No. 14 I May, 2012 DOSSIER

Academic research into Wikipedia

Eduard Aibar
Lecturer at the Arts and Humanities Department (UOC)
Researcher in the IN3’s Electronic Government and Democracy (GADE) research group
eaiabar@uoc.edu
http://www.uoc.edu/webs/eaiabar

Mayo Fuster
Postdoctoral fellow at the Berkman Center for Internet & Society (Harvard University)
and at the Institute of Government and Public Policies (UAB)
Mayo.Fuster@EUI.eu
http://www.onlinecreation.info/

CONTENTS

Presentation ................................................................. 55-58
Eduard Aibar and Mayo Fuster

Wikipedia’s Role in reputation management: An analysis of the best
and worst companies in the USA ......................................................... 59-67
Marcia W.Distaso, Marcus Messner

Using Wikipedia to develop language resources: WordNet 3.0 in Catalan
and Spanish .................................................................................. 68-76
Antoni Oliver, Salvador Climent

Panorama of the wikimediasphere ............................................... 77-87
David Gómez Fontanills

The Truth of Wikipedia ............................................................... 88-93
Nathaniel Tkacz

Wiki Loves Monuments 2011: the experience in Spain and reflections regarding
the diffusion of cultural heritage ...................................................... 94-104
Emilio José Rodríguez Posada, Ángel González Berdasco, Jorge A. Sierra Canduela,
Santiago Navarro Sanz, Tomás Saorín

RECOMMENDED CITATION:
ISSN 1575-2275
Presentation

Academic research into Wikipedia

Eduard Aibar
Lecturer at the Arts and Humanities Department (UOC)
Researcher in the IN3’s Electronic Government and Democracy (GADE) research group
eaiabar@uoc.edu
http://www.uoc.edu/webs/eaiabar

Mayo Fuster
Postdoctoral fellow at the Berkman Center for Internet & Society (Harvard University)
and at the Institute of Government and Public Policies (UAB)
Mayo.Fuster@EUI.eu
http://www.onlinecreation.info/

Published in: May 2012

Last year saw the celebration of the 10th anniversary of Wikipedia. In 2011, 10 years after it was created, Wikipedia became the sixth most visited website on the Internet and one of the most active online communities. At present, it contains nearly 20 million articles, 3.7 million of which can be found in the English version, which is the most extensive, although there are 282 versions of Wikipedia in different languages. It has some 365 million regular readers, around 90,000 regular editors — all voluntary — and hundreds of thousands of people who contribute anonymously without remuneration.

Wikipedia is one of the numerous examples of mass online collaborative projects to follow in the footsteps of free or open source software production and its modus operandi. Some authors see this new type of collaboration as representing an innovative form of social production, given that it operates on the edges of the market and the standard rules of business management, functions successfully without the intervention of hierarchical organisational structures or command management systems, and is developed thanks to the cooperation of thousands — or, in some cases, millions — of geographically dispersed people working voluntarily and without expecting any direct remuneration.

The term commons-based peer production was proposed recently to conceptualise the operation of these online collaborative systems (Benkler, 2006). Indeed, since about 2005, there has been growing interest from the scientific community, and in particular from the fields of social and human sciences, in researching this historically unprecedented phenomenon. A recent review of the scientific bibliography on Wikipedia has identified more than 2,100 scientific articles and 38 doctoral theses with Wikipedia or its sister projects as their object of analysis. The motivation and type of participants, form of organisation and governance, regulatory structure, publishing dynamics, content quality and reliability, teaching uses and role of technology, are some of the most recurrent themes in the scientific literature on Wikipedia (Okoli, 2009; Geert et al., 2011).

The aim of this issue of Digithum is to contribute to the incipient academic literature on Wikipedia. However, unlike most of this literature, which has mainly focused on the English version, this issue includes analyses of other Wikipedias and it also considers aspects little dealt with to date. Viewed as a whole, this issue provides a very clear idea of the wealth of theoretical and methodological approaches for studying Wikipedia and its sister projects.

This issue also provides an insight into the depth and multiplicity of the social and cultural impacts of Wikipedia: the focus of the texts presented range from its development and impact on the commercial world and on the public visibility of large corporations (in the article by Marcia W. DiStaso and Marcus Messner) to its role in the diffusion of culture and monumental heritage (in the article by Emilio José Rodríguez et al.). The article by Antoni Oliver and Salvador Climent looks at the use of Wikipedia as a corpus
of structured knowledge, within the framework of cutting-edge research into natural language processing. Likewise, the article by David Gómez, who proposes the concept of the wikimediaspHERE, shows us that Wikipedia actually forms part of an entire ecosystem full of projects, which, despite having elements in common, function with great autonomy as exceptional nodes in a more extensive network. Finally, the article by Nathaniel Tkacz analyses the practical and epistemological implications of one of the basic pillars of Wikipedia’s operating principles, the Neutral Point of View, and its relation to the specific concept of truth.

In summary, these five articles provide clear evidence of the versatility of Wikipedia as an object of study for contemporary academic research in many different fields of human and social sciences. Indirectly, these articles also highlight the need to develop new rigorous and systematic research into one of the most important open repositories of knowledge in history and, without doubt, one of the largest collective works of humanity.

References


RECOMMENDED CITATION


http://dx.doi.org/10.7238/d.v0i14.1473

ISSN 1575-2275

This work is subject to a Creative Commons Attribution 3.0 Spain licence. It may be copied, distributed, transmitted and adapted, provided you attribute it (authorship, journal name, publisher). The full text of the licence can be consulted here: http://creativecommons.org/licenses/by/3.0/es/deed.en.
Eduard Aibar

Lecturer at the Arts and Humanities Department (UOC)
Researcher in the IN3’s Electronic Government and Democracy (GADE) research group
eaiabar@uoc.edu
http://www.uoc.edu/webs/eaiabar

Eduard Aibar is a lecturer in the Arts and Humanities Department of the Open University of Catalonia (UOC) and a researcher in the Electronic Administration and Democracy (GADE) research group attached to the UOC’s Internet Interdisciplinary Institute (IN3). He specializes in social studies of science and technology. He is a former lecturer in the Department of Logic, History and the Philosophy of Science at the University of Barcelona (UB) and a postdoctoral researcher at the Maastricht Economic Research Institute on Innovation and Technology (University of Maastricht, Netherlands) and in the Political Science Evaluation Group at the University of Salamanca. He is a former vice president of Research at the UOC, a former director of IN3 and a former director of a doctoral programme on the Information and Knowledge Society. He is the co-author, with M.A. Quintanilla, of Cultura Tecnológica: Estudios sobre Ciencia, Tecnología y Sociedad (Barcelona: Horrori, 2002) and of numerous articles and book chapters in the science and technology studies field. He is currently director of a research project on university teacher perceptions, attitudes and practices with regard to Wikipedia.

Estudis d’Arts i Humanitats
Universitat Oberta de Catalunya
Av. Tibidabo 39-43
08035 Barcelona
Mayo Fuster
Postdoctoral fellow at the Berkman Center for Internet & Society (Harvard University) and at the Institute of Government and Public Policies (UAB)
Mayo.Fuster@EUI.eu
http://www.onlinecreation.info/

Mayo Fuster Morell specializes in network communities, commons-based peer production, social movements and public policy and network and participatory action research methods. She recently completed her PhD (Governance of online creation communities: Provision of infrastructure for the building of digital commons) at the European University Institute (Florence) under the supervision of Donatella della Porta. During her doctorate, Fuster Morell was a visiting researcher (under the auspices of Howard Rheingold) at the University of California Berkeley School of Information and was also an assistant in the Department of Communication at Stanford University. In 2011, she was a visiting researcher at the UOC’s Internet Interdisciplinary Institute (IN3). She is currently conducting research with Yochai Benkler at the Berkman Center for Internet and Society at Harvard University. She is a member of the Institute of Government and Public Policy at the Autonomous University of Barcelona (UAB), of the working group on Internet policy of the European Consortium for Political Research and of the Wikimedia Foundation research committee.

Berkman Centre for Internet and Society
23 Everett Street, 2nd Floor
Cambridge, MA 02138
Wikipedia’s Role in Reputation Management: An Analysis of the Best and Worst Companies in the USA

Marcia W. DiStaso
Assistant professor of Public Relations in the College of Communications at Pennsylvania State University
mwd10@psu.edu

Marcus Messner
Assistant professor of Journalism at the School of Mass Communications at Virginia Commonwealth University
mmessner@vcu.edu

Submission date: February, 2012
Accepted date: March, 2012
Published in: May, 2012

Abstract
Being considered one of the best companies in the USA is a great honor, but this reputation does not exempt businesses from negativity in the collaboratively edited online encyclopedia Wikipedia. Content analysis of corporate Wikipedia articles for companies with the best and worst reputations in the USA revealed that negative content outweighed positive content irrespective of reputation. It was found that both the best and the worst companies had more negative than positive content in Wikipedia. This is an important issue because Wikipedia is not only one of the most popular websites in the world, but is also often the first place people look when seeking corporate information. Although there was more content on corporate social responsibility in the entries for the ten companies with the best reputations, this was still overshadowed by content referring to legal issues or scandals. Ultimately, public relations professionals need to regularly monitor and request updates to their corporate Wikipedia articles regardless of what kind of company they work for.

Keywords
Wikipedia, reputation management, USA, monitor, social media

El paper de la Wikipedia en la gestió de la reputació: una anàlisi de les millors i les pitjors companyies dels Estats Units

Resum
Ser considerada una de les millors companyies dels Estats Units és un gran honor, però la bona reputació no eximeix aquestes empreses de trobar aspectes negatius a la Wikipedia, l’enciclopèdia en línia editada en col·laboració. Mitjançant una anàlisi de contingut de les companyies que tenen millor i pitjor reputació dels Estats Units, aquest estudi demostra que el contingut negatiu dels articles corporatius que es publiquen a la Wikipedia supera el contingut positiu independentment de la reputació de l’empresa. S’ha comprovat que tant les millors com les pitjors companyies contenen més continguts negatius que positius. És una qüestió important, ja que la Wikipedia no és només un dels llocs web més populars del món, sinó que s’ha convertit en un dels primers indrets per buscar...
Introduction

Social media has forever changed how corporations “manage” reputations. With social media and the collaboratively edited online encyclopedia Wikipedia, anyone can create and distribute content online, quickly and easily, and often for free. This connects companies and people in ways never possible before, blurring the lines that previously existed between companies, the media, and the public. The result is the public having a louder voice than ever before.

Social media are everywhere, and are increasingly being used by people in every country of the world; Wikipedia, for instance, is now available in 270 languages. No product or corporate reputation can escape the influence of social media, which need to be taken into account in every reputation management campaign. To be successful, a company’s social media activities must be coordinated with broader business goals and strategy, as well as with other communication efforts.

Reputation in the past was often considered as simple as top-of-mind awareness (Fombrun et al., 2003); today, however, the question is whether familiarity breeds trust or contempt (Centurion et al., 2009). To balance and land on the side of trust, a strategic approach to public relations is necessary (DiStaso et al., 2009), especially with today’s low levels of corporate approval and trust (Edelman, 2011). With reputations on the line, companies must include social media in their strategy.

Wikipedia is part of the growing list of social media that public relations professionals need to use and monitor (DiStaso et al., 2010; Paine, 2010). DiStaso et al. (2010) found that Wikipedia was one of the top sites listed when searching for companies using prominent search engines. This places Wikipedia in a position to potentially influence perceptions simply by being easily accessed. Plus, the depth and breadth of information available on Wikipedia far exceeds that of other encyclopedias, putting a plethora of information at the public’s fingertips (Crovitz, 2009).

Since having a good reputation provides companies with a strategic advantage, it makes sense that an aggregation of corporate information in Wikipedia may influence that advantage. Therefore, the purpose of this research is to see if there is a difference in Wikipedia content for companies with good and with bad reputations.

Literature Review

Wikipedia

Wikipedia was founded in 2001 and has quickly grown into one of the most popular websites in the world. According to Alexa.com (2012), it is currently the sixth most popular website in the world and in the USA, with most visitors viewing four pages on average and with visitors spending approximately five minutes on each visit.

Wikipedia, based on the wiki concept, is a free, collaboratively built encyclopedia. The idea that any Internet user can contribute to, or modify, any topic distinguishes Wikipedia from traditional encyclopedias written solely by experts.

Because of the social aspect of its creation, Wikipedia was once criticized, but it has gradually gained credibility. In 2009, the Financial Times pointed out that its editors work to make sure the content is true and, according to Stross (2009), “authority now comes not from a small group of encyclopedia editors and famous contributors, but from Google’s algorithms, which analyze links that point to web pages elsewhere, as well as other clues, to make an educated guess about trustworthiness” (p. B13). Reporters are also increasing their use of Wikipedia, since sources (Shaw, 2008) and news coverage were found to be framed as credible and accurate (Messner et al., 2011).

The growth of Wikipedia’s popularity and its increasing credibility have posed a challenge for public relations professionals (DiStaso et al., 2010); this is because anyone can now influence public opinion about a topic or company. A company’s reputation is no longer largely under its own control and this poses possibly serious threats. Because of this, companies are becoming more aware of the impact of Wikipedia on reputation (Wright et al., 2009).

Corporate reputation

Corporate reputation is “the overall estimation in which a particular company is held by its various constituents” (Fombrun, 1996, p. 37). A comparison of organizations is central to reputation, with one organization likely to have a better reputation than another (Deephouse et al., 2005; Ruef et al., 1998). Essentially it is feedback received by stakeholders.

Corporate reputation is a critical feature of organizations (Carter et al., 1998; Fombrun et al., 1990), but a favorable reputation does not come easily. Companies must work hard at
developing a reputation, which is often a driving force behind public relations efforts (Hutton et al., 2001).

Reputation is an important strategic asset that contributes to continued organizational success (Roberts et al., 2002). Having a favorable reputation has been argued to be one of the best ways to recruit and retain employees (Gatewood et al., 1993), charge premium prices (Milgrom et al., 1992), attract investors (Fombrun et al., 1990), maintain a competitive advantage (Roberts et al., 2002), and retain a reservoir of goodwill (Jones et al., 2000).

The visibility of a company in the minds of its stakeholders has been considered to be one of the strongest factors in determining a favorable reputation (Fombrun et al., 2003). Stakeholders include customers, suppliers, communities, employees, investors, the media, and any group that affects or can be affected by the company (Fombrun, 1996). Stakeholders construct corporate reputations from available information, such as direct communications from the company, the media, interpersonal communications (Fombrun et al., 1990), and the Internet (DiStaso et al., 2010).

Due to the numerous advantages of having a positive reputation, organizations are often concerned with the impact of their actions (Rawlins, 2009). There are many ways an organization can attempt to manage stakeholder perceptions, including press releases, advertising, press conferences, letters to shareholders, annual reports, and interviews in business publications (Fombrun, 1996), but social media and Wikipedia remove that level of control from the company and give high levels of influence to the public.

Research questions

Based on a thorough review of the literature, the research questions described immediately below were posed.

RQ1: Is there a difference in the tone of corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?

RQ1a: Is there a relationship between the overall tone in the corporate Wikipedia articles and reputation?

RQ2: Is there a difference in the amount of content for hot topics (corporate social responsibility and legal concerns/scandals) in corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?

RQ2a: Is there a relationship between the coverage of hot topics (corporate social responsibility and legal concerns/scandals) in corporate Wikipedia articles and reputation?

RQ3: Is there a difference in the use of reputation dimensions (social responsibility, emotional appeal, products and services, workplace environment, financial performance, and vision and leadership) in corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?

RQ3a: Is there a relationship between the amount of reputation sub-attributes covered in the corporate Wikipedia articles and reputation?

Method

This study builds on previous research by DiStaso et al. (2010), who analyzed Wikipedia entries for the top ten Fortune companies. While they identified Wikipedia content for the largest companies in the USA, they did not explore differences based on reputation. The goal of this particular study was to determine whether the formation of public opinion for companies with good and bad reputations differed. Content analysis was performed for 10 companies with the best reputations (Berkshire Hathaway, Johnson & Johnson, Google, 3M Company, SC Johnson, Intel Corporation, Microsoft, The Coca-Cola Company, Amazon.com, and General Mills) and 10 companies with the worst reputations (Delta Airlines, Bank of America, JP Morgan Chase, General Motors, Chrysler, Goldman Sachs, Citigroup, Fannie Mae, AIG, and Freddie Mac) as determined by the Harris Reputation Quotient 2010 (Harris, 2010).

The 2010 Harris Reputation Quotient consisted of 29,963 online interviews conducted between December 29, 2010 and February 15, 2011. It evaluated perceptions for 60 of the most visible companies in America across 20 attributes grouped into six reputation dimensions: products and services, financial performance, workplace environment, social responsibility, vision and leadership, and emotional appeal.

