# A Framework Concept for Profiling Researchers on Twitter using the Web of Data

Selver Softic[1], Martin Ebner[1], Laurens De Vocht[2], Erik Mannens[2] and Rik Van de Walle[2]
[1]*Department for Social Learning, Graz University of Technology, Graz, Austria*
[2]*Department of Electronics and Information Systems - Multimedia Lab, Ghent University - iMinds, Ghent, Belgium*

Keywords: Research 2.0, Science 2.0, Web 2.0, Semantic Web, Social Media, Linked Data, Profiling, Twitter, Microblogs, Web Mining.

Abstract: Based upon findings and results from our recent research (De Vocht et al., 2011) we propose a generic framework concept for researcher profiling with appliance to the areas of "Science 2.0" and "Research 2.0". Intensive growth of users in social networks, such as Twitter* generated a vast amount of information. It has been shown in many previous works that social networks users produce valuable content for profiling and recommendations (Reinhardt et al., 2009; Java et al., 2007; De Vocht et al., 2011). Our research focuses on identifying and locating experts for specific research area or topic. In our approach we apply semantic technologies like (RDF†, SPARQL‡), common vocabularies (SIOC§, FOAF¶, MOAT‖, Tag Ontology**) and Linked Data†† (GeoNames‡‡, COLINDA[a]) (Berners-Lee, 2006; Bizer et al., 2012) .

## 1 INTRODUCTION

Emergence of Social Web evolved many web content producing communities. However information generated in them still resides in isolated "data silos". The main reason for this is lack of standardized approaches for data interlinking. The Semantic Web Technology has well defined stack where appliance of common semantic vocabularies to model data such as SIOC (Semantically Interlinked Online Communities) (Breslin et al., 2005) and FOAF (Friend-Of-A-Friend) leads to generation of interlinked and semantically rich knowledge tanks (Bojars et al., 2008). This knowledge is built upon user profiles and the content they produce. Structuring e.g. microblog information offers potentials on qualitative mining of such data.

Methodology proposed in this paper relies on three main steps: The first step is called "triplification" or "RDFization" where data is extracted and annotated using vocabularies SIOC, FOAF, MOAT and Tag Ontology. The RDF triples as result of this process are stored and made accessible as linked graph instances. The final step includes the publication of the data various formats via SPARQL endpoint in order to provide a data for mining, which is state of the art practice in Semantic Web domain (Bizer et al., 2012; Tummarello et al., 2007; De Vocht et al., 2011). Hereby vocabularies modeling the domain context, used at the same time for structuring and description enable more profound insights on the nature of RDFized data.

Twitter as most known microblog produces 190 million Tweets and 1.6 billion search queries each day [1](2012). Further it is widely accepted in scientific community for communication e.g. at conferences or for discussion purposes (Boyd et al., 2010; Jansen et al., 2009; Zhao and Rosson, 2009; Ebner et al., 2011) what makes it reliable base for researcher profiling process. However Twitter API has some limitation which means that single user timeline includes only last 250 tweets. In order to consider those researchers who tend to tweet more often an alter-

*http://www.twitter.com
†http://www.w3.org/TR/rdf-concepts/
‡http://www.w3.org/TR/rdf-sparql-query/
§http://rdfs.org/sioc/spec/
¶http://www.foaf-project.org/docs/specs
‖http://moat-project.org/ontology
**http://www.holygoat.co.uk/projects/tags/
††http://linkeddata.org/
‡‡http://www.geonames.org/
[a]http://datahub.io/dataset/colinda

[1]http://thesocialskinny.com/100-social-media-statistics-for-2012/

Softic S., Ebner M., De Vocht L., Mannens E. and Van de Walle R..
A Framework Concept for Profiling Researchers on Twitter using the Web of Data.
DOI: 10.5220/0004369504470452
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 447-452
ISBN: 978-989-8565-54-9
Copyright © 2013 SCITEPRESS (Science and Technology Publications, Lda.)

native which includes also previous tweets must be provided.As possible solution to overcome the limits of Twitter API a tool called Grabeeter[2] (Mühlburger et al., 2010) has been implemented by Graz University of Technology. This application serves preservation of social data from Twitter. Grabeeter includes at this certain moment about 1700 profiles mostly from researchers and students and contains currently around 6 Million tweets.

