** 912 bytes

### Variables Created or Modified

- PRE_3
- ZPR_3
- ZRE_3

### Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHQT</td>
<td>.02408</td>
<td>.015967</td>
<td>47</td>
</tr>
<tr>
<td>length</td>
<td>374.40</td>
<td>329.593</td>
<td>47</td>
</tr>
</tbody>
</table>

### Correlations

<table>
<thead>
<tr>
<th></th>
<th>RHQT</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>1.000</td>
<td>-.523</td>
</tr>
<tr>
<td>RHQT</td>
<td></td>
<td>-.523</td>
</tr>
<tr>
<td>length</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>
Correlations

<table>
<thead>
<tr>
<th></th>
<th>RHQT</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig. (1-tailed)</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>length</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>47</td>
<td>47</td>
</tr>
</tbody>
</table>

Variables Entered/Removed

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>length</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

- a. All requested variables entered.
- b. Dependent Variable: RHQT

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.523</td>
<td>.274</td>
<td>.258</td>
<td>.013757</td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), length
- b. Dependent Variable: RHQT

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>.003</td>
<td>1</td>
<td>.003</td>
<td>16.967</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>.009</td>
<td>45</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.012</td>
<td>46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), length
- b. Dependent Variable: RHQT
### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>.034</td>
<td>.003</td>
</tr>
<tr>
<td>length</td>
<td>-2.535E-5</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: RHQT

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>95% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>.027</td>
</tr>
<tr>
<td>length</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: RHQT

### Residuals Statistics

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Value</td>
<td>-.01105</td>
<td>.03164</td>
<td>.02408</td>
<td>.008355</td>
<td>47</td>
</tr>
<tr>
<td>Residual</td>
<td>-.017621</td>
<td>.044852</td>
<td>.000000</td>
<td>.013606</td>
<td>47</td>
</tr>
<tr>
<td>Std. Predicted Value</td>
<td>-4.204</td>
<td>.905</td>
<td>.000</td>
<td>1.000</td>
<td>47</td>
</tr>
<tr>
<td>Std. Residual</td>
<td>-1.281</td>
<td>3.260</td>
<td>.000</td>
<td>.989</td>
<td>47</td>
</tr>
</tbody>
</table>

a. Dependent Variable: RHQT

### Charts
Histogram

Dependent Variable: RHQT

Mean = -4.04E-16
Std. Dev. = 0.989
N = 47
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: RHQT
### One-Sample Statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggested_tags_good_quality</td>
<td>47</td>
<td>3.57</td>
<td>1.850</td>
<td>.270</td>
</tr>
</tbody>
</table>

### One-Sample Test

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
<th>Mean Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Suggested_tags_good_quality</td>
<td>2.128</td>
<td>46</td>
<td>.030</td>
<td>.574</td>
<td>.03</td>
</tr>
</tbody>
</table>

Figure D.2: Spss output
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>Tag quality dimension which checks whether tags are a reflection of the content of the resource. Spamming is not allowed, 80</td>
</tr>
<tr>
<td><strong>Broad folksonomy</strong></td>
<td>Folksonomy where each resource is annotated by several people, 59</td>
</tr>
<tr>
<td><strong>Co-occurrence</strong></td>
<td>Is a technique which is often suggested in literature on tagging. For each annotated resource combinations of tag pairs are made. The tie strength between a pair of tags is increased each time these two tags are used together., 90</td>
</tr>
<tr>
<td><strong>Completeness</strong></td>
<td>Tag quality dimension which checks whether tags describe all the aspects of the resources they annotate, 80</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>Tag quality dimension which checks whether a uniform format for tag use is used, 80</td>
</tr>
<tr>
<td><strong>Controlled vocabulary</strong></td>
<td>A list of predetermined terms that describe a specific domain, 48</td>
</tr>
<tr>
<td><strong>Explicit knowledge</strong></td>
<td>Knowledge which is available in any kind of tangible format, 30</td>
</tr>
</tbody>
</table>
Explicit tag quality measures how well tags describe the resource in terms of the tag quality dimension consistency, 81

Folksonomy is a type of categorization that consists of the aggregation of user-created keywords or tags used to describe a resource, 54

High-quality tag well cleaned-up tags that correspond to the three tag quality dimensions: accuracy, completeness and consistency, 82

Implicit tag quality measures how well tags describe the resource in terms of the tag quality dimension completeness and accuracy, 81

Lemmatization is a technique that transforms words their base or dictionary forms, for example it reduces the word prepared to prepare, 120

Levenshtein edit distance a text similarity metric which calculates the distance between two words. More specifically, it counts how many letters have to be replaced, deleted, or inserted to transform one word into the other. The higher the Levenshtein edit distance, the more different two words are., 120
Glossary

Narrow folksonomy
folksonomy where each resource is annotated by only one person, mostly the author of the resource, 59

Ontology
describes all the concepts, instances and relations from a specific domain mostly expressed in a formal format that is machine-interpretable, 39

Semantic web stack
The W3C has suggested building a Semantic Web Stack to make the semantic web a reality. The Web Stack consists of several technologies that can be used to create the Semantic Web, such as ontologies, RDF and XML, 47

Stemming
is a technique that transforms words into their stems or roots, for example it reduces the words preparation and prepared to prepar, 119

Tacit knowledge
knowledge which resides in individuals, 30

Tag quality dimensions
are introduced in this dissertation to describe the quality of tags. We distinguish three tag quality dimensions: accuracy, completeness and consistency, 80

Taxonomy
belongs to the group of subject-based classification. It puts all the terms in the controlled vocabulary into a hierarchy, 48

Thesaurus
It is an extension of a taxonomy where different relations are included such as equivalence, hierarchical and associative relationships., 50
Bibliography


G.A. Miller. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
UNLOCKING KNOWLEDGE THROUGH CORPORATE TAGS


I. Tuomi. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems*, 16(3), 1999.


R. White and T. Downs. How computers work. Que Corp. Indianapolis, IN, USA, 2008.