The complete sample included 20 articles from Wikipedia (ten companies with the best reputations and ten with the worst reputations). The unit of analysis was the sentence, so each of the 20 Wikipedia articles was analyzed sentence by sentence. The analysis was conducted for all the companies based on their Wikipedia articles as of April 9, 2011.

To increase intercoder and intracoder reliability, a codebook was created, pretested and revised before coding began. All coding was completed by two different coders and 20% was coded by both coders. The data reflected an intercoder reliability coefficient of .91 for tone, .98 for hot topics, and .90 for reputation topics using Scott’s pi (1955).

Framing

Each sentence was analyzed for tonality and topic to determine how the companies were framed. As in previous research, tonality was determined by coding each sentence as positive, negative, or neutral (eg., DiStaso et al., 2010; DiStaso et al., 2007; Michaelson et al., 2005). A sentence was coded as positive if it said something like: “In 1999, Time magazine named Bezos Person of the Year, recognizing the company’s success in popularizing online shopping.” An example of a negative sentence is as follows: “In late 2004, Fannie Mae was under investigation for its accounting practices.” And an example of a neutral sentence is as follows: “The corporation’s headquarters is located in New Brunswick, New Jersey, United States.” Following the methodology in Henry (2008), the overall tone was calculated by dividing the count of
positive sentences minus the count of negative sentences by the count of positive sentences plus the count of negative sentences. Therefore, the maximum value for overall tone was one, and the minimum value was minus one.

This study was concerned with two topics: corporate social responsibility (CSR) (i.e., sentences such as: “The corporation supplied all of its employees with cash incentives to buy hybrid vehicles, and began providing mortgage loan breaks for customers whose homes qualified as energy efficient.”) and legal concerns/scandals (i.e., sentences such as: “As of August 2005, when PBS's Frontline ran a story on the controversy, Coca-Cola strenuously denied all allegations of union-busting and murder of union leaders.”). Each sentence was coded for the presence or absence of either topic.

Reputation

Reputation was analyzed based on the Harris Reputation Quotient (Fombrun et al, 2000) consisting of six attributes with 20 sub-attributes (see Table 1). Each corporate Wikipedia article was analyzed for the presence or absence of each sub-attribute.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Sub-attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Responsibility</td>
<td>Supports good causes</td>
</tr>
<tr>
<td></td>
<td>Environmental responsibility</td>
</tr>
<tr>
<td></td>
<td>Community responsibility</td>
</tr>
<tr>
<td>Emotional Appeal</td>
<td>Feel good about</td>
</tr>
<tr>
<td></td>
<td>Admire and respect</td>
</tr>
<tr>
<td></td>
<td>Trust</td>
</tr>
<tr>
<td>Products and Services</td>
<td>High quality</td>
</tr>
<tr>
<td></td>
<td>Innovative</td>
</tr>
<tr>
<td></td>
<td>Value for money</td>
</tr>
<tr>
<td></td>
<td>Stands behind</td>
</tr>
<tr>
<td>Workplace Environment</td>
<td>Rewards employees fairly</td>
</tr>
<tr>
<td></td>
<td>Good place to work</td>
</tr>
<tr>
<td></td>
<td>Good employees</td>
</tr>
<tr>
<td>Financial Performance</td>
<td>Outperforms competitors</td>
</tr>
<tr>
<td></td>
<td>Profitability record</td>
</tr>
<tr>
<td></td>
<td>Low investment risk</td>
</tr>
<tr>
<td></td>
<td>Growth prospects</td>
</tr>
<tr>
<td>Vision and Leadership</td>
<td>Market opportunities</td>
</tr>
<tr>
<td></td>
<td>Excellent leadership</td>
</tr>
<tr>
<td></td>
<td>Clear vision for the future</td>
</tr>
</tbody>
</table>

Table 1: List of Corporate Reputation Attributes and Sub-attributes

Results

In analyzing the Wikipedia content for the ten US companies with the best reputations and the ten US companies with the worst reputations in 2010, a total of 3,069 sentences were analyzed (1,423 and 1,646 sentences for the companies with the best and worst reputations, respectively).

RQ1: Is there a difference in the tone of corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?

Both the best and the worst companies analyzed had more negative content than positive content. In the corporate Wikipedia articles, the best companies had a higher percentage of positive content (10.1%, n=144) compared to the worst companies (6.5%, n=108), and the best companies had a lower percentage of negative content (15.6%, n=222) compared to the worst companies (29.0%, n=476) ($\chi^2 (1, N=2) = 1.53, p<.05$) (Chart 1).

RQ1a: Is there a relationship between the overall tone in the corporate Wikipedia articles and reputation?

A Pearson correlation was calculated between the reputation quotient and the overall tone to find that there was a moderate statistically significant relationship between the reputation quotient and the percentage of CSR coverage ($r(20)=.558, p<.05$). In other words, as the amount of positive content improved so did reputation.

RQ2: Is there a difference in the amount of content for hot topics (corporate social responsibility and legal concerns/scandals) in corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?

The corporate Wikipedia articles were analyzed sentence by sentence to determine the total number of sentences that referred to the hot topics of CSR and legal concerns/scandals (Chart 2). The companies with the best reputations had more content on CSR (4.6%, n=65 for the best companies versus 2.2%, n=37 for the
worst companies). There was only a slight difference in the amount of content on legal concerns/scandals (13.3%, n=189 for the best companies versus 16.2%, n=266 for the worst companies) ($\chi^2 (1, N=10) = 13.49, p<.05$).

**RQ2a: Is there a relationship between the coverage of hot topics (corporate social responsibility and legal concerns/scandals) in corporate Wikipedia articles and reputation?**

Pearson correlations were calculated between the reputation quotient and the percentage of content on CSR and legal concerns/scandals. There was a moderate statistically significant relationship between the percentage of CSR coverage and the reputation quotient ($r(20) = .465, p<.05$). In other words, as the amount of CSR coverage increased so did a company’s reputation. There was negative non-statistically significant relationship between content on legal concerns/scandals and reputation.

**RQ3: Is there a difference in the use of reputation dimensions (social responsibility, emotional appeal, products and services, workplace environment, financial performance, and vision and leadership) in corporate Wikipedia articles for companies with the best reputations compared to companies with the worst reputations?**

The overall reputation score as determined by the Harris Reputation Quotient scores for the best companies was 16% higher for companies with the best reputations when compared to companies with the worst reputations (see Table 2). All of the reputation attributes were reflected in more content in the Wikipedia articles for companies with the best reputations than for those with the worst reputations, except for the social responsibility attribute. There were 6.7% more companies with the worst reputations with content on social responsibility. Specifically, content on community responsibility was 40% more common for companies with the worst reputations, and content on environmental responsibility was 20% more common for companies with the best reputations.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Sub-attributes</th>
<th>Best</th>
<th>Worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Responsibility</td>
<td>Supports good causes</td>
<td>70%</td>
<td>70%</td>
</tr>
<tr>
<td>Environmental responsibility</td>
<td>70%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Community responsibility</td>
<td>10%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50%</td>
<td>56.7%</td>
<td></td>
</tr>
<tr>
<td>Emotional Appeal</td>
<td>Feel good about</td>
<td>20%</td>
<td>0</td>
</tr>
<tr>
<td>Admire and respect</td>
<td>40%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Trust</td>
<td>10%</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23.3%</td>
<td>6.7%</td>
<td></td>
</tr>
<tr>
<td>Products and Services</td>
<td>High quality</td>
<td>60%</td>
<td>20%</td>
</tr>
<tr>
<td>Innovative</td>
<td>80%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Value for money</td>
<td>30%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Stands behind</td>
<td>60%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57.5%</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>Workplace Environment</td>
<td>Rewards employees fairly</td>
<td>30%</td>
<td>20%</td>
</tr>
<tr>
<td>Good place to work</td>
<td>50%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Good employees</td>
<td>50%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43.3%</td>
<td>26.7%</td>
<td></td>
</tr>
<tr>
<td>Financial Performance</td>
<td>Outperforms competitors</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td>Profitability record</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Low investment risk</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Growth prospects</td>
<td>50%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Vision and Leadership</td>
<td>Market opportunities</td>
<td>50%</td>
<td>10%</td>
</tr>
<tr>
<td>Excellent leadership</td>
<td>60%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Clear vision for the future</td>
<td>20%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>43.3%</td>
<td>16.7%</td>
<td></td>
</tr>
<tr>
<td>OVERALL REPUTATION</td>
<td>44.6%</td>
<td>28.6%</td>
<td></td>
</tr>
</tbody>
</table>

Products and services was the attribute with the greatest gap between the best and worst companies in Wikipedia articles (32.5% difference). The biggest differences between the two groups was content about the company standing behind its product or service (60% of the best companies had this kind of content compared to...
only 10% of the worst companies) and having content about high quality products or services (60% of the best companies had this kind of content compared to only 20% of the worst companies).

The next largest difference was for the vision and leadership attribute (26.6% difference). Content on market opportunities and excellent leadership had a 40% difference (respectively, 50% and 60% of the best companies had content compared to only 10% and 20% of the worst companies).

The companies also differed for the emotional appeal and workplace environment attributes (16.6% difference for both). Although content for emotional appeal was low overall, 20% of the best companies had content about making people feel good and 10% had content about trust, while the worst companies had no content on either topic. The biggest difference in the workplace environment attribute was with content regarding good employees (50% of the best companies had this kind of content compared to only 20% of the worst companies).

None of the companies in this analysis had content regarding low investment risk, and all of them had content regarding a record of profitability at some point in their history. Approximately 50% of the best companies talked about growth prospects compared to only 20% of the worst companies.

RQ3a: Is there a relationship between the amount of reputation sub-attributes covered in the corporate Wikipedia articles and reputation?

There was a strong statistically significant relationship between reputation and the number of reputation sub-attributes given coverage in the Wikipedia articles (r(20)=.627, p<.01). In other words, the more reputation content was included in Wikipedia, the better the reputation.

Discussion

We analyzed Wikipedia articles to see if there was a difference in content for 10 US based companies with the best reputations (Berkshire Hathaway, Johnson & Johnson, Google, 3M Company, SC Johnson, Intel Corporation, Microsoft, Coca-Cola Company, Amazon.com, and General Mills) compared to 10 US based companies with the worst reputations (Delta Airlines, Bank of America, JP Morgan Chase, General Motors, Chrysler, Goldman Sachs, Citigroup, Fannie Mae, AIG, and Freddie Mac).

Not only were the Wikipedia articles of the companies with the best reputations found to be slightly shorter than those for the companies with the worst reputations, they also contained slightly more neutral content. This could possibly be a reflection of having more to talk about for the worst companies, but that seems unlikely since there should be plenty to talk about for the best companies as well.

It was not surprising to find that the companies with the worst reputations had more negative content; they had, in fact, almost double the amount of negative content, although only slightly less positive content. Both types of companies had more negative than positive content. This indicates that even if a company is considered to have a good reputation, it is still very vulnerable to having its dirty laundry aired on Wikipedia. In other words, legal difficulties and scandals are popular Wikipedia topics, irrespective of the company's reputation. In fact, the amount of content on corporate social responsibility (CSR) efforts was about 10% less than the amount of content on legal issues or scandals. It is possible that the companies in this study had more negative aspects, but it seems unlikely that even the companies with the best reputations had more negative than positive aspects. Further research is needed to identify the CSR and legal issues/scandals that do or do not make it into corporate Wikipedia articles.

Looking at the reputation attributes and sub-attributes, the differences between the best companies and the worst companies were more obvious. Both types of companies had CSR content and, as mentioned previously, the best companies had more content dedicated to their efforts; in the end, both kinds of companies were fairly equal in the fact that they both had content that influences reputation. Where the companies truly differed was in content about products and services, vision and leadership, and emotional appeal. Taking into account the fact that all the companies in this study were considered to be among the most visible in the USA, not having content about how they stand behind their products and services or have high quality is surprising. A lack of excellent leadership in the companies with the worst reputations could be part of their problems, so not having content about this issue would be consistent. Emotional appeal is an attribute where both types of companies lacked content. It was rare for companies to have content about trust or feeling good, which only existed for the best companies. Admiration and respect was also low, especially for the companies with the worst reputations.

Overall, the attributes and sub-attributes that the public used to determine, through the Harris Interactive Reputation Quotient, the reputation of the companies in this study were also reflected in the Wikipedia articles. However, if a company’s reputation was to be judged by this content, the worst companies would still be the worst, but the best companies might be considered to have more of a moderate reputation since, on average, they only accounted for about 45% of the reputation sub-attributes. This begs the question: why don’t the best companies have more reputation building content in Wikipedia?

Public relations implications

Companies need to pay close attention to the content in their Wikipedia articles, because this social medium is often the first place people look when seeking information about companies. It is also important that public relations professionals move beyond monitoring. There are editing rules that must be followed, but this
does not mean that changes cannot be made (see the CREWE Wikipedia Engagement Flowchart on Wikimedia for guidance on making edits). The bottom line is that if information is incorrect, or if balance or positive news about a company needs to be added, requests for edits can be made with the proper references.

—

Conclusion

Through an analysis of the ten best companies and the ten worst companies in the USA as determined by the Harris Interactive Reputation Quotient, this study identified differences in Wikipedia content based on reputation. Although there were some differences, most were not great and would hardly indicate a clear dichotomy of best versus worst.

These findings highlight the importance of public relations professionals monitoring and requesting updates to Wikipedia articles about their companies. Being one of the best companies does not mean that a company can afford to ignore its Wikipedia content.

As with all research, this study had limitations. A key question that this study cannot answer – because it is beyond the scope of the research – is whether the companies in this study have different Wikipedia content because of their reputation or because they simply do/did good or bad things? Future research needs to explore this issue and study the influence of corporate Wikipedia articles on public opinion.

—

References

ALEXA (2012). “Wikipedia.org” [online article]. [Date of citation: 05/07/12].


This work is subject to a Creative Commons Attribution 3.0 Spain licence. It may be copied, distributed, transmitted and adapted, provided you attribute it (authorship, journal name, publisher). The full text of the licence can be consulted here: http://creativecommons.org/licenses/by/3.0/es/deed.en.
Marcia W. DiStaso
Assistant professor of public relations in the College of Communications at Penn State University
mwd10@psu.edu

Marcia W. DiStaso (PhD, University of Miami, 2007) is an assistant professor of communication in the Department of Advertising and Public Relations at Penn State University. Her research has been published in the Journal of Public Relations Research, Public Relations Review, Public Relations Journal, Journalism Studies, in over ten different books, and through the Institute for Public Relations. She is chair of the Public Relations Society of America (PRSA) Financial Communications section, co-chair of the PRSA National Research Committee, research chair of the PRSA Educator's Academy, member of the PRSA Speaker's Board, editorial board member for Public Relations Journal, and a board member for the International Public Relations Research Conference. Her research focuses on exploring and informing the practice of financial communication, investor relations and social media in public relations. She has been conducting research into Wikipedia since 2006.

College of Communications
Pennsylvania State University
200 Carnegie Building, University Park, PA 16802

Marcus Messner
Assistant professor of journalism at the School of Mass Communications at the Virginia Commonwealth University
mmessner@vcu.edu

Marcus Messner (PhD, University of Miami, 2009) is an assistant professor at the School of Mass Communications at the Virginia Commonwealth University. His research has been published in Journalism Studies, Journalism Practice, Newspaper Research Journal, Public Relations Journal as well as in The Handbook of Computer Networks, Comparative Studies on Media Consumption and World Youth, and New Media & Public Relations. He is a former head of the Communication Technology Division of the Association for Education in Journalism and Mass Communication and serves on the advisory board of the international Social Media Club Education Connection. His research interests are in new and social media and how they are adopted and used in journalism and public relations. He has been conducting research into Wikipedia since 2006.

School of Mass Communications
Virginia Commonwealth University
PO Box 842034, Richmond, VA 23284-2034
Abstract
We describe the state of the art in the use of Wikipedia for natural language processing tasks and also describe three applications of our own that enrich a powerful language resource: WordNet version 3.0 in Catalan and Spanish. Researchers have for many years sought applications that would take account of world knowledge in a more or less structured way, as this kind of knowledge has proven to be crucial to satisfactorily solving certain language processing tasks. Wikipedia may be the answer to the provision of this kind of information, as it is constantly updated and access is free.

Keywords
Wikipedia, WordNet, Natural Language Processing, linguistic resources

Ús de la Viquipèdia per al desenvolupament de recursos lingüístics: el WordNet 3.0 català i castellà

Resum
En aquest article presentem l’estat de la qüestió en l’ús de la Viquipèdia per a tasques relacionades amb el processament del llenguatge natural i tres aplicacions que hem creat per a l’enriquiment d’un recurs lingüístic de gran abast: el WordNet versió 3.0 per al català i castellà. Els investigadors en aquesta àrea fa anys que cerquen vies perquè les aplicacions integrin informació sobre coneixement del món, d’una manera més o menys estructurada, ja que aquest tipus de coneixement ha demostrat ser molt important per a resoldre de manera satisfactoria moltes tasques de processament del llenguatge. La Viquipèdia pot respondre perfectament a aquesta demanda d’informació amb l’avantatge del seu accés lliure i la seva actualització constant.

