Dr. Gabor Mihaly Toth

Archiving the World Wide Web from the Perspective of Digital Humanities

/ New Questions, Old Heuristics /

Thank you very much for inviting me to this very interesting workshop on web archiving. When I was preparing for this talk, I had two questions in my mind:

First:

1. What can we, students of Humanities and Social Sciences, do with the entire Web as gigantic data archive? In other words, what kind of new research questions can we raise and answer?

Second:

2. Why is the archiving of the world wide web is more than crucial for research in the 21st century?

Of course, I did not know the answer, and I hardly believe that anyone can today give an exact answer. So I asked myself, who could help me? A strange person came to my mind, a person whom I first encountered in the Biblioteca Riccardiana in Florence, and later here in the Bayerische Staatsbibliothek. This odd person is named Benedetto Dei, an enigmatic protagonist of the Italian Renaissance. Benedetto, who was born in fifteenth century Florence, was a diplomat and most probably a spy. The Bayerische Staatsbibliothe keeps a late copy of Benedetto's diary. In this diary Benedetto collected a highly heterogeneous set of information: geography, politics, history, as well as personal information. Apparently, Benedetto was passionate about collecting information; his fifteenth century passion for information was pointing towards a new culture, known as ‘culture of curiosity.’ The elite of Europe between the 16th and 18th century was deeply enthusiastic to accumulate information about the natural, social, historical, and political world. Perhaps, this excitement culminated in the French *Encyclopédie*: an enterprise the purpose of which was to put together a universal set of knowledge and information. Here rose the question:

What can we learn from this enterprise or project in the twenty-first century?

I feel that our aim to archive the entire world wide web is somehow similar to the goals of Rousseau, Descartes, D'Alambert, and so on.

But, with the help of a systematically built web archive and with the help of technology, what shall we realize that they could not?

This is the question I shall address today, which will eventually lead to some answers to the question I raised at the beginning of this talk:

What can we, students of Humanities and Social Sciences, investigate in the data archive of the entire world wide web?

To offer you answers, first of all, I needed a method, or a simple heuristic process. In Digital Humanities we like experimenting, so I said, let's experiment with an old procedure, one that is generally ascribed to St. Augustinus: *tolle et legge*! Take and Read! As he is accounting it in his *Confessions*, Augustinus took the Bible, opened it and started to read at a random place. This led to his revelation and conversion. My idea was to repeat this procedure but with the French *Encyclopédie* and then with Google: let's open a random volume of the French *Encyclopédie* at a random place, and let's read a random entry. After this, let's search for this random entry with Google.

What I was hoping was that a simple comparison of the hits by Google with the random entry in the French *Encyclopédie* can actually give me ingredients for reflecting on the question, what shall we achieve with the help of technology in the twenty-first century that Enlightenment could not?

As a next step, I went to the Rare Book Department of our Library, and I told the librarian, fetch a random volume of the French *Encyclopédie*. He brought the 9th volume, I opened it, and the random entry was *Lampedusa*, the tragic but beautiful island of Lampedusa. There was only short and concise entry about Lampedusa. This short entry gave account of the geography of the island, and provided me with some details on the history of the Island.

After this, I googled Lampedusa.

First of all, what grabbed my attention was the beautiful images of the Island: nice sea shores, deserts, yachts, and so on. Then I was looking at maps, and listening some music from Lampedusa. Finally, I was reading a bit about the local culture of Lampedusa. As a whole, in 5 minutes I could nourish my eyes, my ears and my

brain. I was exposed to a great variety of information channels or mediums such as maps, music, images and so on.

Of course, our eighteenth century friends had no access to such variety of information channels. This is as obvious as the fact that we in the 21st century have access to a great diversity of information. But this obvious manifold of information is leading to a terribly difficult research question to be investigated through a web archive:

How do different types of information channels correlate and influence each other?

When I was reading news about Lampedusa, the key topic was of course migration crisis. Similarly, the recurrent theme on the images of Lampedusa was migration. But is there a systematic correlation between key topics in written and visual media? Or is this correlation only due to google and its algorithm? Similarly, are salient themes in blogs also outstanding on the images of Flickr and Instagram? Or do the most popular hashtags in Twitter have any impact on visual attentions and visual culture?

