ICT - Information and Communication Technologies
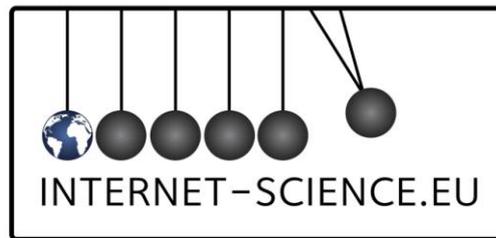
FP7-288021

Network of Excellence in Internet Science

OPEN CALL PROJECT: Contropedia

D2.1 Functional specifications of the Contropedia platform

Due Date of Deliverable: M3

Actual Submission Date: 6 June 2014

Revision: FINAL

Start date of project: November 1st 2013   **Duration:** 12 months

Organization name of lead contractor for this deliverable: UvA

Editor: Erik Borra

Contributor: Tomaso Venturini

Project Information

PROJECT	
Open Call Project name:	Real time analysis and visualization of controversies around Wikipedia articles.
Open Call Project acronym	Contropedia
Project start date:	01/11/2013
Project duration:	12 months
Contract number:	288021
Activity:	THEME ICT-2011.1.1: Future Networks
DOCUMENT	
Document title:	Contropedia D2.1 Functional specification of the Contropedia platform.
Document type:	Report
Deliverable number:	D2.1
Contractual date of delivery:	Month 3, 31.01.2014
Calendar date of delivery:	Month 7, 19.05.2014
Editor:	Erik Borra
Authors:	Tommaso Venturini
Workpackage number:	2
Workpackage title:	Hackathon 1: Individual Controversy Streams
Lead partner:	UvA
Dissemination level:	CO
Date created:	15 May 2014
Updated:	6 June 2014
Version:	2
Total number of Pages	10
Document status:	Final

TABLE OF CONTENTS

Executive Summary

Together with deliverable D2.2 this report describes the ongoing project Contropedia that aims to build a platform for the analysis and visualization of such controversies in Wikipedia.

This deliverable (D2.1) describes the functional specifications of the Contropedia platform, which allowed us to focus our efforts during the first hackathon (WP2) that took place at the University of Amsterdam from 11 to 14 February 2014. The objective of the complete work package 2 was to specify the Contropedia platform and generate the components for the first prototype of the Contropedia. The hackathon was organized around four tasks:

- One for analyzing the sociological context and specifying and operationalizing the research questions we will address and answer with the project (T2.1)

- Two tasks for extracting controversy metrics from the discussion and edit activity separately (T2.2 and T2.3)

- One for building visualizations around these metrics (T2.4).

For an elaborate description of the tasks we refer to the EINS DoW. This deliverable D2.1 focuses specifically on the determination of the relevant questions to answer with the Contropedia platform (T2.1). The results of T2.2, T2.3 and T2.4 can be found in deliverable D2.2.

1. Introduction

Following common practices in user-centered design and marketing, the functional dialogue is written up as the interior monologue of a user (persona) interested in exploring climate change controversies on Wikipedia.

The accompanying legend is as follows:

- Crucial function: should be implemented at all cost

- Nice function: implementing it would be very cool, but it is not necessary

- Optional function, should be implemented if we have some time left

Like more and more people in the world, I am a frequent reader of Wikipedia. Though I know that it is someway different from classic encyclopedias, so far I have always used it in a pretty traditional way, to find basic and reliable information about all kind of things. Sometimes, I have gone directly to Wikipedia in search for information. Most of the time, I have been referred to it as one of the first results of a search engine's query.

Today, however, I am on different quest. After having been repeatedly exposed to personal and media debates on climate change, I decided that I want to know more about this controversy. I am not looking for answers, not straight away: I am fed up with people tell me what's true and what's false about climate. For the moment, I just want to know more about the debate itself. After that, maybe, I'll try to make up my mind on the question, but first I want to understand better what the question is.

I know that Wikipedia is an 'open' encyclopedia, edited collaboratively by thousands of persons and I thought: if climate change is really controversial, then all those people cannot possibly agree on its definition. Maybe I can read the debate among the Wikipedians as a sort of proxy of the larger debate on climate change. However, when I try to find traces of the controversy in the pages dedicated to the discussion in Wikipedia (the 'history' and the 'talk' of my article), I get lost. Such pages are far too complicated for someone like me, who knows little about how Wikipedia works.

And this is where Contropedia comes in. I have found this wonderful online platform that promises to decrypt for me all the quarrels of Wikipedia. Fantastic, exactly what I was looking for, let's see how it works.

2. What is controversial about climate change?

The first thing I would like to know is what is controversial and to what extent. I am sure that there are things on which everyone agrees (no one doubts earth has an atmosphere, right?) and I am sure that not everything is equally problematic. So I would like to see my Wikipedia page augmented with some indicators that tell me which parts are more controversial and which are less. In fact, I would need at least two indicators: how much is a give part of the page discussed (to appreciate the lack of consensus on it) and how many people are discussing it (maybe it is just some crazy guy repeatedly vandalizing the page or a flame war between two users). Now that I think about it, there is also a third indicator that is very important: it is when Wikipedians themselves marks a page or a passage as controversial. I want this to be highlighted as well.