Paraules clau
Viquipèdia, WordNet, processament del llenguatge natural, recursos lingüístics
Introduction

A key goal of artificial intelligence (AI) is to equip machines with the ability to understand human language. Whereas language processing at the morphological and syntactic levels has reached a satisfactory stage in development, a major challenge is for computers to understand and represent meaning in order to be able to intelligently complete tasks.

Linguistic theory, following on from Langacker (1987), postulates that semantic processing cannot be considered to be linguistic but also encyclopedic, that is, word knowledge is inseparable from world knowledge. From this perspective, words are points of access to vast repositories of related knowledge. Many natural language processing (NLP) and AI researchers adopt the same starting point and seek methods and information sources that enable use of the knowledge collected and organized by humans.

In the last decade, Wikipedia (WP) has developed into a major semantic knowledge resource for NLP. WP is a vast, reliable and structured knowledge corpus that contains far more knowledge than any other manually created ontology or knowledge base, such as the extensive CYC (Lenat et al., 1990) or the widely used WordNet (Fellbaum, 1998). WP also has all the NLP utilities of the typical corpus, primarily in regard to how it makes use of statistical methods. In addition, the simple and well-structured XML format and an ongoing dumping policy (backups of packages of all of WP up to a specific cutoff point) both greatly simplify processing.

WP’s structure facilitates automatic and semiautomatic exploitation of its knowledge. WP is composed of articles organized as an article per concept and a concept per article. Titles are representative and introductions, relevant compact descriptions and infoboxes all reflect related knowledge structures. There are indirect links between equivalent terms, hyperlinks between terms and other articles, disambiguation pages for homonyms and synonyms and an underlying category structure by means of which each article is assigned to one or more categories. Wikipedia is also multilingual, with connected up structures for the different languages. WP articles may also have links to external URLs.

Hence, for the first time, NLP has available to it vast quantities of quality knowledge for exploitation. Below we review the current state of research in this area and then describe multilingual strategies of our own for extracting monolingual semantic knowledge.

1. The state of the art

Thomas and Amit (2007) have demonstrated how the quality of the information in WP articles can be inferred from the editing history: there is a clear correlation between stability in terms of production changes and WP quality, which is considered to be comparable to the Encyclopaedia Britannica (cf. Giles, 2005), just ten times larger (information referring to the year 2009).

Below we describe the main uses being made of WP for NLP research purposes, drawing especially on work reported in Medelyan et al. (2009), Gabrilovich et al. (2009) and Vivaldi et al. (2010).

1.1. Lexical disambiguation, semantic similarity and automatic document classification

Reasoning about the similarity between two words, phrases or documents is a routine task for humans but a challenge beyond the scope of machines. The most common way to establish the similarity between two documents is to statistically analyse the degree of coincidence between words, weighted according to the occurrence of these same words in text corpora; this approach is called latent semantic analysis (LSA) (Deerwester et al., 1990). Gabrilovich et al. (2007) also developed a technique for measuring the degree of association between the words in a particular document and WP entries, with the comparison producing an estimate of similarity that has proved more accurate than the LSA.

Calculating similarity is potentially useful for many tasks, such as detecting plagiarism and automatically classifying documents; the latter task, in turn, can be used to route news and emails and to identify spam.

A limiting factor in reasoning regarding text string similarity is disambiguating lexical meaning, that is, deciding whether two words that are identical in form have the same meaning. This is a key task in all applications for recovering, extracting and processing information. It is also the most difficult task, because language is inherently ambiguous: humans constantly disambiguate between possible meanings according to the context in which words occur. For NLP purposes, WordNet (Fellbaum, 1998) is the knowledge base most used as an inventory of words with meanings organized as related concepts. However, this knowledge base has drawbacks, namely, the excessive proliferation of meanings, bias towards the common lexicon and low content in proper names and specialized terms.

Authors such as Medelyan et al. (2008) use techniques such as those described above to calculate similarity between documents, comparing concepts (represented by WP entries) that are candidate meanings for the analysed word with the words that constitute the context for the sentence in which the word appears.

WP is currently the largest known repository of proper nouns and specialized vocabulary. Searches link words in texts with WP articles, thereby disambiguating and categorizing words; in a certain context, for example, “Chicago” is the band and not the city and not the Broadway show.

Another linguistic task necessary for all intelligent applications is knowing when different words in a text refer to the same concept, e.g., when “the King” is synonymous with “Juan Carlos
1.2. Retrieving information and obtaining answers

Retrieving information and obtaining answers from collections of documents is a task that is iteratively performed nowadays using popular search engines to pose questions and using the Internet as a collection of documents. WP is used to improve this approach in different ways.

First, search questions can be automatically expanded in WP using synonyms, alternative forms and related concepts so as to retrieve relevant documents. These operations had already been implemented in WordNet but are optimized with WP (eg., Gregorowicz et al., 2006); this is due to the sheer size and dynamism of WP which, compared with other static or slowly evolving resources, is constantly growing in size and in coverage regarding the emerging technologies.

Question answering (the automatic answering of a question posed in natural language) is an information retrieval approach that calls up not a collection of documents but a relevant sentence or fragment providing a specific answer. Authors like Kaisser (2008) include, in their usual Internet search procedure, structured WP information that helps establish a key factor: the type of entity sought as a response.

Wu et al. (2007), among others, have proposed, for the purpose of intelligent web searching, revitalizing the semantic web project (Berners-Lee et al., 2001), until now impractical given the magnitude of the task of semantically tagging the entire web. These authors propose semantically tagging the WP hyperlink structure itself to convert it into a machine-readable structure of knowledge and could constitute the core for a real semantic web of the future.

1.3. Expanding and improving semantic knowledge bases

The WP structure facilitates the maintenance and expansion of knowledge structures such as the thesauri and ontologies that are essential for many tasks and applications. Exploitation of hierarchical links and redirect links enables synonyms and inclusion relations between concepts to be established and also enables new concepts not covered by other ontologies to be explored. Infoboxes and the categories structure enable WP content to be converted directly into relational databases. Moreover, opening paragraphs in WP articles provide glosses (descriptions of concepts), interlingual links make it possible to create equivalent knowledge bases in other languages and comparison of the WP structure with that of another ontology permits expansion of both.

Information extraction, aimed at building structured knowledge bases from unstructured text, is optimized by exploitation of the encyclopedic nature of WP and its uniform writing style. A more ambitious goal is exploitation of the structure of WP links and categories in order to explicitly represent facts and rules and augment them with legacy systems that capture the overall semantic structure of language. This structure, which can be extracted from WP, is less costly and offers more coverage than manually developed knowledge bases such as WordNet and CYC. The three main approaches currently being investigated are extraction and tagging of WP categorical relations, extraction and organization of infobox information and enrichment of manually developed resources with information extracted from WP.

The main project in this regard is YAGO (Suchanek et al., 2007), a large taxonomy created by enriching concept structures in WordNet with categories and articles from WP. These authors are also enriching WordNet with semantic relations that were not previously in WordNet but were in WP, such as “born_in_year” and “born_in_place”. The authors claim to have created a knowledge base containing 20 million facts attributed to two million entities and equipped, moreover, with a highly expressive descriptive logic that allows inferences to be drawn from the facts. Another project along the same lines is DBpedia (Auer et al. 2007), which has transformed WP infoboxes into a giant set of logical propositions (103 million).

1.4. Educational applications

Sawaki et al. (2007) developed an answer search alternative that could potentially be applied to teaching: the extraction of questions, answers and clues from biographical WP articles with which to automatically generate puzzles of the kind “Who is this person?”. An interesting line of work in this regard is the ordering of clues according to difficulty so that the puzzle can be adapted accordingly.

Moré (2009) from the UOC exploits the hierarchical structure of WP categories to generate semantic concept fields (which can be multilingual) that can be used to prepare teaching and other academic materials. Moré et al. (2010) also use WP as an online teacher support resource via techniques that obtain answers based on students’ messages as questions and using WP and other sources as a collection of documents in which to find answers.
1.5. Multilingualism

Given its multilingual nature and the interlingual connection between entries, WP is both a multilingual dictionary and an aligned corpus. It can therefore be potentially used for machine translation purposes or for interlingual information retrieval (retrieval of information in a language other than that in which the query is launched) through the multilingual creation and expansion of questions.

Adafre et al. (2006) have developed another interesting use of WP. Working in English and Dutch, they compare machine translations of WP articles with the corresponding article written by a human and create bilingual lexicons that are used to identify pairs of sentences in different languages with the same or similar meaning. Erdmann et al. (2008) exploit WP's structure in a sophisticated way by going beyond direct interlingual links to create bilingual dictionaries.

Below we describe research of our own that exploits multilingualism in WP – currently underexplored – for the purpose of creating knowledge bases.

2. WordNet

As mentioned earlier, WordNet (Fellbaum, 1998) is a lexical database of English vocabulary, in which open-category words (nouns, verbs, adjectives and adverbs) are organized as sets of synonyms called “synsets”. Each synset represents a lexicalized concept that is connected with other synsets through semantic relations. The main WordNet relations are as follows: hyponymy (a type-of relation between a hyponym and a more generic hypernym), antonymy (a relation that implies a directly opposite meaning), meronymy (a part-of relation between a part and a whole) and troponymy (a relation between verbs that is similar to hyponymy in nouns).

For example, the WordNet 3.0 English synset, as identified by an offset and grammar category 02958343-n, has several variants: car, auto, automobile, machine and motorcar. Each synset is assigned a gloss or definition and examples of usage are frequently provided. For this synset, for example, the information provided is a motor vehicle with four wheels; usually propelled by an internal combustion engine along with the example he needs a car to get to work. This synset has 31 hyponyms, including 02701002-n (ambulance) and 03594945-n (jeep, landrover). It also has a hypernym, namely, 03791235-n (motor vehicle, automotive vehicle). An example meronym is 02685365-n (airbag).

To give an example of troponymy, we need to consider a verbal synset, for example, 01928579-v (sprint; run very fast, usually for a short distance) and 02055649-v (travel rapidly, speed, hurry, zip; move very fast).

Synonymy occurs between all the lexical forms (variants) of a specific synset; for example, car, auto, automobile, machine and motorcar are synonyms in that they are variants of the synset 02958343-n.

WordNet has become a standard resource for all kinds of NLP semantic research and applications. WordNet in English is free and can be downloaded from the Princeton University website (http://wordnet.princeton.edu). In the remainder of this article, we will refer to this Princeton WordNet as PWN. The current version is 3.0, released in December 2006. Table 1 shows the number of synsets available in versions 1.5, 1.6 and 3.0 of the PWN.

<table>
<thead>
<tr>
<th></th>
<th>Version 1.5</th>
<th>Version 1.6</th>
<th>Version 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>76,705</td>
<td>99,642</td>
<td>118,695</td>
</tr>
<tr>
<td>Nouns</td>
<td>51,253</td>
<td>66,025</td>
<td>83,073</td>
</tr>
<tr>
<td>Verbs</td>
<td>8,847</td>
<td>12,127</td>
<td>13,845</td>
</tr>
<tr>
<td>Adjectives</td>
<td>13,460</td>
<td>17,915</td>
<td>18,156</td>
</tr>
<tr>
<td>Adverbs</td>
<td>3,145</td>
<td>3,375</td>
<td>3,621</td>
</tr>
</tbody>
</table>

Table 1. Number of synsets for different versions of the Princeton University WordNet

As can be observed, the number of synsets has increased with each new version. Such updates also require word nets for other languages to be updated to ensure comparability.

Not all word nets are built and edited under a free license. Bond (2012), who lists existing word nets for different languages and the associated licences, indicates that the languages for which free word nets are available are English, Finnish, Russian, Thai, Danish, Japanese, Catalan, Irish Gaelic, Hindi, French, Malay, Indonesian, Spanish, Arabic and Hebrew.

2.1. Word net building strategies

Below we review some of the strategies that have been used to build word nets for different languages (excluding that for English, considered the original word net). Vossen (1998) distinguishes between two general approaches to building word nets:

- **Merge**: An ontology is first generated for the language and interlingual relations are subsequently generated between this word net and the PWN.
- **Expand**: The variants associated with the PWN synsets are translated using various strategies. In this case it is not necessary to establish interlingual relations because the word net and the PWN are parallel.
Each strategy has associated advantages and disadvantages (Vossen, 1996). The expand strategy is technically easier and ensures greater compatibility between word nets in different languages. However, such word nets are heavily influenced by the PWN and propagate all the latter’s errors and structural weaknesses. The merge strategy is more complex but allows for optimal use of available ontologies and thesauri.

2.2. The Catalan and Spanish word nets

The first word nets for Catalan (Benítez et al., 1998) and Spanish (Atserias et al., 1997) were based on applying an expand strategy to PWN 1.5 based on translating the variants corresponding to PWN synsets. This translation was performed using bilingual dictionaries, which described relations between English words and Catalan (or Spanish) words, although not between synsets and Catalan (or Spanish) words. To establish levels of confidence in assigning Catalan (or Spanish) variants to synsets: (i) relations between English words and synsets were assigned to monosemic and polysemic groups, and (ii) relations between English and Catalan (or Spanish) words were grouped according to the number of translations. This allowed synsets and Catalan (or Spanish) words to be directly related (thus obtaining variants in Catalan or Spanish) and a confidence level to be assigned to each relation. Catalan and Spanish word net versions 1.6 were generated by mapping synsets from version 1.5 to synsets in version 1.6.

The main difference between the Catalan and Spanish word nets is the licence used for distribution. The Catalan word net is distributed under a free license (GNU-GPL), whereas the Spanish word net is available under a proprietary license, although a small portion is distributed under a free licence along with the Freeling analyzer (the complete version is freely available for research).

3. Using Wikipedia to build word nets

Below we describe three WP-based applications that we developed to enrich Catalan and Spanish word nets version 3.0 with new variants or concepts: data exploitation in a knowledge base associated with WP called Babelnet; WP used as a bilingual dictionary; and WP used as a source of invariable proper names.

3.1. The Babelnet project

The aim of the Babelnet project (Navigli et al., 2010) is to create a large-scale semantic network that relates word-net lexical knowledge to WP encyclopaedic knowledge. In Figure 1 we can see how Babelnet relates synset 02958343-n for the variants motorcar, automobile and car with the English WP entry for car.

Based on this approach, whereby we have a relation between a synset \( s \) and an entry \( w_{\text{eng}} \) in the English WP, we can use interlingual links to establish the same relation for the equivalent entries in other languages (\( w_{\text{cat}} \), \( w_{\text{spa}} \), etc).

Figure 2 shows the interlingual links for the WP entry for car extracted from an XML dump. With this information we can relate the synset 02958343-n with the Catalan variant automòbil.

We can then find other variants for the original entry. WP contains a number of redirect pages that link up entries to a main entry to which they are equivalent. As can be seen in Figure 3, which shows the structure in XML of the Catalan entry for cotxe (car), this entry is redirected to automòbil. In other words, if
we look for cotxe, the information displayed will be the entry for automòbil. This redirect system allows us to establish, for example, that cotxe is a valid Catalan variant for the 02958343-n synset.

**Figure 3.** Information on the Catalan Wikipedia entry for cotxe

In practice, however, it is difficult to take full advantage of this feature as redirect pages have a broader use. For automòbil in Catalan, for example, we have the following synonyms: automoció, cotxe, cotxes, elements aerodinàmics en l’automòbil. This example indicates that the information cannot be automatically used, although it certainly could be useful but provided the proposals are subsequently manually reviewed.

In some cases, the language of interest may not have a corresponding entry in WP. To obtain variants in this case, Navigli et al. (2010) use Google Translate to translate a number of sentences in English containing variants from the synset s. These sentences are taken from two sources: SemCor and WP.

- **The SemCor corpus** (Miller et al., 1993). This corpus of English is semantically tagged (that is, most, if not all, nouns, verbs, adjectives and adverbs have a tag that indicates meaning). Tags for this corpus are the synsets of WordNet. Figure 4 shows an example sentence from this corpus with its tags (note that the real format of the corpus is different). The SemCor corpus can be downloaded from http://www.cse.unt.edu/~rada/downloads.html.

**Figure 4.** Example sentence from the SemCor corpus (format adapted for easier comprehension).

Then 00117620r he noticed 0215408v that the dry 02551380a wood 15098161n of the wheels 0457499n had swollen 00256507v.