This paper describes the concept architecture for the researcher profiling framework. It is aiming at gaining more knowledge and mining usable data out of the social context of microblogs for researcher profiling using findings from (Reinhardt et al., 2009; Java et al., 2007; Letierce et al., 2010; Boyd et al., 2010; Honeycutt and Herring, 2009; De Vocht et al., 2011).

## 2 RELATED WORK

The importance of microblogs, in recent years, is gaining on importance significantly every day (Zhao and Rosson, 2009). Most favored among them is Twitter, which induced a new culture of communication (McFedries, 2007; Java et al., 2007). Restricted 140 characters long Twitter messages are comparable with a short message internet-based services. Java et al. (Java et al., 2007) defined four main user behaviors why people are using Twitter: for daily chats, for conversation, for sharing information and for reporting news. Twitter is generating vast of tweets and search queries each day. According to recent reports [3](2012), Twitter has over 225 million users. Around 50 milliion of them use the Twitter each day. This makes the Twitter worth to be researched more into detail (Kwak et al., 2010). Usage of Twitter at conferences helps to increase information awareness around the event as well it supports spontaneous conversation between the conference participants, which can be used for networking and experience exchange. Nowadays very often so called conference related Twitterstreams based upon a hashtag search reflect the ongoing occurrences within the actual event (Reinhardt et al., 2009). Twitter info-walls placed at the conference location also support the conference administration, communication and discussion between the scientific tracks and sessions(Boyd et al., 2010; Jansen et al., 2009; Zhao and Rosson, 2009; Ebner et al., 2011). Applied in this manner, microblogging becomes a valuable reporting and exchange service. This finding is also confirmed various different publications before (Reinhardt et al., 2009; Ebner et al., 2010).

Communicational patterns in microblogs are easily mappable into a tripartite structure (De Vocht et al., 2011; Mika, 2005). Tripartite relations to data corresponds to the basic idea of RDF Framework and graph based data relation. Regarding Twitter, recently there have been some efforts like Semantic Tweet[4] to bring the data about Twitter users into a wise semantic form. In current research efforts for mapping of relations between the users, widely used FOAF (FrienOf-A-Friend) vocabulary is recommended to be used, and it will be considered by our architecture paradigm. For posts description and relations around microblogs like topic, author, content Semantic Web community offers a vocabulary called SIOC (Semantically Interlinked Online Communities) (Bojars et al., 2008; Breslin et al., 2005) along with Dublin Core[5]. Dealing with tags established MOAT(Passant, 2008) and Tag Ontology as good ontologies in this realm (Softic et al., 2009). Currently there are also some scientific projects that adress the issue of semantic microblogging platforms. Most remarkable of them is named (Semantic MicrOBlogging) or recently also known as SMOB2 (Passant et al., 2010; Passant et al., 2008). It provides a SPARQL API and relies on vocabularies like FOAF, SIOC, MOAT and OPO (Online Presence Ontology)[6]. Additionally it offers interfaces to the semantic search engines like Sindice[7] and to the Linked Data Cloud (LOD). Twitter based User Modeling Service (TUMS)[8] infers semantic user profiles from the tweet messages. This platform provides a topic detection and entity extraction for tweets. Further it allows an enrichment by linking tweets to news articles related to the context of these tweets (Tao et al., 2011). According to the emerging trends, there are some proven domain vocabularies as FOAF, SIOC, DC (Dublin Core)[9], MOAT, Tag Ontology or semantic retrieval standard protocols like SPARQL provided by the Semantic Web Community, which can be used for semantic description and quering of semantically enriched microblog data from Twitter (Mendes et al., 2010; Softic et al., 2010). For description of conference data through label, description, start and end date and location SWRC (Sure et al., 2005) Ontology along with the GeoNames Ontology covers all needs for COLINDA as primary mining source.