When I say parts of the page, I think first of all to controversial sections. I am not sure that I want to read the whole page (it's kind of long), so I'd like to know where the debate concentrates. Then I can zoom in and see what words, images, links… are specifically controversial in those sections.

Now that I have explored this one page, I'd like to know if there are other related pages that are also controversial and I would like to know that in two ways. First, while I read the page, I would like the hyperlinks that point toward other controversial pages to be highlighted (and I don't want to be confused, so I want to distinguish whether the link is controversial or the page that it points to). Second, I'd like to zoom out a little more and see a bigger picture of the pages about climate change: how many are there? How are they connected? Which are the oldest? Which are the most cited? Which are the most edited? Which involve most people? Which are most controversial? In particular, I want to know if there are groups of pages that are both highly linked and highly controversial, because that would be the core of the quarrel, right?

This last information is important because, in the rest of my exploration of Contropedia, my focus will not necessarily be a single page. It is possible that some controversies are confined to a single page of Wikipedia, but most debate (and certainly that on climate change) will spread to a few dozens of pages. Contropedia should help me to identify these pages and define them as my corpus of analysis.

Finally, since Wikipedia is neither the whole web nor the whole world, I'd also like to know which other documents (web pages or scientific references) are most cited and/or most controversial about climate change.

3. Who is debating about climate change (and how are they connected?)

Now that I have learned which questions constitute my controversy, I want to discover who the main actors of it are. Being open-minded, my definition of actor is pretty broad: of course I am thinking about individuals, but I acknowledge that institutions can be actors too, as well as instruments, natural elements, ideas... basically anything that has a name and does something. So now I select one page or a series of pages and I hope Contropedia will tell me what actors appears on those pages and how relevant and controversial they are (please note that relevance and controversiality are to me two different things and I don't want them mixed up). Since I am a demanding user, I also want to know how these actors are connected, for example if they are often mentioned together or if they have pages in Wikipedia that links one another.

After having extracted the actors *in* the debate, I now wish to extract the actor *of* the debate, that is to say, in general, the Wikipedians that have most edited the pages that I am focusing on and, specifically, the Wikipedians that have been most active in rousing the controversy. And I'd like to know a bunch of things about these people: Which role do they have in Wikipedia? When they have started editing and when they stopped it they did? Which other pages do they edit? Do they focus on climate change or do they edit everywhere in Wikipedia? With what frequency do they edit the pages on climate change? How frequently do they edit at all? Do their edits remains on the pages or are they often deleted/modified?

I am also curious of understanding how debating Wikipedians are related. Do their profile pages link each other? Are they parts of the same interest groups in Wikipedia? Do they edit the same pages? Do they oppose the same opponents?

4. Where are the controversies discussed?

After what is debated and who is debating, the question that Contropedia should answer is where is the debate is taking place – both geographically and thematically. As for the geography is concerned, things are relatively simple: what I want to know is which toponyms occur most often in the pages dedicated to climate change and whether their mentions are controversial or not.

Also, since Wikipedia exists in many different languages, Contropedia could help me compare the debate over different linguistic space. Of course the best way of doing this would be that I create different corpuses of pages (one for each of the pages that I want to study), but since I am lazy, I'd like Contropedia to do the job for me by following the 'languages' links in the left column of each article. The result will be less precise, but it will save me a hell lot of time.

5.   When are controversies controversial?

So far Contropedia has given me information on my controversy in the present, or to be more precise 'up to present' (for example, by telling me who has modified a page from day it was created). But the same information would be even more interesting if 'periodized'. Who has debated one page in the last month? Who did one year ago? What has been controversial in the last week? And what is not controversial anymore? Did controversy move from one language to another? And since the relevant periodization (day, week, month, year...) may be different for different controversies, I'd like Contropedia to ask me over what period I want to aggregate the data.

Once data has been periodized a couple of interesting things become possible. First, Contropedia can tell me about the trajectories of different controversies (they are mounting, defusing; they go up and down as a rollercoaster? The opposition between the editors is becoming more complex or more binary?). Second, Contropedia can detect events: periods in which the trajectory of the controversy has suddenly changed.

6.   Conclusions

As written in our proposal, we foresee a variety of users such as Wikipedians, scientists, teachers, decision makers, communicators, and citizens. During our preliminary conversations about climate change it was suggested by the design team that personas are a great way to identify user needs. As such, this document presents one such persona which takes into account the needs of these various users who want to get insight into the controversies as they enfold on Wikipedia. The document was written by a leading expert in the field, who himself is a teacher, scientist and citizen. Through discussions with a Wikipedia administrator and journalists we have made sure that the system would also be relevant for those user groups.

So far we have provided techniques and visualizations for all crucial functions described in this document (the passages marked red), except for a) identifying which hyperlinks are themselves controversial (section 2), b) periodization and the trajectories of controversies (section 5). These will be amongst the things tackled during our second hackathon in July.