- **WP sentences containing links to the page weng.** WP texts contain many hyperlinks to WP pages. Figure 5 shows a fragment of the Catalan entry for automòbil with hyperlinks marked in blue. These hyperlinks can also be considered as a kind of semantic tag, in that ambiguous words are linked to pages with the correct meaning. For example, the word bateries in Figure 5 could be considered ambiguous, but WP provides information about ambiguous words in a disambiguation page. Figure 6, for example, shows disambiguation information for bateria. The link for bateries in Figure 5 goes directly to the page giving information on electrical batteries and so confirms the meaning of this word in the text.

**Figure 5.** Fragment of the Catalan entry for automòbil with hyperlinks marked in blue.

L’automòbil (comunament cotxe o votura a la Catalunya Nord) és usualment un vehicle de quatre rodes destinat al transport de persones, amb capacitat entre dos i vuit seients. Es desplaça gràcies a un motor d’explosió a base d’una mescla de gasolina, gasoil i aire. En alguns països el combustible es fabrica a partir de determinades plantes en forma d’alcohol etílic. Recentment s’han començat a produir automòbils que funcionen amb motor elèctric, si bé l’autonomia d’aquests vehicles és encara limitada a causa del pes de les bateries. Les rodes davanteres dels automòbils poden moure’s cap a ambdós costats per a fer girs i prendre les corbes.

**Figure 6.** Disambiguation information for the Catalan word bateria.

**Bateria** té els significats següents.
- **Electricitat:** Bateria elèctrica, conjunt d’acumuladors connectats en sèrie o en paral·lel
- **Exèrcit:** Bateria (unitat militar), agrupació tàctica i de tir elemental composta per un conjunt d’artillers i de canons
- **Música:** Bateria (instrument musical), conjunt d’instruments de percussió que és tocat per un sol instrumentista
- **Eines:** Conjunt d’estris de cuina. Vegeu la Categoria: Estris de cuina

Once these sentences have been machine-translated, the most common translation is identified and is included as a variant corresponding to the language of interest.

3.2. Using Wikipedia as a bilingual dictionary

Encyclopaedic bilingual dictionaries can be generated from WP, that is, dictionaries containing both general language words of encyclopedic interest and also the names of people, geographic
locations, etc. Creating such dictionaries is simply a matter of following interlingual links. Just to provide readers with an idea of size, bilingual English-Catalan and English-Spanish dictionaries with 233,130 entries and 481,105 entries, respectively, can be generated from the English WP.

Such dictionaries can be used to directly assign Catalan and Spanish variants on the basis of English variants assigned to a specific synset. This assignment can only be determined with some degree of certainty in cases where a specific variant is assigned to a single synset, that is, when a word is monosemic. Table 2 summarizes monosemy and polysemy statistics for variants in WordNet 3.0 in English; note how the majority (82.32%) of the variants are, in fact, monosemic.

<table>
<thead>
<tr>
<th>Number of meanings</th>
<th>Variants</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>123,228</td>
<td>82.32</td>
</tr>
<tr>
<td>2</td>
<td>15,577</td>
<td>10.41</td>
</tr>
<tr>
<td>3</td>
<td>5,027</td>
<td>3.36</td>
</tr>
<tr>
<td>4</td>
<td>2,199</td>
<td>1.47</td>
</tr>
<tr>
<td>5 or more</td>
<td>3,659</td>
<td>2.44</td>
</tr>
<tr>
<td>Total variants</td>
<td>149,691</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2. Monosemy and polysemy for variants in WordNet 3.0 in English

Of these monosemic variants, a certain proportion have corresponding interlingual links (from which the bilingual dictionary is determined) and so can be directly assigned a Catalan or Spanish variant (which should, however, be checked manually). Using this method, we generated 3,719 and 5,260 variants for Catalan and for Spanish, respectively.

3.3. Detecting invariable proper names

WordNet contains many variants corresponding to proper names; WordNet 3.0 in English, for example, contains over 40,000 variants written with a capital letter. Many of these variants are names of people and geographic locations that are written exactly the same in English, Spanish and Catalan; they are not listed, however, as entries in the Catalan or Spanish WPs and so do not have interlingual links to these languages.

We can use interlingual links to other languages to check whether, as well as in English, the variant in question is written the same in, say, five languages. In our initial experiments using this strategy, performed only with multi-word units, 2,858 and 2,928 variants were obtained for Catalan and Spanish, respectively, with an accuracy rate of above 87%.

4. Conclusions

We have described the current situation regarding the use made of WP for NLP-related tasks and, more specifically, for the creation of language resources. In regard to language resources, we have described three ways in which we use WP to automatically generate lexical knowledge bases for Catalan and Spanish from WordNet 3.0, by now a standard resource in semantic processing. The approaches have the advantage that they can be applied to any language with an associated WP. Although they will not enable the creation of complete word nets, they can be usefully combined with other manual or automated techniques.

References


Using Wikipedia to develop language resources...

Antoni Oliver
Lecturer of the Department of Arts and Humanities
and director of the postgraduate course on Translation and Technologies (UOC)
aoliverg@uoc.edu

Antoni Oliver is lecturer of the Department of Arts and Humanities and director of the postgraduate course on Translation and Technologies at the Open University of Catalonia (UOC). His research interest is computational linguistics, especially in relation to machine translation.

Salvador Climent Roca
Lecturer of the Department of Arts and Humanities (UOC)
scliment@uoc.edu

Salvador Climent is lecturer of the Department of Arts and Humanities at the Open University of Catalonia (UOC). He coordinates the Catalan language and literature degree course and teaches general and cognitive linguistic subjects. He conducts research into computational linguistics and cognitive linguistics.
Abstract
The term wikimediasphere is proposed to refer to the group of WikiProjects, communities of editors, guidelines and organisations structured around the Wikimdia movement to generate free knowledge that is available to everyone. A description is made of the wikimediasphere, presenting the main projects and their characteristics, and its community, technological, regulatory, social and institutional dimensions are outlined. The wikimediasphere is placed in context and reference is made to its blurred boundaries. An explanation is provided of the role of the communities of editors of each project and their autonomy with respect to each other and to the Wikimedia Foundation. The author concludes by offering a panoramic view of the wikimediasphere.

Keywords
Wikimedia, Wikipedia, MediaWiki, free community good, communities of editors, editorial autonomy, free software
Introduction

The wikimediasphere is formed by wiki-based projects, principally promoted by the Wikimedia Foundation, in addition to the technologies that make these possible and communities of editors that work on and manage these projects.

Why wikimediasphere?

When mentioning this group, the media tend to refer simply to Wikipedia, as the academic literature often does; and following the community of editors, reference is also made to Wikipedia and its sister projects or the Wikimedia projects. If its social dimension is alluded to, reference may be made to the Wikimova movement or, more specifically if the focus is on a particular project, the community of editors. All these terms are metonymic (they mention the part to refer to the whole) and therefore they emphasise a particular dimension (the projects, the communities, the social movement).

The term wikimediasphere or wikimedia sphere can be useful in order to refer to some shared practices and some work and communication networks that are based on a group of interconnected wikis, complemented by other socio-technical devices. Theories of communication have used terms like logosphere (the universe of the spoken word), graphosphere (the universe of the written word) and videosphere or iconosphere (the universe of visual communication) to refer to the shared area of communication and meaning that is created around a mode of communication (Debray, 1994). In the field of the social use of digital networks, blogosphere has been used to refer to the panorama of blogs that are interconnected through links, back links and exchanged comments, which through these subtle interactions create a space for shared communication (Estalella, 2006). Wikisphere has also been used to refer to the equivalent concept in the field of wikis and various initiatives have been made to establish relations between them and ways to interconnect them, generally with less success than in the case of blogs.

With respect to the wikimediasphere, I propose using the name of the organisation (Wikimedia Foundation) and the social movement (Wikimedia movement) that drives and hosts it as the technical host, not the editor; the latter task is performed by the community of participants in each project (Paumier et al., 2010). The ideological glue that binds all the players involved is the aim of generating and supplying free content, an aim that is summarised in the vision statement of the Wikimedia Foundation, which in practice has become the motto of the entire movement:

"Imagine a world in which every single human being can freely share in the sum of all knowledge. That’s our commitment.”

This future vision materialises on the one hand in the mission to encourage people to participate in the creation of this content, providing them with the resources to do so, and on the other hand, in a set of values that are intended to guide this task (freedom, accessibility and quality, independence, a commitment...
to openness and diversity, transparency, a community focus. This vision also involves a set of founding principles of the Wikimedia projects as a community, which include the neutral point of view in the content, giving everyone the opportunity to edit, free licensing of content, the “wiki process” of decision-making (which means seeking consensus), an editorial environment that welcomes newcomers, and the creation of room and mechanisms for the resolution of conflicts.

---

### The communities of editors

In order to gain a better understanding of what the wikimediasphere is and how it is organised, it is important to point out that each of these projects includes a set of subprojects in different languages. With the exception of Wikimedia Commons and Wikispecies, which are multilingual wikis on a single installation, for each language there is a separate installation of the Mediawiki system, an autonomous project and a community that takes independent decisions about the rules of operation and the actions to be taken. Thus it is in the intersection between types of project (encyclopedia, dictionary, text books, etc.) and the language (English, German, Spanish, Catalan, French, Galician, etc.) where an autonomous community of wiki editors is formed. I would add that it is the language and not the country or territory that defines the project.

Any one can launch a version of these projects in a new language, first in the incubator wiki, intended for the first stages of development of a project, and then by getting a community together and transferring the project to an independent installation. In the case of Wikipedia, there are versions being edited for more than 270 languages.

In the wikimediasphere, the dynamics of the autonomous language communities that create their own rules, working methodologies and decision-taking mechanisms have a considerable influence on the specific implementation and operation of each project. As a result of this autonomy, each community adapts its processes to the volume of editors and to the aims it sets at any given time. However, this autonomy also means that each community has its own path of development and that there are variations in style, guidelines and policies.

Thus it would be possible to partially understand and study the wikimediasphere by looking at this intersection between project type and language used from different perspectives. On the one hand, it is possible to explore the language variations of the same project (the wikimediasphere, wikibooksphere, etc.), observing how they are interrelated and looking at the similarities and differences in operation and results of each community. On the other hand, it is also possible to review the different projects in the same language and, in a manner similar to that adopted with blogs – e.g. the Hispanic blogosphere (Cerezo, 2006) – talk about the wikimediasphere in English, Spanish, Catalan or any of the languages used.

---

### The wiki projects

Wikipedia is the flagship project of the wikimediasphere. It is a universal and multilingual encyclopedia created collaboratively by thousands of people around the world. The project was begun in January 2001 and it uses wiki technology to make it easy for anyone who wishes to publish material, create new content or improve what others have written before them.

In the years that followed, other projects were launched, all of them based on wiki and with the same open editing and community management approach:

- **Wiktionary**: a project begun in 2002. An encyclopedia is not a dictionary, and Wiktionary was created as a lexical companion to Wikipedia. It includes definitions, pronunciation, examples of use, etymology, synonyms and antonyms, and multilingual equivalents.
- **Wikibooks**: a project begun in 2003. It is a platform for the creation of text books for various areas of knowledge and various educational levels. In contrast to Wikipedia’s encyclopedia entries, the content can be structured into chapters and sections following a reading itinerary.
- **Wikiquote**: a project begun in 2003. It is a compendium of quotations, well-known phrases, popular sayings and set phrases, with attribution and contextual information.
- **Wikisource**: a project begun in 2003. It is a repository of public domain or freely-licensed texts. These include novels, poetry, plays, songs, manifestos, essays, ...
- **Wikimedia Commons**: a project begun in 2004. It is a repository of public domain or freely-licensed images and multimedia files (audio recordings, videos). The initial aim was to avoid duplicating files published on the various Wikipedias or wiki projects. The file is saved just once on Commons and it can be embedded into the pages of any of the projects. In terms of the number of editors and pageviews, it is the second most important project in the wikimediasphere after Wikipedia.
- **Wikinews**: a project begun in 2004. A space where news is published, usually drawing together news published in various media. The aim is to expand on the news and to offer a neutral point of view.
Besides the projects that have been mentioned, there are others that form part of the wikimediasphere and which are developed on a single multilingual MediaWiki installation, with English as the main language and translations of the content into other languages. Of these, we have already mentioned Wikimedia Commons, Wikispecies and Incubator. Another is MediaWiki.org, the wiki that serves as the official website for the software used by all the wikis.

The technical component

The MediaWiki.org wiki is a resource that provides information about the program and instructions for its installation, update, maintenance and administration. It is also the site where the extensions of the program developed by independent programmers, university teams, development companies or communities of programmers are documented. It is through MediaWiki that the wikimediasphere spreads and diffuses beyond the projects hosted by the Wikimedia Foundation; for the program, modulated for the needs and the working methods of its projects, in addition to the social relevance of these, has an influence on a large number of independent wiki projects that use it. At the same time, the administrators and users of the independent projects have their own needs and develop or demand characteristics for MediaWiki that often lead to one of the projects promoted by Wikimedia.

A project that complements MediaWiki.org is Translatewiki. This is not a project hosted on the servers of the Wikimedia Foundation, but an independent project maintained by two developers, Niklas Laxström and Siebrand Mazeland. It is a MediaWiki with a self-developed extension (Translate) that facilitates the translation and revision of software translations. Translatewiki is used for making collaborative translations of the interface and the system messages of the MediaWiki program, its extensions and many other free web-oriented software projects.

Another technological resource of note is Toolserver. And there are at least two reasons for its importance: because it hosts various technological “tools” that are used in the projects, and because it is an infrastructure provided by Wikimedia-Deutschland (the German chapter of the movement) with the assistance of the German chapter of the movement) with the support of other centres, rather than being provided directly by the foundation. Thus Toolserver shows the capacity and the autonomy of a local MediaWiki chapter to provide an infrastructure of key importance to the wikimediasphere as a whole, and it also offers the glimpse of a possibility that the latter is able to work while shifting part of the responsibility for the infrastructure on to organisations outside the foundation. Toolserver consists of a cluster of UNIX servers.

Of the tools that these host, some are experimental and under development, while others are stable tools that are used in the projects. Among the tools hosted are programs that provide statistical and analytical information about participation in the projects, “patrol” tools that assist maintenance, and bots used by the editors to perform repetitive tasks.

The technical aspect is one of the characteristic ingredients of the wikimediasphere. The most evident consideration is the fact that the same software is shared, but there are other aspects that contribute to its integration, such as the sharing of images (through Wikimedia Commons) or the possibility of inserting interwikis, a type of link shown as an internal link, but which connects to pages in other languages or other projects.

The legal and regulatory dimension

The legal and regulatory dimension of the wikimediasphere involves both its relation to the national laws and the editorial and operational rules that it creates itself through the communities. The production and implementation of rules is an important activity performed by the wikimediasphere. The first stages of the projects resulted in the drafting of some general guidelines, such as the 5 pillars of Wikipedia, which guide the activity of the editors. However, the production of content constantly means that the project communities have to deal with conflict, with situations arising in which a solution has to be defined to ensure editorial coherence, or a procedure needs to be created for the operation of these communities, and this leads to policies being established which, although they can be reviewed, will have to be followed by new editors that arrive at a later stage.

The free licenses used regulate the conditions in which the content is published. The Wikimedia projects represent one of the most significant examples of mass collaborative production of free content on a world scale. There are close links and relations with the organisations that promote the licenses, such as the Free Software Foundation or Creative Commons, and the use of the copyleft licenses for Wikipedia triggered the freedomdefined initiative to clearly establish what is understood by free cultural production (Chen, 2011). The incorporation of images and multimedia resources into Wikipedia, and especially Wikimedia Commons, represents constant interaction with the regulation of the copyright and intellectual property rights in various countries, detecting and standardising a range of issues related with the implementation of the licenses, the limits of the copyright, the scope of the public domain and the freedom of panorama.

Thus the wikimediasphere is self-regulated through the use of free licenses and rules created through internal processes of negotiation, consensus and decision-making. And, at the same time, it has a growing influence on social uses, rules and regulatory mechanisms that affect cultural production in various countries.
The social and institutional dimension

The wikimediasphere also has a dimension as an institution, association and social movement, which is developed, as mentioned earlier, around the Wikimedia Foundation. This is a non-profit organisation based in California (US), created in June 2003 by Jimmy Wales in order to assure the long-term future of the Wikipedia project, which he had begun in his company Bomis, and to provide a response to the demand made by some editors that it should not be a profit-making organisation that provided Wikipedia with its infrastructure (Chen, 2011). The governing body of the Foundation is the Board of Trustees, formed by 10 people, some of whom are chosen by mechanisms that include the participation of the community of editors. The Board is advised by the Advisory Board, an international network of experts. The Foundation has a relatively small, but growing number of staff (130 people as at April 2012) and its principal support comes from the sharing of tasks between thousands of volunteers and unpaid collaborators.