[2]http://grabeeter.tugraz.at

[3]http://thesocialskinny.com/100-social-media-statistics-for-2012/

[4]http://semantictweet.com/

[5]http://dublincore.org/documents/dcmi-terms/

[6]http://online-presence.net/ontology.php

[7]http://sindice.com/

[8]http://wis.ewi.tudelft.nl/tums/

[9]http://dublincore.org/documents/dces/

Linked Data movement (Berners-Lee, 2006) turned meanwhile the LOD Cloud (Linking Open Data Cloud) (Bojars et al., 2008; Bizer et al., 2012) as result of it into a reliable data source of graph based data offering data sets like e.g. GeoNames The GeoNames is a semantic version of a location service. For identifying conferences, linked data set called COLINDA (COnference LInked DAta)[10] offers a appropriate SPARQL endpoint[11]. COLINDA is linked to GeoNames and contains information about conference name, label, description, location and time when this event happened. Similar data sets also exist about books, publications, science etc. Linking semantic sources using simple principles described in (Berners-Lee, 2006; Bizer et al., 2012) turns the web into a large database, not only available for human, but also to intelligent agents (Bojars et al., 2008). Bringing implicit knowledge from Twitter data into this infrastructure would enhance the LOD Cloud and offer solid information base for research on Research 2.0 and Science 2.0 issues.

SCIENCE AND TECHNOLOGY PUBLICATIONS

# 3 ARCHITECTURE

## 3.1 Use Case and Design Specification

The main use case for framework is illustrated by: ”the conference case” depicted in Figure 1 already presented in our previous work (De Vocht et al., 2011). The idea of this scenario resides on fact that researchers are interested in very specific topics and events and that most of them report about these events via blogs or tweets (Reinhardt et al., 2009; Ebner et al., 2010) what creates huge opportunities for profiling. Agile development suggests to work with use

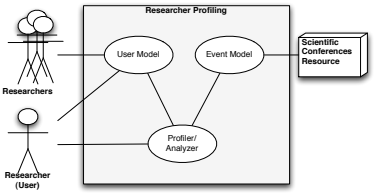


Figure 1: Use Case for ”Researcher Profiling” as proposed in (De Vocht et al., 2011).

cases. The framework has to support at least the ”Researcher Profiling” application that meets the requirements to the use case presented in figure 1. According to the current research work about semantic extraction frameworks (Softic et al., 2010) for data mining the conceptual design should include three basic layers: a data extraction layer, an interlinking layer and an analysis layer.

**Extraction Layer.** Extracts data from various several data sources and describes and relates them to a specific data context using the ontologies.

**Interlinking Layer.** Is feeded with annotated data (triples) and creates a SPARQL endpoint for it. It is responsible for requesting more data if needed for a certain information query.Futher it interpres and handles high level queries and translates them from/to SPARQL.

**Analysis Layer.** Deliver the results from interlinking layer adding some metrics to rank and evaluate the returned results.

## 3.2 Extraction

The extraction layer collects data from a profile from Twitter and the Grabeeter (Mühlburger et al., 2010) and maps then into two models: the ”User Microblogs Model” and the ”User ProfileModel”. The ”User Microblog Model” gathers all data from the tweets it gets from Grabeeter and describes them semantically using SIOC and Dublin Core vocabularies. The ”User Profile Model” is built upon Twitter user profile data with FOAF ontology. If a user not exists in Grabeeter, then the user profile and microblogs will be retrieved directly from Twitter. The data from Twitter is retrieved with the help of a Twitter API[12]. Finally, hashtags are identified by simple regular expressions and linked to the microblog data (text, creation date, author) and user profile( author data, social connections) using the Tag Ontology. These models serve a component named ”Triplifier”, that creates semantic instances of graph by assembling the data using relevant entities from ontologies. The result of the extraction is a collection of forms semantically annotated data into triples that describe the tweet wise content with time stamps and links to user profile. These triples are finally stored in a RDF Store. Figure 2 illustrates this extraction layer.