After looking at images and maps, and listening to music, I wanted to read a bit about Lampedusa. The entry in the algorithm is giving some hints into the history and geography of the Island, but not too much. I like nature, so I was reading about the flora and the fauna of Lampedusa. I learned that dears were taken to the island by the noble family, named Tomasi in the 19th century. I was also reading about the relationship between fishery, fishes living around the island and history. I could thus very quickly gather information on how history, nature, tourism, economy and so on, intersect.

I contend that the systematic division of knowledge and information into different branches and fields, as it is done in the French *Encyclopédie*, is what we need to exceed in the twenty first century. I do believe that by archiving the world wide web, we can study the following question:

How do different areas of life and domains of knowledge affect each other?

But at this point, you might say, the French *Encyclopédie* is not everything. That is true. So I looked up Lampedusa in other lexicons. I checked Lampedusa in a Konversationslexicon. There was a short and not very informative description. I then checked the word Lampedusa in the *Encyclopedia Judaica*. To my surprise, there was no entry on Lampedusa. Does this mean that Lampedusa have no Jewish

connections? Of course not. Again, I googled the words, *ebreio* and *isola di Lampedusa*. There was a great number of hits. I started to look at them, and after 5 minutes I realized that there was a fundamental problem. I am sure, you all know what this problem is. The ranking by Google seems to be completely arbitrary. On the one hand, it is the result of a non transparent algorithm about which we have very few pieces of information. On the other, it must reflect the commercial interest of Google. If I do a research on the connection between the Island of Lampedusa and Jewish people, I do not care about the commercial interest of Google.

The problem is clear: even if I am a knowledgable person in information retrieval, I cannot write and run my own algorithm to mine information from the world wide web. I am at the mercy of a multi-national company. For this reason, archiving the world wide web, and make it available for scientific research is essential. This leads to another question.

We like thinking that information and knowledge are becoming less authoritative and more pluralistic in the 21st century. Whereas in the 18th century, it was a privileged group of intellectuals who told what one had to think about Lampedusa, today, we tend to think that the Internet is giving rise to a broad democratization of knowledge; the web is presenting an extremely large set of information from many different perspectives. Theoretically, I can read for instance different perspectives on the migration crisis and Lampedusa. I can read the opinions of the local fishers and habitants. I can read about how migrants perceive the island. But isn't this entire process of democratization of knowledge and access to information only an illusion? This is in fact a key question in Political and Social Sciences today. For me, for the data scientist, the question is how can I study the assumption that World Wide Web is leading to democratization of knowledge and information with the help of data? But the only source of help I have is again Google. But Google does not help me to distinguish different types of information such as comments from original web content, and it is presenting information in a seriously biased way.

Finally, a last point: knowledge and information are not static. They are both changing in space and time. I wanted to read a bit about how Lampedusa was described in travel blogs before the migration crisis. I couldn't. Google did not offer me any option to search for web content before 2005. Again, this can be researched only if the web is systematically archived.

In summary, at the beginning of this talk, I set the goal of coming up with some new questions that we can investigate only by archiving the entire World Wide

Web. These are the broad research questions that can lead to more tangible and concrete investigations to be studied with data science:

1. How do different information channels correlate and influence each other?
2. How do different areas of life and domains of knowledge intersect?
3. Is the birth of Internet leading to democratization of knowledge and information?
4. How does information change in space and time?

My other goal was to address why the archiving of the world wide web by a public institution is essential. As I have pointed out the common and obvious experience, Google and other commercial search engines are not appropriate tools for research: they are presenting information in a highly biased way that we, researchers, are unable to assess.

As a short conclusion, I would like to highlight the absolute importance of two things. First, the importance of integrating randomness into our research practices. The size of the world wide web is so incredibly large that we do not even know where and how to start investigations. In this situation, randomness is what can lead to new and surprising results. But there is a key problem with this. If I submit an application that uses random heuristics to a funding agency such as Europian Research Council or DFG, it is very likely that my application will be rejected. How can I include unpredictability into a project plan that has to present the expected outcome of my project?

Second, in the last decades or perhaps in the last centuries, we have developed a hubris: we think that we have to look towards the future and neglect the wisdom of the past. We would hardly apply the *tolle et lege* approach by Augustinus, and try to answer the questions of present and the future by investigating the knowledge construction by the Enlightenment. I think, this is mistaken. If there is something to learn from Benedetto Dei and his generation, who renewed human civilization by drawing on the achievement of the antiquity, is that to face the challenges of the present and the future, we need the wisdom and the practices of the past as springboard.