There are various grass-roots associations that support the development of the Wikimedia projects around the world, and the Wikimedia Foundation has agreements with these (39 recognised associations as at April 2012) as National and sub-national chapters or as Wikimedia Thematic Organizations. There are also organisations that sympathise with the free knowledge movement — Movement partners — with which stable collaborations are created, in addition to User groups that do not have to be formally established as associations.

The costs incurred by the Foundation consist of maintaining the technical infrastructure and paying the staff. Financing mainly comes from thousands of small contributions collected in annual campaigns, in addition to contributions made by individual and corporate benefactors through cash sums, servers and hosting. To date, the fundraising campaigns have succeeded in covering the costs of the Foundation’s staff and maintaining its infrastructure, while they have also provided financial assistance for trips and initiatives undertaken by individuals and organisations linked with the movement.

An important part of the organisational activity is also structured through wikis at a single multilingual installation, with English as the main language for each of these:

- Wikimedia Foundation: this serves as the official website of the Foundation. It is also a MediaWiki installation, but configured so that editing is limited to those persons authorised to do so. It is possible to comment on any content or contribute to it through a parallel web page created at Meta-Wiki.
- Meta-Wiki: the wiki dedicated to coordination and documentation for the global community of participants in the various projects in the wikimediasphere. On this wiki, different projects were developed that have subsequently ended up by having their own wiki, eg. MediaWiki.org, Wikimedia Strategy and Wikimedia Outreach.
- Wikimedia Strategy: this is used as a platform for discussion, deliberation and development of the five-year strategic plan for the Wikimedia movement. It was used for the first time for this purpose between July 2009 and July 2010.
- Wikimedia Outreach: this is a resource for outreach initiatives that provide information about the projects in the wikimediasphere. It is a collection of good practices, outreach materials and collaborative experiences with cultural and academic institutions.

The wikimediasphere is also structured on the Internet by means of other resources besides wikis. There is the official blog of the Foundation, in addition to the blogs of the various community members that are compiled on the aggregator Planet Wikimedia. Various IRC channels and specific mailing lists are also used. There are common channels and lists for different spheres, while each project community and each local chapter usually have their own.

The encyclopedia editors who work on Wikipedia call themselves Wikimedians. Wikimedians is a more generic term to refer to the participants in any of the Wikimedia projects. Wikipedia is not only the Foundation, because it also sees and projects itself as a social movement, the Wikimedia movement. For some time now, Wikimedians from each language version have been holding wiki meetings face to face to get to know each other personally and to coordinate work on the encyclopedia. They also organise workshops and other activities to promote the projects and to teach other people to edit. Since 2005, Wikimania has been the annual congress which brings together an important part of the Wikimedia community and serves as a platform for the presentation of initiatives and research relating to Wikipedia and its sister projects.

Blurred boundaries and context

Use of the term the wikimediasphere to refer to this entire framework of projects, communities and technology leads one to imagine it as being fairly structured and interconnected. But the wikimediasphere is not very clearly defined, falling within the context of society and the life of those who take part in it. In particular, the wikimediasphere forms part of a broader movement that works towards the “liberation” of technology, knowledge and cultural production, and which is in tension with other social forces that are working in an opposing or different direction. The operational characteristics of Wikipedia and other Wikimedia projects are largely the result of transferring to the field of knowledge what had been successfully tried out in the field of free software. The evolution of the wikimediasphere participates...
in and interacts with this movement, which uses free software, free and/or open licenses for content, and open, transparent and participative forms of organisation; it is also a movement that fosters social and technological autonomy. Together with the formation of communities of participants with the capacity for collective action, there is a process of forming non-profit organisations (Wikimedia Foundation, Mozilla Foundation, Free Software Foundation, Blender Foundation, LibreOffice Foundation, Creative Commons, P2P Foundation, etc.) that become players in the cultural production and social use of technologies. Some of these are as technologically and socially relevant as the corporate players such as Google, Yahoo and Microsoft; however, their mode of operation and impact is very different.

Editorial autonomy of the communities

From the outset, the versions of Wikipedia in different languages were created as independent editions and not as translations from one production language, such as English, to other languages. Thus each Wikipedia is written anew and it does not have to have the same articles or articles written in the same way. The articles can be translated (and the more complete languages with more articles are often translated into other languages), but this is a decision taken by each editor who chooses to contribute. In keeping with this independence of the versions, the communities of editors around each one are also free to take their own decisions. The definition of the role of the Wikimedia Foundation as an infrastructure supplier that does not intervene in editorial policy consolidates this position. The editorial policies on which each community can make decisions cover a very wide range of aspects: from the decision-making process itself, the way in which the content is organised (front pages, categories, navigation templates), and text format and style, to technical aspects and decisions on whether to include images. In practice, the communities have a mutual influence on each other in the way they do things, and when something works well in one community, it can serve as a reference for another. In particular, the Wikipedias that have a large number of articles and many participants usually have to find solutions to issues that are subsequently adopted or adapted by other communities.

An example of these variations can be seen in how the articles are distinguished in terms of quality. If three communities are compared (English, Spanish and Catalan), we may observe that in all three cases this process is begun with a request for peer review. This is optional, but it is generally considered to be necessary. Once this review has been completed, an active user (who has been registered for at least one month and has completed 100 editing assignments) can make the proposal which is discussed by the other active editors, who can declare themselves in favour, against or raise objections. In the English Wikipedia, the person who determines the length of this process and decides whether a consensus has been reached is the Featured Articles director (FA director) or someone delegated by this person.37 In the Spanish Wikipedia, this evaluation (in a discussion in which a minimum of 6 editors take part) is made by an Administrador de Candidaturas a Artículo Destacado (ACAD — Featured Article Candidature Administrator), a post awarded following a vote.38 And in the Catalan Wikipedia the proposal must receive a minimum of 8 votes in favour and no more than 2 objections; if these conditions have been met within a month, it is considered approved, but there is no specific post for making this evaluation.39 In all three cases, a distinction is drawn between featured and good articles. Thus we can observe similar processes with slight variations related with the type of community, its experience in the evaluation of articles, and its size.

Some recent examples also demonstrate the autonomy enjoyed by the communities when taking decisions, even when these have a considerable impact on society and in the media:

• “Strike” called by the Italian Wikipedia on 4 October 2011: The community of editors of Wikipedia in Italy decided to block access to the pages of the encyclopedia in protest against a law passed by the Italian parliament restricting freedom of expression on the Internet, which was a potential threat to Wikipedia.40
• “Strike” called by the English Wikipedia on 18 January 2012: The community of editors of the English Wikipedia decided (following a vote made by some 1,800 editors, the highest number to take part in a decision making process up until that time) to block access to the encyclopedia for 24 hours in protest against the SOPA and PIPA “anti-piracy laws” presented to the US Congress; in this way they took part in joint action with other websites and key Internet players. The Wikipedia communities in other languages and the communities of other projects (such as Commons, Wiktionary and Wikinews) discussed whether to back this initiative and 37 of them adopted some kind of measure in support; most of them posted an informative announcement on all the pages, but did not block access to articles. The Wikimedia Foundation offered and gave its technical support to the communities so that they could implement the decisions taken, which the foundation respected. 41 42 43
• Proposed “fork” of the German Wikipedia; autumn 2011: The discontent of the German Wikimedia community at the plans of the Wikimedia Foundation to promote a filter of “indecent”, “dangerous” or culturally “unacceptable” images led it to consider the possibility of creating a fork of the German Wikipedia in October and November 2011; it would do this by making a copy of it and operating it from servers outside the foundation.44 45
In practice, the Wikimedia Foundation plays a very important role, to the extent of championing this autonomy of the communities. In this respect, the following interventions were significant: the message of Sue Gardner, executive director of the foundation, defending the decision (and the decision-making process) of the Italian community; her message of support for the blackout in protest against SOPA and PIPA, accompanied by an explanation of the decision-making process; the press release from the foundation about this blackout; and the active role played by both the foundation in compiling the positions of the editors, and by the founder of Wikipedia, Jimmy Wales, in promoting the protest and explaining the reasons for it.

The actions taken by the foundation on some issues also generate tension related with the editorial autonomy of the communities. One example here is the initiative that caused the discontent of the German Wikimedians and of various Wikimedians in other communities: the proposal to create a filter so that users could make certain images invisible. The drive behind the proposal of the foundation has been seen as interfering with editorial autonomy. Likewise, the foundation’s “imposition” of a common policy on biographies of living people in April 2009 was seen as interference. As Shun-Ling Chen has pointed out (Chen, 2011), in the Wikimedia movement it is considered that the source of authority is in the communities and that the Wikimedia Foundation, at the appropriate moment, must implement the decisions taken by these communities; however, in practice the foundation has played an active role, both in internal processes and in external affairs with other social players. Due to a community-focused vocation, one of the values of the foundation, decision taking seeks at all times to open up participative processes and procedures to facilitate the search for consensus. The communities, however, urge the foundation not to over-extend its role.

Therefore, the communities of editors of each project and language version consider themselves as autonomous with respect to their editorial policies and they customarily take collective decisions about these. Even so, in some socially relevant aspects or projects that may have legal implications, the Wikimedia Foundation has defended the positions and the autonomy of the communities, but at the same time it has not refrained from promoting certain policies that may have a significant impact on the practices of the communities or even from imposing these in specific cases.

## Conclusions

The wikimediasphere is an interrelated unit of people, technological devices and self-generated regulations that operate mainly on the Internet. The people work on projects whose aim is the production of a common good related with knowledge and making this knowledge accessible to the whole of humanity. They form communities according to the type of project and language, with autonomy in their editorial policies, although they are heavily influenced by each other. The technological dimension is extremely relevant: the projects share the same software, MediaWiki, with variations in its configuration, while its main characteristic is the possibility to edit any content collaboratively. The wikimediasphere also has a social and institutional dimension. The Wikimedia Foundation is the principal supplier of the technological infrastructure and also the principal instrument for obtaining economic and organisational resources. Various grass-roots organisations disseminate and support the projects by organising activities, making them known in various social sectors and acting as interlocutors with other organisations and institutions. Together, the communities of editors and organisations form the Wikimedia movement, which perceives itself as a social movement in support of free knowledge. Despite its considerable activity and its strong interconnectedness, the wikimediasphere is not very strictly defined, and its influence and relations extend to projects and similar initiatives with which technologies or aims are shared.

If we look at the wikimediasphere as a whole, despite its blurred boundaries and its “fragmentation” or internal “modularity”, it becomes easier for us to understand how its various dimensions (content, values, community, social, institutional, regulatory, technological) are integrated and related.

## Notes

1. As with the term “logospher e”, here “sphere” (from the Greek ὁφαίρα, “sfaira”) refers to the “universe of” or “world of” as something that constitutes a coherent group.
5. Wikisphere (definition) – GetWiki http://getwiki.net/-WikiSphere
7. Wikindex http://wikitracer.com/Welcome
9. On Wikimedia Meta-Wiki the Wikimedia Movement is defined as the totality of people, activities and shared values which revolve around Wikimedia projects, including the Wikimedia Foundation, its chapters or sections, other similar-minded organisations and anyone who makes a
10. In 2005, in a text about Wikipedia, I referred to this as a community project that produces a common good, taking the definition of P. Kollock, according to whom a common good is a good from which anyone can benefit, independently of whether or not they have contributed to its production. See: GÓMEZ, D. (2005). “Wikipedia, un projecte comunitari en xarxa”. In: COL·LECTIU INVESTIGACCIÓ. Recerca Activista i moviments socials. Ed. El Viejo Topo. Barcelona http://www.enlloc.org/dvd/recerca/mmss/wikipedia_pcomu.pdf


11. Mayo Fuster has described Online Creation Communities (OCCs) as those in which people communicate, interact and cooperate by means of an Internet-based platform with the common goal of knowledge-making and sharing. The OCCs, among which she cites Wikipedia as an emblematic case, constitute a digital common good in themselves, as an integrated resource of information and knowledge, collectively owned and freely accessible to those who do not form part of the community. See: FUSTER, M. (2010) Participation in Online Creation Communities: Ecosystemic Participation? European University Institute.

12. Benjamin Keith Johnson considers Wikipedia to be a public good, insofar as it is an open community resource that requires voluntary contributions in order to be able to provide its content. See: JOHNSON, B.K. (2007). Wikipedia as collective action: personal incentives and enabling structures. P. 24. Thesis for Master of Arts, Michigan State University.

13. In reality, Wikipedia began to be published in 2001 under the GFDL copyleft license, and then in 2009 it was also published under the Creative Commons Attribution-ShareAlike license. Other Wikimedia projects also use these licenses, publish public domain content or, in the case of images and multimedia files, use a variety of free licenses.


15. The vision statement for the future of the Wikimedia Foundation is published on its website: http://wikimediafoundation.org/wiki/Vision

16. The mission statement of the Wikimedia Foundation is published on its website: http://wikimediafoundation.org/wiki/Mission

17. [11d] The values that guide the Wikimedia Foundation in its initiatives are published on its website: http://wikimediafoundation.org/wiki/Values

18. The Wikimedia projects as a community share certain principles that are published on Meta-Wiki: http://meta.wikimedia.org/wiki/Founding_principles

19. You can access any language version of Wikipedia from the following address: http://www.wikipedia.org


22. Toolserver is presented as “a service of Wikimedia Deutschland e.V. with assistance from the Wikimedia Foundation, Inc. and supported by Wikimedia UK, Wikimedia Switzerland, Wikimedia Austria and Wikimedia Sweden”. The access page to Toolserver is: http://toolserver.org

The page about Toolserver on Meta is: http://meta.wikimedia.org/wiki/Toolserver

The Toolserver wiki where information about the tools offered can be found is: https://wiki.toolserver.org

23. As the user Yavidaxiu remarked to me in a comment about the first version of this text in the Spanish Wikipedia Café, the communities of editors and the wiki technology are necessary so that the wikimediaphere may exist, but the existence of the Wikimedia Foundation is not essential; it is a useful instrument as an institutional space for the movement, but the wikimediaphere would also be possible without it. This said, it might then be different, since the institutional characteristics of the foundation also contribute to shaping it.

24. See the composition of the cluster that forms Toolserver at: https://wiki.toolserver.org/view/Servers

25. Any official Wikipedia policy must be based on the 5 pillars: Wikipedia is an encyclopedia, strive for the neutral point of view, the content is free, follow rules of etiquette, no rule is carved in stone. These can be found at: http://ca.wikipedia.org/wiki/Vi quip%C3%A9dia:Els_cinc_pilars

The first version of these, drawn up on the English Wikipedia in 2005, can be found at: http://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=13207659

26. Freedomdefined provides an agreed definition of what a “free cultural work” means. This is an independent definition of the license, and therefore it can cover various licenses and situations that achieve the “freedoms” enumerated. See the text at: http://freedomdefined.org


29. There is information about the Board of Trustees on this page: http://wikimediafoundation.org/wiki/Board_of_Trustees

30. There is information about the Advisory Board on this page: http://wikimediafoundation.org/wiki/Advisory_Board

31. Information obtained from http://wikimediafoundation.org/wiki/Frequently_asked_questions#How_is_the_Wikimedia_Foundation_run.3F In February 2011, when I wrote the first draft of this text, there were 57 employees. One year later, in February 2012, there were 104, almost double. As at April 2012, there were 130.