How the result of this module does look like at the end of the process can be seen in listing 1.

## 3.3 Interlinking

The interlinking layer accesses the stored triples created in the extraction layer via SPARQL protocol and tries to interlink them to COLINDA and GeoNames. It is impossible to create a generic framework that covers all data domains, but we can create a system that supports a broad range of use cases for a specific domain like e.g. research. For now we are focusing on

[10]http://www.colinda.org

[11]http://data.colinda.org/endpoint.php

[12]https://dev.twitter.com/docs/api

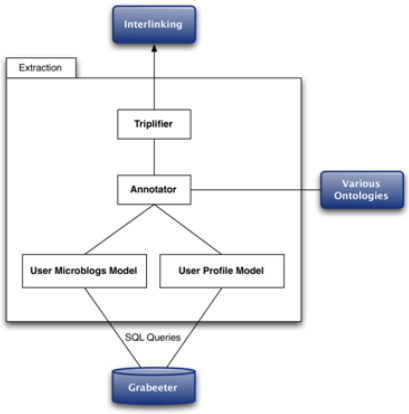


Figure 2: Extraction layer.

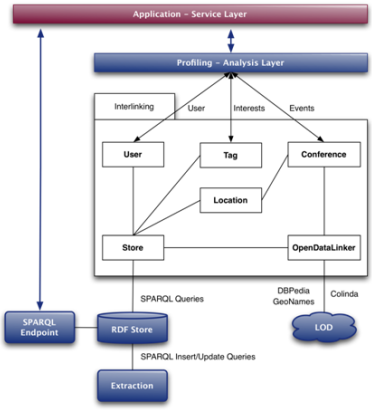


Figure 3: Interlinking layer.

Listing 1: Sample of RDFized data from microblog in N3 notation.

```
<http://twitter.com/someuser/status/21606926237>
  rdf:type sioct:MicroblogPost ;
  sioc:content
  "Great talk about #web #intelligence" @ #WEBIST2012 by @otheruser";
  sioc:has_creator <http://twitter.com/someuser/> ;
  foaf:maker <http://grabeteer.tugraz.at/foaf/someuser/> ;
  dcterms:created "2012-11-19" ;
  rdfs:sameAs <http://grabeeter.tugraz.at/tweet/199272> ;
  tag:tagName "WEBIST2012" ;
  tag:tagName "web" ;
  tag:tagName "intelligence" ;
  tag:taggedResource
  <http://twitter.com/someuser/status/21606926237> .

<http://twitter.com/someuser/>
  rdf:type foaf:Person ;
  foaf:name "Some User" ;
  foaf:depiction
  <http://a0.twimg.com/p_img/someuser.jpg> ;
  foaf:knows <http://twitter.com/friend_x_y/> ;
```

"Researcher Profiling", thus we distinguish two basic entities:

**User.** Social Microblogs, annotated data from twitter users (SIOC, FOAF, Dublin Core,Tag Ontology). Since we are doing profiling, data from the user is an absolute must.

**Domain.** Scientific Conferences, annotated data of scientific conferences (COLINDA) to enable the framework to recognize and link to conferences and scientific events.

Current state of Interlinking layer is depicted in Figure 3.

This module can also handle simple requests such as "give me tags for a user", "describe a user", "give me all friends of a user" same as Twitter API however the retrieved data is represented as RDF graph triples beside the state of the art formats like XML, CSV or JSON. Further it tries to identify which tags are scientific conferences. Finally the general knowledge is added and verified by linking tags or entities of conferences that occur in the result set from Linked Open Data. This process happens in the "OpenDataLinker" module. Adding additional Linked Data sets to this layer extends the appliance range of this framework. For now framework interprets following concepts that a researcher may be interested like: "persons", "topics", "events" and "locations". The interlinking layer translates queries concerning these concepts via the Linked Data sets, in our case COLINDA as shown in listing 2.