32. [19b] The information about the approved Wikimedia organisations and those that are being prepared can be found at: http://meta.wikimedia.org/wiki/Wikimedia_chapters

33. During the drafting and revision of this text, the Wikimedia movement is in the throes of an important debate about movement roles, with a range of proposals on the table (local chapters, partner organisations, associations, affiliates), and also about the role that these grass-roots organisations should play in fund raising and the distribution of funds. Finally, the various proposals have been synthesised in an agreement by the Board, whereby 5 models of affiliation are defined (Movement Partners, National or Sub-national Chapters, Thematic Organizations, User groups) without a hierarchical relationship between them: http://wikimediafoundation.org/wiki/Resolution:Recognizing_Models_of_Affiliations

34. At the 2012 Wikimedia Conference held in Berlin at the end of March, the Wikimedia chapters decided to initiate the constitution of a Chapters Council or Chapters Association to coordinate the Wikimedia organisations. This body, which will have its own legal status, has the potential to play an important institutional role in the future. See: http://meta.wikimedia.org/wiki/Wikimedia_Chapters_Association/Berlin_Agreement

35. The Wikimedia Foundation publishes reports on the sums of money raised and how they are used. From the most frequent questions section on its website it is possible to access the most recent reports: http://wikimediafoundation.org/wiki/Frequently_asked_questions#Where_can_I_find_more_financial_information.3F

36. See WikipediaOS (Detecting and validating the parallelisms between Wikipedia’s peer-production model and the way programmers and developers interact and work together in collaborative Open Source environments.) Enric Senabre (Esenabre) at Wikiversity: http://en.wikiversity.org/wiki/WikipediaOS


38. On the determination of the quality of articles in Spanish Wikipedia, see: http://es.wikipedia.org/wiki/Wikipedia:Candidatos_a_art%C3%A1culos_destacados

39. On the determination of the quality of articles in Catalan Wikipedia, you can view the content “The process for determining the quality of articles” on this page: http://ca.wikipedia.org/wiki/Viquip%C3%A8dia:Proposta_de_distinci%C3%B3_d%27alta_qualitat_d%27un_article

40. See the announcement made by the Italian Wikipedia community about the strike on 4 October 2011 at: http://it.wikipedia.org/wiki/Wikipedia:Comunicato_4_ottobre_2011 The page where the decision was taken: http://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Comma_29_e_Wikipedia And the page where an evaluation was made and reopening was discussed: http://it.wikipedia.org/wiki/Wikipedia:Bar/Discussioni/Sciopero_il_punto_della_situazione


42. Article on English Wikipedia about the protests against SOPA and PIPA. Section on the involvement of the Wikimedia community. http://en.wikipedia.org/wiki/Wikipedia:SOPA_initiative/Actions_by_other_communities

43. List of actions taken by the various Wikimedia project communities on the day of action against SOPA and PIPA on 18 January 2012: http://en.wikipedia.org/wiki/Wikipedia:SOPA_initiative/Actions_by_other_communities

44. Discussion on German Wikipedia about the possibility of a fork and its viability: http://de.wikipedia.org/wiki/Wikipedia:Fork


46. Message from Sue Gardner about the strike on 4 October
47. See: http://wikimediafoundation.org/wiki/English_Wikipedia_anti-SOPA_blackout
50. Why Wikipedia went down at midnight CNN, 17 January 2012
51. See the cited resolution about biographies of living people that is applied to all the language versions of Wikipedia: http://wikimediafoundation.org/wiki/Resolution:Biographies_of_living_people/ca
52. I would like to thank the user Montgomery, who in some observations about the first version of this text focused my attention on these tensions related with the autonomy of editorial policy.

RECOMMENDED CITATION
ISSN 1575-2275

This work is subject to a Creative Commons Attribution 3.0 Spain licence. It may be copied, distributed, transmitted and adapted, provided you attribute it (authorship, journal name, publisher). The full text of the licence can be consulted here: http://creativecommons.org/licenses/by/3.0/es/deed.en.
David Gómez Fontanills
Bachelor’s degree in Fine Arts (University of Barcelona)
Master’s degree in the Information and Knowledge Society (UOC)
dvd@enloc.org

Awarded a Bachelor’s Degree in Fine Arts by the UB (1991) and a Master in Information and Knowledge Society by the UOC (2006). He is currently working on his doctoral thesis, directed by Elisenda Ardèvol.

Educational collaborator and author of educational material for the Multimedia Degree course at the UOC (since 2000). Wiki consultancy as an independent professional (since 2007); he is currently technical and support manager of Viquilletra and Wiki EOI.

Member of the artistic and research group TAG Taller d’Intangibles (since 1996), with which he has taken part in several artistic initiatives and net.art projects (enloc.net). Member of the Friends of Catalan Wikipedia Association, Spanish Wikimedian and the Catalan Association of Visual Artists.

C/ E.Granados, 69
08330 Premià de Mar, El Maresme
Catalonia, Spain
The Truth of Wikipedia

Nathaniel Tkacz
Assistant Professor, The University of Warwick (UK)
nathanieltkacz@gmail.com

Submission date: April, 2012
Accepted date: April, 2012
Published in: May 2012

Resum
What does it mean to assert that Wikipedia has a relation to truth? That there is, despite regular claims to the contrary, an entire apparatus of truth in Wikipedia? In this article, I show that Wikipedia has in fact two distinct relations to truth: one which is well known and forms the basis of existing popular and scholarly commentaries, and another which refers to equally well-known aspects of Wikipedia, but has not been understood in terms of truth. I demonstrate Wikipedia's dual relation to truth through a close analysis of the Neutral Point of View core content policy (and one of the project's "Five Pillars"). I conclude by indicating what is at stake in the assertion that Wikipedia has a regime of truth and what bearing this has on existing commentaries.

Paraules clau
Wikipedia, Neutral Point of View, NPOV, Truth, Collaboration

La veritat de la Viquipèdia

Abstract
Què significa l’afirmació que la Viquipèdia té una relació amb la veritat? Que la Viquipèdia conté, malgrat que sovint es digui el contrari, tot un dispositiu de la veritat? En aquest article vull mostrar que, en realitat, la Viquipèdia té dues relacions diferents amb la veritat: una de ben coneguda i que constitueix la base dels comentaris populars i acadèmics, i una altra que es refereix a aspectes també ben coneguts de la Viquipèdia, però que no s’ha entès en termes de la veritat. Demostró aquí la doble relació de la Viquipèdia amb la veritat a través d’una detallada anàlisi d’una de les principals normes que regueixen a la Viquipèdia, el «punt de vista neutral» (que actualment és un dels «cinc pilars» del projecte). Finalitza l’article demostrant el que es posa en qüestió quan diem que la Viquipèdia té un «règim de la veritat» i com això repercuteix en els comentaris existents.

Keywords
viquipèdia, punt de vista neutral, PVN, veritat, col·laboració
What does it mean to assert that Wikipedia has a relation to truth? That there is, despite regular claims to the contrary, an entire apparatus of truth in Wikipedia? In this article, I show that Wikipedia has in fact two distinct relations to truth: one which is well known and forms the basis of existing popular and scholarly commentaries, and another which refers to equally well-known aspects of Wikipedia, but has not been understood in terms of truth. I demonstrate Wikipedia’s dual relation to truth through a close analysis of the Neutral Point of View core content policy (and one of the project’s “Five Pillars”). I conclude by indicating what is at stake in the assertion that Wikipedia has a regime of truth and what bearing this has on existing commentaries.

The Abandonment of Truth

Both supporters and detractors regularly suggest that truth is not central to Wikipedia. For Joseph Reagle, what he describes as the “abandonment of ‘truth’” (2010, p. 56) is central to the success of Wikipedia’s mode of collaboration and also what distinguishes it from its predecessors:

Historically, reference works have made few claims about neutrality as a stance of collaboration, or as an end result. While other reference works have had contributions from thousands of people, they were still controlled by a few persons of a relatively homogenous worldview. Indeed, a preoccupation of traditional references is their authoritative, quite different from Wikipedia’s abandonment of “truth”. (Reagle, 2010, p. 56)

Reagle notes how the authoritative stance of previous encyclopaedias is replaced in Wikipedia by the Neutral Point of View (NPOV). This policy, Reagle asserts, is what allows people who do not share “a relatively homogenous worldview” to collaborate on the same Wikipedia entry. The NPOV and related core content policies (Verifiability and No Original Research) are what seemingly replace the criterion of truth — the requirement that an entry or a particular statement is true — and what allows these people to get along. This is also how Jimmy Wales sees the function of NPOV, and Reagle makes use of the following passage to support his own claims:

The whole concept of the neutral point of view, as I originally envisioned it, was the idea of a social concept, for helping people get along: to avoid or sidestep a lot of philosophical debates. Someone who believes that truth is socially constructed, and somebody who believes that truth is a correspondence to the facts of reality, they can still work together. (Wales, cited in Reagle, 2010, p. 53)

These positive accounts of the replacement of truth with NPOV (and Verifiability) are met with more critical perspectives. The most common objection is simply that replacing more typical notions of truth with the NPOV as the threshold of inclusion for a statement means that things that are true might be excluded from the encyclopaedia if they do not also meet the NPOV and related policy requirements. Such objectives are usually manifested in cases involving recognised experts, such as the 2005 incident involving climate change expert William Connelly (see Schiff, 2006) and more recently, the incident involving Timothy Messer-Kruse and the article on the “Haymarket affair” of 1886 (see Messer-Kruse, 2012). In both instances, these experts complained that their “true” contributions were either watered down or deleted from the encyclopaedia. It is these kinds of situations that led Simson Garfinkel to write critically about the same developments described by Reagle and Wales:

Wikipedia has evolved a radically different set of epistemological standards – standards that aren’t especially surprising given that the site is rooted in a Web-based community, but that should concern those of us who are interested in traditional notions of truth and accuracy. On Wikipedia, objective truth isn’t all that important, actually. What makes a fact or statement fit for inclusion is that it appeared in some other publication — ideally, one that is in English and is available free online. (Garfinkel, 2008)

I want to suggest, however, that by accepting Wikipedia’s own claim that truth has been replaced by the NPOV, both sides of this debate have missed the most crucial dimension of Wikipedia’s relation to truth. Contrary to the consensus, there is no escaping the truth of Wikipedia. To explore these (two) relations to truth, I turn to its supposed replacement, the Neutral Point of View.

The Neutral Point of View

Iterations of NPOV are coextensive with the history of Wikipedia. In his oft-cited Slashdot essay “The Early History of Nupedia and Wikipedia: A Memoir”, project co-founder Larry Sanger recalls how a version of NPOV was already established in Wikipedia’s precursor, Nupedia:

Also, I am fairly sure that one of the first policies that Jimmy and I agreed upon was a “nonbias” or neutrality policy. I know I was extremely insistant upon it from the beginning, because
neutrality has been a hobby-horse of mine for a very long time, and one of my guiding principles in writing “Sanger’s Review.” Neutrality, we agreed, required that articles should not represent any one point of view on controversial subjects, but instead fairly represent all sides. (Sanger, 2005)

The original Nupedia policy was titled “Lack of Bias” and its core elements are strikingly similar to the current Wikipedia equivalent. Some especially pertinent excerpts include:

Nupedia articles are, in terms of their content, to be unbiased … This requires that, for each controversial view discussed, the author of an article (at a bare minimum) mention various opposing views that are taken seriously by any significant minority of experts (or concerned parties) on the subject. In longer articles, of course, opposing views will be spelled out in considerable detail. In a final version of the article, every party to the controversy in question must be able to judge that its views have been fairly presented, or as fairly as is possible in a context in which other opposing views must also be presented as fairly as possible…

On any controversial issue, it is usually important to state which views, if any, are now (or were at some time) in favour and no longer in favour (among experts or other specified group of people). But even this information can and should be imparted in such a fashion as not to imply that the majority view is correct, or even that it has any more presumption in its favor than is implied by the plain fact of its popularity.

To present a subject without bias, one must pay attention not just to the matters of which views and arguments are presented, but also to their wording or the tone in which they are mentioned. Nupedia articles should avoid describing controversial views, persons, events, etc., in language that can plausibly be regarded as implying some value judgment, whether positive or negative, except when the judgment is on some relatively innocuous matter and is virtually universal. It will suffice to state the relevant (agreed-upon) facts, to describe various divergent views about those facts, and then let readers make up their own minds about what the correct views are. (Nupedia Editors, 2000)

Sanger notes how the “Lack of bias” policy from Nupedia was translated into Wikipedia very early on as a “Rule to Consider” and shortly thereafter into the Neutral Point of View policy by Wales — a name disliked by Sanger for its oxymoronic status (2005). Despite the name, the core elements of the initial policy remained intact.

The NPOV page has existed since 2001, and as of October 2011 it has been edited roughly 4,500 times (toolserver.org, 2011). During this time it has undergone many transformations and the current version is much longer than the first revised version of November 2001 (the earliest version available on Wikipedia). Nonetheless, the core aspects of NPOV, its most forceful statements, remain largely unchanged. The current “Wikipedia:Neutral point of view” page has a text box near the top noting its status as an English Wikipedia policy and a “widely accepted standard that all editors should normally follow” (Wikipedia Contributors, 2011a). A square box to the right signals that NPOV is one of the five pillars and a “core content policy”. The opening text reads:

Editing from a neutral point of view (NPOV) means representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources. All Wikipedia articles and other encyclopedic content must be written from a neutral point of view. NPOV is a fundamental principle of Wikipedia and of other Wikimedia projects. This policy is non-negotiable and all editors and articles must follow it.

“Neutral point of view” is one of Wikipedia’s three core content policies. The other two are “Verifiability” and “No original research”. These three core policies jointly determine the type and quality of material that is acceptable in Wikipedia articles. Because these policies work in harmony, they should not be interpreted in isolation from one another, and editors should try to familiarize themselves with all three. The principles upon which this policy is based cannot be superseded by other policies or guidelines, or by editors’ consensus. (Wikipedia Contributors, 2011a)

The first sentence summarises the epistemic stance of Wikipedia, with the rest of the passage indicating the force of this stance (“non-negotiable”, “cannot be superseded”) and pointing to its two key allies, “Verifiability” and “No original research”. While the orientation towards bias has been slightly weakened — “as far as possible without bias” — and the concept of “reliable sources” has been included, much of this opening statement closely mirrors the first iteration from Nupedia.

What has been crucially preserved from the outset is a particular two-fold relation to truth. Neutrality, defined interchangeably as non-bias or lack of bias, attempts to distance itself from the truth-battles of the outside world, that is, from contests of truth that take place outside of Wikipedia. For example, it no longer matters if a statement, “Jesus was resurrected”, corresponds to an actual reality of the figure Jesus rising from the dead. Such distancing from these battles in turn enables an inclusiveness, where competing truths — reconfigured as conflicting “points of view” — can all be subsumed into the encyclopaedic mode, albeit under quite specific conditions. For example, “Jesus was resurrected” might appear instead as “Most Christians believe in the resurrection of Jesus” or “The belief that Jesus was resurrected is a core component of
Christian faith” and such statements therefore become compatible with other, non-Christian perspectives on Jesus. Indeed, the entry on Jesus contains six main “religious perspectives”, a sub-section on “other [religious] views”, and a section dedicated to “historical views” (Wikipedia Contributors, 2011b).

This first relation to truth, one of distancing and inclusion, is captured in the part of the statement that requires “representing fairly … all significant views”. It is restated in the “this page in a nutshell” box (at the top of the page) as: “Articles mustn’t take sides, but should explain the sides, fairly and without bias”. How does this actually work? What is the truth of this statement? Another way of putting it is: how is the first relation to truth, one of distance and inclusion, established and confirmed? The NPOV entry provides most of the answers.

After the sentence, “Observe the following principles to achieve the level of neutrality which is appropriate for an encyclopedia”, there is a list of six principles:

- Avoid stating opinions as facts.
- Avoid presenting seriously contested assertions as facts.
- Avoid presenting uncontested assertions as mere opinion.
- Prefer non-judgemental language.
- Accurately indicate the relative prominence of opposing views. (Wikipedia Contributors, 2011a)

These principles spell out a precise relation to “outside” truth, at the level of individual statements. The first two require weakening the truth-value — or “facticity” — of anything contested. Any statement whose truth-value is contested “should be attributed to the text to particular sources, or where justified, described as widespread views, etc.” (Wikipedia Contributors, 2011a). The fourth principle is similar to these first two, but directs the focus to language and authorial voice. Another way of putting it would be: “avoid adding opinions to facts”. The third principle is the converse of the first two principles: if something is uncontested, don’t weaken its truth-value. This principle comes very close to participating in “truth battles” with the distinction, perhaps, that what matters is not whether or not a statement is actually true, but whether its truth is contested. The above example of Jesus’ resurrection captures this distinction: The statement “Jesus was resurrected” is clearly contested and thus cannot be included. By reformulating it as “Most Christians believe in the resurrection of Jesus”, the statement is weakened, at least in relation to the resurrection of Jesus.

However, the focus of the truth-value of the statement has actually been redirected from the resurrection of Jesus to whether or not most Christians believe this to be true. Because this element of the statement, “Most Christians believe”, is an “uncontested assertion”, it therefore cannot be presented in a weakened form: “It is the opinion of X that most Christians believe....”. And despite Wikipedia’s “abandonment” of truth, it nonetheless has a whole regime (of truth) in place for determining whether or not this transformed statement, with its altered focus on “Most Christians believe”, is in fact uncontested.

The final principle requires all of these newly formed statements to be ordered in relation to one another, and this order is determined by an outside reality. Together, these five principles explain how NPOV is established at the level of individual written statements, but by no means do they represent the limits of this content pillar. “Neutrality procedures” also apply to the naming of articles, the structure and arrangement of articles, the “weight” given to particular perspectives (eg, 1,000 words on a minor perspective, while the majority view is only 100 words long), research methods for acquiring sources, and so on.