Listing 2: Retrieving conference location from COLINDA as proposed in (De Vocht et al., 2011).

```
SELECT *
{
  ?x rdfs:label "WEBIST2012";
     swrc:location ?loc.
  OPTIONAL
  {
  ?loc gn:name ?city;
       gn:countryName ?country;
       geo:lat ?lat;
       geo:long ?long.
  }
}
```

As result of interlinking process a hashtag that was extracted by triplification and tagged by Tag Ontology, if the matching with conference label has been detected, is attached to the COLINDA linked data set using the MOAT Ontology as presented in 3

Listing 3: Tagging recognised conference into RDF graph.

```
moat:tagMeaning <http://colinda.org/connference/123>;
tag:tagName "WEBIST2012" ;
tag:taggedResource <http://twitter.com/someuser/status/21606926237> .
```

In this way a link to the conference enables the resolution of all conference related aditional data, like description, date or conference location regarding the attached profile and microblog content.

## 3.4 Analysis and Result Delivery

The analysis is currently limited to a demonstration where two twitter users can be compared based on

similar hashtags they use. An evaluation rating is now simply the ”Cosine Similarity” between the two sets of hashtags. Further identified matched conferences with all additional data like topic, event and location are attached to result. The identification will just use the hashtag and see if it matches a conference name or abbreviation. Then a new metric giving weights to this conference matches will be calculated. A screen shot of a demo that uses the mining functionality of the profiling framework can be seen in Figure 4. The ranking is evaluated by simple count of corresponding entities. A similar function in framework is applicable for other topic, locations, links, mentions and friends.


Figure 4: Demo matching.

## 4 CONCLUSIONS AND FUTURE WORK

Approach presented in this paper, aims at mining usable information out of social microblogs, with a framework driven methodology. It is based upon Semantic Web standards and Linked Data. Introducing the interesting aspects about microblogs, authors tried to answer how far they this data can be used for other research areas like Science 2.0, Research 2.0. The authors also outlined the importance and relevance of such or similar efforts by examples and arguments from current research and with example of recent own work. In the near future, we want to answer questions like: Which researcher fit to me? Which conferences that cetrain researcher visited recently and they want to visit in the future fit to my interest area? Generation of forecast reports about scientist and specific conferences as well about upcoming conferences that match the own research focus is also an issue that will be supported in the future with proposed framework. Further linking the scientists automatically to sub communities based on their interests can be a thinkable extension of proposed context. Hereby we are aiminig at using common techiques for community distinction provided by network science like hierarchical clustering or minimal cut ratio methods. Considering the technical improvements we want to expose our proof of concept implementation as REST based API as done in previous work for interface aggregation (De Vocht et al., 2011) and run some more accurate tests on retrieval metrics like precision,recall and F-measure in order to evauate the quality of proposed solution.

## ACKNOWLEDGEMENTS

The research activities that have been described in this paper were funded by Graz University of Technology, Ghent University, iMinds (an independent research institute founded by the Flemish government to stimulate ICT innovation), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

## REFERENCES

Berners-Lee, T. (2006). Linked data - design issues.

Bizer, C., Richard, C., and Tom, H. (2012). How to publish linked data on the web.

Bojars, U., Breslin, J. G., Finn, A., and Decker, S. (2008). Using the semantic web for linking and reusing data across web 2.0 communities. *Web Semantics*, 6(1):21–28.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS ’10, pages 1–10, Washington, DC, USA. IEEE Computer Society.

Breslin, J. G., Harth, A., Bojars, U., and Decker, S. (2005). Towards semantically-interlinked online communities. In Gomez-Perez, A. and Euzenat, J., editors, *European Semantic Web Conference (ESWC)*, volume 3532 of *Lecture Notes on Computer Science*, pages 500–514. Springer.