There is, therefore, a whole other relation to truth to be found in the NPOV pillar. The truth-value of a statement is by no means rejected, merely redirected. And while I have described this regime of truth as the internal truth of Wikipedia, in actual fact the twofold relation to truth cannot be grasped entirely in terms of an inside (a truth of NPOV) and outside (the truth battles beyond the encyclopaedia). Instead, the reach of NPOV extends well beyond the limits of the encyclopaedia. NPOV must be understood as a grid of intelligibility, a set of forceful statements that circumscribe a world beyond the encyclopaedia as well as the precise manner in which to engage with it; it is an internal truth with an external reach. It is also at this point that the NPOV policy’s allies become especially important.

Alongside NPOV, “No original research” and “Verifiability” make up Wikipedia’s three core content policies, which are designed to work in unison: “Because these policies work in harmony, they should not be interpreted in isolation from one another” (Wikipedia Contributors, 2011a). “No original research” establishes a pre-existing outside world as the only legitimate source of encyclopaedia statements. But the outside invoked by “No original research” is very specific: “Wikipedia does not publish original thought: all material in Wikipedia must be attributable to a reliable, published source” (Wikipedia Contributors, 2011c). The pre-existing outside world is purely discursive, a world comprised solely of sources. It is on this level, or in regard to this outside that Wikipedia engages in battles for truth. While I won’t go into detail, there are extensive criteria for what constitutes a reliable source, a published source, and indeed, a source in itself. The function of “Verifiability” in this regard is to establish the reality of this outside world of sources and the method for connecting to it (via citation).
NPOV is the pillar of all content policy, working in “harmony” with “No original research” and “Verifiabilty”. Together, these three core content policies sit atop a whole body of related policies, guidelines and essays, which all work to define the contours of the project: the precise rules of a statement’s formation and the threshold of statement inclusion; the arrangement of and relation between statements; and what constitutes the “source” world beyond the encyclopaedia formation and how to approach it. While outside battles for truth are explicitly rejected — “The threshold for inclusion in Wikipedia is verifiability, not truth” (Wikipedia Contributors, 2011d) — Wikipedia nonetheless has a whole body of forceful statements whose function is to establish the truth of any particular statement; a truth of what is neutral, non-original, published, reliable, attributable, and verifiable.

A Return to Truth

Recognising Wikipedia’s other relation to truth, its regime for establishing the truth of neutrality, non-originality, verifiability and so on, permits a reframing of the initial debate about the abandonment of truth. Most notably, Wales’ and Reagle’s position that collaboration in Wikipedia is successful because NPOV permits a “sidestepping” of philosophical debates about truth needs to be reversed: Collaboration in Wikipedia is possible precisely because there is a clear regime of truth in place for sorting and filtering competing contributions. Rather than abandoning the truth, collaboration is possible because the truth of how to contribute is in place and well established. (As we saw, iterations of the NPOV are co-extensive with the history of Wikipedia, extending even to Wikipedia’s precursor, Nupedia.) While this might seem like a minor adjustment, it isn’t. Those who celebrate Wikipedia’s abandonment of the truth do so because they see in this abandonment a form of collaboration largely devoid of major antagonism. There is in this position the recognition that the quest for truth can divide people and lead to heated and often irresolvable conflicts. If the problem of the truth can be avoided, formerly divided people can be brought together: “Someone who believes that truth is socially constructed, and somebody who believes that truth is a correspondence to the facts of reality, they can still work together” (Wales, cited in Reagle, 2010, p. 53). More than this though, it is the dream of Wikipeidians that people with radically different “points of view” — not different definitions of the truth, but different versions of it — can all be brought together under the umbrella of neutrality. Once it is recognised that even neutrality requires its truth procedures, the dream of universal collaboration through neutrality must be abandoned. This isn’t at all to suggest that Wikipedia’s truth-based collaboration is evil or wrong, just that it exists because of and not in distinction to the truth. And because Wikipedia’s knowledge cannot be divorced from truth, neither can it be divorced from all the messiness that comes with contests for it.

Conversely, we can also see that those who lament Wikipedia’s abandonment of the truth have misdiagnosed the problem. The cases mentioned above (Connelly and climate change; Messer-Kruse and the Haymarket affair) are not instances of recognised experts in possession of true knowledge being shunned because the actual truth doesn’t matter in Wikipedia. Rather, these instances involve a clash of truth regimes and related expertise. They are episodes where the truth of Wikipedia is in conflict with the truth of the scientist or the truth of the historian and where the expertise of these individuals plays off against the expertise possessed by Wikipedia’s editors. Rather than dismiss Wikipedia for abandoning the truth, we would be better served by attending to the procedures in which the truth of Wikipedia is established. It is from such a vantage point that we might begin to understand how power is distributed throughout the project; from where authority derives; how some contributions are accepted over others; and how the project maintains order and coherence in the face of its “spontaneous division of labour” (Shirky, 2008, p. 118) and “ad-hocratic” (Bruns, 2008, p. 25; Konieczny, 2010) mode of governance. If we wish to avoid the hype about collaboration without replicating conservative arguments about the death of the expert, we must return to the truth of Wikipedia.

References


Dossier “Academic research into Wikipedia”

Wiki Loves Monuments 2011: the experience in Spain and reflections regarding the diffusion of cultural heritage

Emilio José Rodríguez Posada
Computer Engineering Bachelor degree and predoctoral student (University of Cadiz)
emiliojose.rodriguez@uca.es

Ángel González Berdasco
Computer Engineering Bachelor degree (University of Oviedo)
UO179253@uniovi.es

Jorge A. Sierra Canduela
President of Wikimedia Spain
Degrees in Chemistry and Theology and diploma in Therapeutic Pedagogy
Secondary school teacher
jorgesierra@wikimedia.org.es

Santiago Navarro Sanz
Member of Wikimedia Spain
Student of the degrees in Chemical Engineering, Chemistry (Jaume I University of Castellón) and History (UNED)
santiagonavarro@wikimedia.org.es

Tomás Saorín
Lecturer at the Faculty of Communication and Documentation (University of Murcia)
tsp@um.es

Submission date: February, 2012
Accepted date: April, 2012
Published in: May, 2012

Abstract
Wikipedia came into being in cyberspace. Its early years were marked by asynchronous work by users located all over the world who hardly ever related on a personal level outside the net. With time, some of the volunteers met at what were called wikimeetups, encounters initially aimed at tightening bonds which did not bring about any direct improvement to the project content. Face-to-face initiatives later took place that involved not just volunteers but also cultural entities. The most recent event and the one with the greatest impact was Wiki Loves Monuments 2011, a competition to photograph monuments in 18 European countries, including Spain. The high level of participation led to 160,000 photographs of monuments being taken, with Spain occupying the third place in terms of number of photographs. In this paper we explore the origins, implementation, development and results of Wiki Loves Monuments. The success of the 2011 edition and requests from other countries has led to organization of Wiki Loves Monuments 2012, which will be held at the global level.

Keywords
Wiki Loves Monuments, cultural heritage, image banks, Wikimedia Commons, free knowledge, photography
Wiki Loves Monuments 2011: the experience in Spain and reflections…

Introduction

Wikipedia is shaped by content and by people. Supporting the encyclopaedia and sister projects is an active community of users who writes articles, oversees changes, promotes projects, resolves conflicts and maintains tools, with the aim of progressively improving quality and scope. Wikipedia came into being in cyberspace. Its early years were marked by asynchronous work by users located all over the world who hardly ever related on a personal level outside the net. This virtual community focused on a very specific task which it performed using a radically innovative horizontal cooperation mechanism. Maturation has led to the incorporation of an organizational and personal dimension, first through the constitution of the Wikimedia Foundation to support and foster projects and then through the organization of regular face-to-face meetings attended by editors of Wikimedia projects (Wikimedians). Some volunteers began to meet at meetups although the first edition of Wikimania was not held until 2005, as an annual conference designed to bring together the hundreds of participants in Wikimedia projects. An annual WikiSym symposium is now also held in which information on research into wikis is shared, with Wikipedia work occupying centre stage.

Initially the Wikimedian meetings had a social objective and were attended by people who had worked together on projects but who had never met. Although these encounters may have strengthened ties between members of the community, they did nothing to directly improve project content. Gradually, however, activities that combine both objectives have taken shape.

In addition to the free encyclopaedia, an enormously vital resource in the form of an images and media bank called Wikimedia Commons was brought about. Created in 2004, this multimedia repository had, by March 2012, more than 12 million files donated by volunteers and, increasingly, by cultural institutions under a free licence. Other projects exists that make vast amounts of free multimedia material available, such as the Geograph Britain and Ireland project, which has 2.8 million photographs, and Flickr, with over 51 million photographs (or 215 million if all the possible Creative Commons licences are included). Since it was centralized for all Wikimedia projects, Commons growth became steady, with ongoing improvements in quality and diversity, applying the policies of collaboration and reuse.

1. The Wikimedia Commons image bank

While Wikimedia Commons covers all kinds of media, this paper will refer only to images and, in particular, photographs. Embedded images need to be selected for their informational value in supplementing encyclopaedia articles. Wikipedia uses the adjective educational to differentiate its bank from other social image banks: images must have a value in terms of “providing knowledge; [being] instructional or informative”.

Illustration is not just a complement for the main textual information. An image itself conveys information, reinforces explanations and improves the overall readability of any text. However, there are limitations in terms of potential reuse, as most images are protected by copyright, unlike properly cited literature sources, which, in summarized form, can be used to create articles (Saorín; Pastor-Sánchez, 2011).

Wikipedia Commons is not like other collaborative digital image archives, since most have or potentially have concrete illustrative purposes. Wikipedia Commons content can also be used in any other context, digital or otherwise, under the same kind of free licence. It is a resource bank for both Wikimedia projects and any other social, editorial or commercial communication project.

Wikipedia Commons has a very careful policy regarding the use of images. It only accepts free content. The images are stored and organized on Wikimedia Foundation servers and can then be used for Wikipedia articles or other Wikimedia projects. Unlike Wikipedia, Commons is a centralized and multilingual resource that involves users in many languages. Initially, each Wikipedia had its local repository of images, but as the encyclopaedic project became more mature, the need to combine all graphic resources became clear, given their universal value. Even so, however, many editions of Wikipedia still allow images to be uploaded locally.

2. Wikipedia and cultural institutions

Although Wikipedia is written and constructed as a social initiative, public and private social bodies can also directly and indirectly contribute. Wikipedia may exist in digital space, but it is not independent of the social infrastructure that supports its editors and readers. The need for involvement in Wikipedia to impregnate cultural sectors is increasingly becoming apparent. As Antonio Lafuente (2011) pointed out, all studies on the development of free software have demonstrated "the importance of certain public or private organizational structures in the sustainability of projects and also show how Wikipedia would be impossible without the existence of public libraries, universities and museums".

Institutions are slowly beginning to become aware that they can better fulfil their social function by dedicating a portion of their resources to building the digital commons – of which Wikipedia is one of the most outstanding examples. Librarians, researchers, museum curators, heritage experts and educators are being called on to change the knowledge diffusion model, not only as citizens but also as agents commissioned with a social mission regarding education and culture transmission. Institutions, taking their conflict of interest policies into account, could each develop a model of participation adapted to their resources and interests. Wikimedia Outreach compiles and evaluates best practices in order to expand the scope of the Wikipedia project by working within the real-world cultural ecosystem.

One such activity is called GLAM (Galleries, Libraries, Archives and Museums). In a broader sense, also including science museums, interpretation centres, public broadcasters, etc. The GLAM-Wiki conference of 2009 developed a set of recommendations for collaboration between Wikipedia and cultural institutions, aimed at addressing the gap between concepts regarding cultural institutions and open culture. The recommendations covered aspects such as project financing, moral and reproduction rights and technologies, and also incorporated assessment of degree of access as an indicator of public culture policies.

An example is that of the Wikipedians in residence (for example, at the British Museum or the Museu Picasso), commissioned with the task of improving articles referring to institutions. Several meetings have been held in Spain (all in Catalonia) to promote cooperation between Wikipedians and cultural institutions. The first was a GLAM-Wiki meeting held in Barcelona in March 2011, when Wikimedia volunteers met with representatives of Museu Picasso, MACBA, Creative Commons Spain and patrimoni.gencat.5 Other meetings have been held in cities like Lleida and Girona, and there also have been collaborations with institutions such as the Joan Miró Foundation in Barcelona.

Cultural institutions committed to the free dissemination and use of information have added valuable collections to Wikimedia Commons. The label “Commons partnership” is applied to organizations that have made material contributions. One major example of the release of photographic and visual heritage material is the State Library of Queensland’s provision of 50,000 images and their metadata.

Recent years have witnessed initiatives in which Wikimedians are given a visible identity (Wikipedia Multimedia Events),6 which, as well as increase content, seek to highlight the work of volunteers for mainstream institutions and the media and to strengthen community ties. Experiences in collaboration with museums and other initiatives such as Wiki Loves Art are evidence of this.7 The latest initiative and the one with the greatest impact to date is the subject of this article: Wiki Loves Monuments.8

5. http://www20.gencat.cat/portal/site/Patrimonii/menuitem.6a2dec9a300f68a8cd0181db00e1a0/?vgnextoid=079256e10b371110VgnVCM1000008d0c1e0aCRCD&vgnextchannel=079256e10b371110VgnVCM1000008d0c1e0aCRCD&vgnextfmt=detail&contentid=3f0372da4f4ce210VgnVCM2000009b0c1e0aCRCD
3. Wiki Loves Monuments

3.1. Origins

The first edition of Wiki Loves Monuments was held in the Netherlands in September 2010, although it had been preceded by other initiatives. In Wiki Loves Art Netherlands, held in June 2009, 40 Dutch museums allowed volunteers to visit and photograph their collections. Of the photographs taken, 5,413 were uploaded to Flickr and subsequently transferred to Wikimedia Commons. The success of the event led to a search for new targets to be photographed. Wikipedia Nederland had Wikimedia Commons. The success of the event led to a search for new targets to be photographed. Wikipedia Nederland had a project underway to create lists of Dutch windmills and write, organize and categorize articles on them. 10 Building on this initiative, volunteers, in coordination with Wikimedia Nederland, obtained data on national heritage sites (Rijksmonumenten) from Rijksdienst voor het Cultureel Erfgoed, the organization responsible for protecting and preserving Dutch national heritage. These data were transformed into lists by provinces, which were then edited by Wikipedians. The following year Wiki Loves Art Netherlands became Wiki Loves Monuments, taking the form of a photography competition. Thanks to this initiative, over 12,000 images of 8,000 monuments were uploaded by over 250 users to Wikimedia Commons.

3.2. Wiki Loves Monuments Europe 2011

Towards the end of 2010, Wikimedia Nederland proposed a Wiki Loves Monuments 2011 to other European chapters of Wikimedia and the organization of an international competition with selected photographs from each country.

A Wikimedia Chapters Meeting, held in Berlin between 25 and 27 March 2011, presented Wiki Loves Monuments. Discussion focused on expanding the competition to other European countries, potential problems, the participation of the Wikipedian and photographer communities (Flickr, Panoramio) and the role of the Wikimedia chapters in contest organization. Underlined was the need for the chapters to act as intermediaries between the communities and the cultural institutions and also the need to provide appropriate software tools that facilitated file uploads. Another coordination meeting was held in Berlin from 13 to 15 May, where it was agreed to create upload wizards, locate possible sponsors at the European level and develop judging criteria. The 2011 edition was held in 18 countries: Germany, Andorra, Austria, Belgium, Denmark, Spain, Estonia, France, Hungary, Luxembourg, Norway, Netherlands, Poland, Portugal, Romania, Russia, Sweden and Switzerland. Each country nominated ten photographs to participate in the European phase of the competition (except Belgium and Luxembourg, which shared their nominations). A jury evaluated the material according to criteria of quality, originality and usefulness of the image to Wikipedia and then selected a winning photograph and eleven finalists. The winning photograph in 2011 showed Chaïnja monastery on the outskirts of Bucharest. A Spanish entry showing stables in the castle of Sant Ferran in Figueres (Girona) was among the runner-up photographs.

Europeana, the European open-access digital library, awarded a special prize for the best Art Nouveau photograph; the winning photo, chosen by voting through Flickr, showed the inside of the Museum of Applied Arts in Budapest. There was also a prize for the best photograph from the Pyrenees-Mediterranean Euroregion, formed by the Spanish regions of Catalonia and the Balearic Islands and the French regions of Languedoc-Roussillon and Midi-Pyrénées; the jury selected five photographs of Languedoc-Roussillon, three of Catalonia and two of Midi-Pyrénées and chosen as the winner was the church of Notre-Dame-des-Anges in Collioure. Also awarded was a prize for the best video, which showed a wallpaper printing machine from 1877 (the first to use 26 colours), as displayed in the old Leroy factory in Saint-Fargeau-Ponthierry (France).