De Vocht, L., Softic, S., Ebner, M., and Mühlburger, H. (2011). Semantically driven social data aggregation interfaces for research 2.0. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW ’11, pages 43:1–43:9, New York, NY, USA. ACM.

Ebner, M., Altmann, T., and Softic, S. (2011). @twitter analysis of #edmedia10– is the #informationstream usable for the #mass. *Form@re Open Journal*, 11(74).

Ebner, M., Mühlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., and Wheeler, S. (2010). Getting Granular on Twitter: Tweets from a Conference and

Their Limited Usefulness for Non-participants. In Reynolds, N. and Turcsányi-Szabó, M., editors, *Key Competencies in the Knowledge Society*, volume 324 of *IFIP Advances in Information and Communication Technology*, pages 102–113. Springer Boston.

Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. In *HICSS*, pages 1–10. IEEE Computer Society.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Letierce, J., Passant, A., Breslin, J., and Decker, S. (2010). Understanding how twitter is used to widely spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*.

McFedries, P. (2007). Technically speaking: All a-twitter. In *Spectrum, IEEE*, volume 44, pages 84–84.

Mendes, P. N., Passant, A., and Kapanipathi, P. (2010). Twarql: tapping into the wisdom of the crowd. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2010*.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V. R., and Musen, M. A., editors, *The Semantic Web - ISWC 2005, Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer.

Mühlburger, H., Ebner, M., and Taraghi, B. (2010). twitter try out# grabeeter to export, archive and search your tweets. *Research 2.0 approaches to TEL (2010)*.

Passant, A. (2008). Laublet p.: Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the Linked Data on the Web (LDOW2008) workshop at WWW2008*.

Passant, A., Bojars, U., Breslin, J. G., Hastrup, T., Stankovic, M., and Laublet, P. (2010). An overview of smob 2: Open, semantic and distributed microblogging. In Cohen, W. W. and Gosling, S., editors, *ICWSM*. The AAAI Press.

Passant, A., Hastrup, T., Bojars, U., and Breslin, J. (2008). Microblogging: A semantic web and distributed approach. In Bizer, C., Auer, S., Grimmes, G. A., and Heath, T., editors, *4th Workshop on Scripting for the Semantic Web co-located with ESWC2008*, Tenerife, Spain.

Reinhardt, W., Ebner, M., Beham, G., and Costa, C. (2009). How people are using twitter during conferences. In *Hornung-Pr ”ahauser, V., Luckmann, M.(Hg.): 5th EduMedia conference, Salzburg*, pages 145–156. Citeseer.

Softic, S., Ebner, M., Mühlburger, H., Altmann, T., and Taraghi, B. (2010). twitter mining# microblogs using# semantic technologies. *InProceedingsof6th Workshop on Semantic Web Applications and Perspectives*.

Softic, S., Taraghi, B., and Halb, W. (2009). Weaving social e-learning platforms into the web of linked data. *Proceedings of I-SEMANTICS*, pages 559–567.

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., and Oberle, D. (2005). The SWRC ontology - Semantic Web for research communities. In Bento, C., Cardoso, A., and Dias, G., editors, *Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005)*, pages 218–231.

Tao, K., Abel, F., Gao, Q., and Houben, G.-J. (2011). Tums: Twitter-based user modeling service. In Garcia-Castro, R., Fensel, D., and Antoniou, G., editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 269–283. Springer.

Tummarello, G., Oren, E., and Delbru, R. (2007). Sindice.com: Weaving the open linked data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, pages 547–560, Berlin, Heidelberg. Springer Verlag.

Zhao, D. and Rosson, M. B. (2009). How and why people twitter: the role that micro-blogging plays in informal communication at work. In Teasley, S. D., Havn, E. C., Prinz, W., and Lutters, W. G., editors, *GROUP*, pages 243–252. ACM.