3.3. Wiki Loves Monuments Spain 2011

In Spain there were two parallel events: Wikimedia Spain organized a competition for Spain and Amical Viquipèdia organized another for Catalan-speaking areas. It was decided that the competition would centre on objects of cultural interest in the monuments category, totalling more than 15,000 items.
Since most of the objects of cultural interest in Barcelona and Valencia had already been photographed, Wikipedia in Catalan included monuments listed in the regional inventories, namely, Cultural Assets of Local Interest for Catalonia and Buildings of Local Relevance for Valencia. These photographs would not be counted for the “Largest number of monuments photographed”, prize in Spain.

Wikimedia project pages, a mailing list and virtual meetings were the means used to coordinate activities between Wikimedia Spain and Amical Viquipèdia. A number of Catalan volunteers also met in Barcelona in April 2011. To promote the competition, Amical Viquipèdia also organized an event called Viquipèdia Prem Barcelona.20

Although the files were uploaded to Wikimedia Commons, Wikipedia was used to create the lists of monuments. Heritage projects were launched in the Spanish21 and Catalan22 versions of Wikipedia. A project on heritage was not created in Wikipedia in Galician, although a list of monuments in Galicia was drawn up. A WikiProject on heritage with a list of monuments in Aragon was also created in Wikipedia in Aragonese,23 although the Aragonese volunteers were not involved in organizing the competition. Initially it was intended to draw up lists of monuments by provinces in Wikimedia Commons, so as to bring all users together regardless of language; ultimately, however, the work was done via the different language versions of Wikipedia, where visibility was higher.

One of the requirements for creating tables and subsequently processing databases was that each monument would have a unique identifier. The Spanish Wikipedia lists were created from data available from the website of the Ministry of Culture.24 The Catalan Wikipedia lists were based on data from the regional administrations of Catalonia, the Balearic Islands and Valencia. All the lists were improved and corrected by volunteers, who administrations of Catalonia, the Balearic Islands and Valencia. Catalan Wikipedia lists were based on data from the regional inventories, namely, Cultural Assets of Local Interest for Catalonia and Buildings of Local Relevance for Valencia. These photographs would not be counted for the “Largest number of monuments photographed”, prize in Spain.

To assist in locating the monuments, a map was prepared that geotagged monuments for which coordinates were available (around 50% of the total).29 This tool proved very useful, as participants were able to locate nearby monuments for photographing and to plan routes to visit as many monuments as possible. Figure 1 shows an early screenshot for the area of Madrid, with the red dots indicating monuments for which images are available and the white dots indicating monuments without images. The map was updated daily and by the time the regions, even though these properties should, in any case, be included in the General Register of Properties of Cultural Interest of the Sub -Directorate General for the Protection of Historical Heritage. The information provided by the different regional administrations was therefore rather uneven.

This difficulty was aggravated by the fact that the identification codes used by the regional administrations are very different and the fact that some regional portals do not even list these codes (a fundamental identification requirement for the competition). This was further compounded by the complexity of identifying generically declared properties of cultural interest, which in many cases were not assigned any code. This was the case of castles and defensive works, protected by a decree dated 22 April 1949; badges, emblems, heraldic stones, pillory stones, boundary crosses and similar items of historical interest, protected by Decree 571/1963; granaries of Asturias and Galicia, protected by Decree 499/1973; and megalithic monuments, prehistoric caves and other prehistoric remains on the islands of Mallorca and Menorca, protected by Decree 2563/1966. For these reasons, most of the lists only feature properties that appear on the Ministry of Culture website, in the expectation of being able to obtain further data for future editions.

Flickr was also used for file uploads; two groups were created, one for Spain called Wiki Loves Wikimedia Monuments 2011 Spain25 and another for Viquipèdia Amical called Wiki Loves Monuments CAT.26 Of the ten selected photographs for Spain, four were uploaded to Flickr directly and not to Wikimedia Commons. The Department of Culture of the Generalitat of Catalonia decided to include some of the photographs uploaded to its architectural heritage inventory.29

To assist in locating the monuments, a map was prepared that geotagged monuments for which coordinates were available (around 50% of the total).29 This tool proved very useful, as participants were able to locate nearby monuments for photographing and to plan routes to visit as many monuments as possible. Figure 1 shows an early screenshot for the area of Madrid, with the red dots indicating monuments for which images are available and the white dots indicating monuments without images. The map was updated daily and by the time the
The competition was over, the number of monuments with images had grown considerably. Even so, despite the high participation level (to be discussed in the results section) and given the number of architectural heritage items, coverage for the rest of the country was uneven; images were available for only just over 30% of the monuments, leaving a great deal of work for future issues.

Figure 1. Map showing monuments for the Madrid area

The Town Council of Plasencia (Cáceres) provided a venue for the awards ceremony and nominated one of the jurors. In the absence of financial support, the competition costs were borne by Wikimedia Spain with a contribution from Wikimedia Nederland. All the information was centralized in a site specifically created for the purpose. The Spanish jury, which consisted of three members, published its decision on 1 November. The awards were presented on 5 November at the old Convento de las Claras in Plasencia. The winning photograph showed the baths of Doña María Padilla in the Alcazar of Seville. Amical Viquipèdia presented the awards for Catalonia on 24 November at the Palacio Moja in Barcelona; the winning photograph show the Torre de Sant Joan in Amposta.

3.4. File uploading

Historically, file uploading for Wikimedia projects has been a problem for inexperienced users, making the usability of Wikimedia Commons problematic. To address this problem, an improved upload form was prepared, with a custom configuration for the event.

The photographs had to be the person’s own work, had to be published under a free licence compatible with Wikimedia Commons (recommended was Creative Commons Attribution-ShareAlike 3.0) and had to be uploaded in September. Contestants had to supply an email for future communications with the organizers and had to include the identification code for the monument in the upload form.

The file organization system in Wikimeditas Commons, which is based on a system of general categories that grow increasingly specialized as browsing deepens, is difficult to use given the size of the category tree, representing more than 12 million files. For example, 11 sub-categories, including subcategories such as Culture, Cultural Heritage and Cultural Heritage Monuments in Spain, have to be navigated to see the photograph, labelled Monasterio de Veruela.jpg, the second prize winner. Although this structure makes it easier to find images on the same topic, the search for individual files usually has to be done using the site search.

Due to the complexity of the category tree with its categories named in English, participants were not asked to categorize images; rather, this task was delegated to Wikimedia Commons volunteers. A minority of users did add detailed categories for their images, but most were uploaded to a default category called “Cultural Heritage Monuments in Spain”, which quickly filled with thousands of images. From day 4, a bot (software that performs tedious tasks automatically) began to classify the images by province using the monument identification codes, thereby significantly reducing the amount of manual work. However, further sub-categorization had to be done manually and had not yet been completed by early 2012.

4. Results

Results for the Wiki Loves Monuments 2011 experiment were analysed from a dataset created from all the relevant data for the uploaded images. This dataset contains several fields, among them, image name, author, date created, date uploaded, file size in bytes, width and height in pixels, the corresponding monument (if indicated on uploading) and country. The Wikimedia Commons MediaWiki API was used to build the dataset for the “Images from Wiki Loves Monuments 2011” category containing all the images for the competition. The creation of this dataset and its release under the CC-BY-SA 3.0 licence (metadata) and GPLv3 (code) allows anyone to reproduce the results or conduct their own analysis.

33. See the list of FAQs: http://commons.wikimedia.org/wiki/FAQ
37. http://commons.wikimedia.org/w/api.php
Although some users uploaded images just a few hours before judging began or after it ended, the analysis was restricted to the month of September, as originally planned. It also excluded the images uploaded through Flickr, whose transfer to Wikimedia Commons took place later.

Between 00:00:00 UTC on 1 September 2011 and 23:59:59 UTC on 30 September 2011, 5,347 users uploaded 163,251 files, accounting total for 448,307 megabytes, (more than 437 gigabytes) to Wikimedia Commons for the Wiki Loves Monuments 2011 event. Just for comparative purposes, the largest ever photography competition in the world to date, according to the Guinness Book of World Records, handled a total of 126,501 images.

Uploads over the month can be seen in Figure 2. Participation on a daily basis was very steady (no day had fewer than 1,500 uploads) and the average was 5,400 files a day; in the last three days, uploads exceeded 10,000 images a day, with 22,000 images uploaded on day 30.

As regards participation by country, Spain ranked third with 16,708 images (10.2% of the total). Figure 3 shows how some countries carried more weight in the competition; for example, the top four countries (Germany, France, Spain and Poland) together accounted for over 50% of all the files.

The purpose of the event was to photograph as many monuments in the participating countries as possible so as to create photographic material to illustrate Wikipedia articles. Using GLAMorous, a tool designed to calculate the use of Wikimedia Commons images in Wikimedia projects, it was possible to determine that 40,000 images (just over 20%) from the category “Images from Wiki Loves Monuments 2011” and 2,387 images (13%) from the category “Images from Wiki Loves Monuments 2011 Spain” are already being used in some language edition of Wikipedia. The percentages may seem low, but (at the time of writing) only five months have passed since the end of the competition; furthermore, an article about the monument first needs to exist and that implies a longer-term project. There are also monuments with a large number of photographs that cannot all be shown, although they can be retained as an extra resource.

Although more than 52,000 images (around a third of the total) as yet have no monument identifier code (identification is a task still being completed by volunteers), if we consider those that do contain this metadata (around 111,000 images), over 36,000 photographs have been taken of monuments in Europe and around 3,600 of monuments in Spain. The most photographed monument, with more than 400 images, was the Abbey of Sainte-Foy (code PA00093999), located in Conques (France). The most photographed monuments in Spain were the Sagrada Familia in Barcelona, Tarragona Cathedral and the Cathedral of Cordoba (a former mosque), with more than 100 photographs each.

Looking at the distribution of images by regions, Figure 4 shows that participation was greatest in Catalonia, which accounted for more than 45% of the images.

40. 415 images of Spain were uploaded to Flickr in September but were uploaded to Wikimedia Commons after a manual review in October.

Figure 2. Number of files uploaded on each day of the competition

Figure 3. Proportion of images uploaded by country
Figure 4. Photograph distribution by region of Spain

The image upload rate by country throughout the month of September is depicted in Figure 5. By day 9, the country with the highest number of pictures was Germany. Spain, Poland and Portugal had a similar number of uploaded images, but Spain was eventually placed third after Germany and France. Hungary was a special case, as participation commenced on 21 September, therefore the completion date for the local competition extended until 21 October. By late October, 5,043 images for Hungary had been uploaded. Note, however, that comparisons between countries are not realistic given the fact that population and user community sizes differ for each of the countries.

Figure 5. Country shares over the month

A survey was conducted after the competition. Responses were mostly positive, but noted the need to improve the uploading to that of sites like Flickr and difficulties to locate monument identification codes. The main reasons for participating in the competition were the interest in collaborating with free-culture projects, the prizes and the fact that photographs would be viewed and used by thousands of internet users.

Media coverage of the event varied considerably from one country to another. Wikimedia Russia organized an exhibition of the 37 best photographs, using the QRpedia system to link to Wikipedia articles on the monuments from mobile phones. In Poland, the competition, which went by the name Wiki Zabytki Lubi, was nominated for the Historical Event of 2011 plebiscite organized by the Museum of Polish History and the site Historia.org.pl.

5. Conclusions and the future

We have described the Wiki Loves Monuments 2011 experience, in which 18 European countries, including Spain, participated. This edition was held in response to the favourable reception given to the 2010 edition in the Netherlands and also to the growing interest in Wikimedia community-based cultural initiatives and activities, aimed at expanding not just text but also multimedia content.

The outcome has been more than satisfactory, as indicated by the more than 160,000 images uploaded by participants. For Spain, more than 16,000 photographs of some 3,600 monuments were uploaded, representing 10.2% of the total. The success of the 2011 edition and requests from other countries has led to the organizers to begin work on a worldwide Wiki Loves Monuments for 2012. Iberocoop, which coordinates Spanish and South American Wikimedia chapters, is organizing a working group, participated in by ten countries, which will organize the competition in Latin America and which will also provide legal advice in countries, like Argentina, where legislation regarding freedom of panorama and free licences is ambiguous.

The outcome of this initiative in terms of the dissemination of heritage is undeniable, to which can be added the preservation in images of monuments that may be damaged or destroyed. Recent earthquakes in Haiti and Chile, for example, damaged part of their heritage. Neither of these countries participated in any editions to date, although Chile will participate in Wikimedia Wiki Loves Monuments 2012.

The role of the Wikimedia chapters is crucial, as they bring together large numbers of experienced and active Wikipedians, who act as the ideal nexus between the communities and...
cultural institutions, which, with their technical means, can publicize projects and encourage participation by offering awards.

Finally, Wiki Loves Monuments demonstrates that there are alternative strategies for developing knowledge and for documenting our common cultural heritage. The growing influence of such initiatives on European cultural and digital policies is evident. The enormous advantages of working with licences that are as open as possible and that allow data and content to be reused is becoming increasingly understood, as it makes our heritage truly available to us as digital citizens.

References


RECOMMENDED CITATION

ISSN 1575-2275

This work is subject to a Creative Commons Attribution 3.0 Spain licence. It may be copied, distributed, transmitted and adapted, provided you attribute it (authorship, journal name, publisher). The full text of the licence can be consulted here: http://creativecommons.org/licenses/by/3.0/es/deed.en.

Emilio José Rodríguez Posada
Computer Engineering Bachelor degree and predoctoral student (University of Cadiz)
emiliojose.rodriguez@uca.es

He has been awarded grants by the Free Software and Open Knowledge Office of the University of Cadiz, has helped organize several free software seminars and has given talks and workshops on Wikipedia at the University of Cadiz. He has been an editor of Wikimedia projects since 2009 and has developed free software for wikis such as the anti-vandalism bot AVBOT (winner of the 3rd University Free Software Competition in the Best Community Project category) and the StatMediaWiki and WikiEvidens applications to generate statistics and to view these communities online. He recently founded the WikiTeam project which has already preserved hundreds of wikis, and he has created the WikiPapers project to collect all the academic literature referring to this research field.

Universidad de Cádiz
C/. Ancha, 10
11001 - Cádiz

Wiki Loves Monuments 2011: the experience in Spain and reflections…

Ángel González Berdasco  
Computer Engineering Bachelor degree (University of Oviedo)  
UO179253@uniovi.es

Holder of a computer engineering degree from the University of Oviedo, he received an award for his project on an offline wiki reader for MediaWiki. He has collaborated in the organization of Wiki Loves Monuments 2011 and is currently preparing his computer engineering course project and working in the private sector. His interests are security, web technologies and free knowledge.

Universidad de Oviedo  
Campus de Viesques, s/n  
33204 Oviedo

Jorge A. Sierra Canduela  
President of Wikimedia Spain  
Degrees in Chemistry and Theology and diploma in Therapeutic Pedagogy  
Secondary school teacher  
jorgesier@wikimedia.org.es

Born in Palencia in 1981. Professor specializing in special education needs and educational innovation. An actively collaborator with several NGOs, including Amnesty International and PROYDE and a member of the Free Software Foundation. Volunteer for Wikimedia projects since 2006 and president of the Spanish Wikimedia chapter since 2011. Has given numerous talks and workshops on free knowledge, fair trade and involvement in Wikipedia. He participated in Wikimania 2010 and 2011 and represented Spain at the Wikimedia Chapters Meetings in Berlin in 2011 and 2012.

Colexio La Salle Ferrol  
Rúa San Roque s/n  
15406 Ferrol (A Coruña)

Santiago Navarro Sanz  
Member of Wikimedia Spain  
Student of the degrees in Chemical Engineering, Chemistry (Jaume I University of Castellón) and History (UNED)  
santiagonavarro@wikimedia.org.es

Born in Vila-real in 1981. Wikipedia administrator and bureaucrat in Spanish and volunteer for other Wikimedia projects such as Wikimedia Commons and Viquipèdia (in Catalan) since 2006. Founding member and member of the board of Wikimedia Spain. Has given numerous talks and workshops on Wikipedia and other Wikimedia projects. Member of the organizing committee for Wiki Loves Monuments 2011 Spain and representative of Wikimedia Spain in the Wikipedia Chapters Meeting in Berlin in 2012.

Avda. Cedre, 37  
12540 Vila-real (Castellón)
Tomás Saorín
Lecturer at the Faculty of Communication and Documentation of the University of Murcia.
tsp@um.es

Documentalist for the Autonomous Government of Murcia and lecturer at the Faculty of Communication and Documentation at the University of Murcia, teaching information management, museology and documentation. Doctor of philosophy whose thesis was on the subject of library portals. Has published on the subject of the mass culture image of libraries, content management systems and information architecture. In recent years he has developed introductory workshops on the creation of Wikipedia articles for students and lecturers within programmes for information technology innovations in teaching. A member of the Spanish chapter of Wikimedia.

Facultad de Comunicación y Documentación
Campus de Espinardo
Universidad de Murcia
3071 Murcia